Pontifícia Universidade Católica
do Rio de Janeiro

Andre Nascimento Alcantara Pereira

Informal housing, spatial spillovers, and labor market access in Brazil

Dissertação de Mestrado

Thesis presented to the Programa de Pós–graduação em Economia, do Departamento de Economia da PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Economia.

Advisor: Prof. Thierry Andre Louis Verdier

Rio de Janeiro
April 2022

Pontifícia Universidade Católica
do Rio de Janeiro

Andre Nascimento Alcantara Pereira

Informal housing, spatial spillovers, and labor market access in Brazil

Thesis presented to the Programa de Pós–graduação em Economia da PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Economia. Approved by the Examination Committee:

Prof. Thierry Andre Louis Verdier
Advisor
Departamento de Economia – PUC-Rio

Prof. Gabriel Lopes de Ulyssea
University College London

Prof. Juliano Junqueira Assunção
Departamento de Economia – PUC-Rio

Rio de Janeiro, April the 11th, 2022

Andre Nascimento Alcantara Pereira

B.Sc. Physics, Universidade Federal de Minas Gerais (UFMG), 2017; M.Sc. Physics, University of Waterloo, 2019, and São Paulo State University (UNESP), 2020.

Bibliographic data

Pereira, Andre Nascimento Alcantara

Informal housing, spatial spillovers, and labor market access in Brazil / Andre Nascimento Alcantara Pereira; advisor: Thierry Andre Louis Verdier. – 2022.

47 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Economia, 2022.

Inclui bibliografia

1. Economia – Teses. 2. Economia urbana. 3. Habitação. 4. Spillovers espaciais. 5. Acesso ao mercado de trabalho. 6. Região Metropolitana de São Paulo. I. Verdier, Thierry Andre Louis. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Economia. III. Título.

CDD: 330.9173/2

I dedicate this to my wife, for her unwavering
support and the happiness she brings me.

Acknowledgments

I would like to first thank my advisor, Prof Thierry Verdier, for his support and outstanding mentorship. I would also like to thank the committee members, Profs Gabriel Ulyssea and Juliano Assunção, for their participation and insightful comments. And finally, I wish to thank Prof Gustavo Gonzaga, for the significant role he played in my education at PUC-Rio.

Then, I wish to thank the partial financial support of PUC-Rio, which made this work possible. O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo qual sou grato. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Finally, I am grateful to my family and friends, for helping keep life light and fun during this period. Among these, I am especially thankful to my classmates and colleagues at PUC-Rio, for creating a great collaborative environment, even with the challenges of a long distance program and a pandemic.

Abstract

Pereira, Andre Nascimento Alcantara; Verdier, Thierry Andre Louis (Advisor). **Informal housing, spatial spillovers, and labor market access in Brazil**. Rio de Janeiro, 2022. 47p. Dissertação de Mestrado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

In this work, I study the supply and demand for housing in the São Paulo Metropolitan Area, a major city in Brazil. Using detailed commuting data, I estimate a quantitative spatial model, in which agents make decisions on residence and workplace based on local rents, wages, commuting costs, and amenities. I propose an extension of the usual framework with a formal housing supply sector to include a competing informal one, an important institutional characteristic present in many developing countries. I quantify the spatial spillovers of this informal housing, and investigate its role in providing residents with improved access to the local labor market.

Keywords

Urban economics; Housing; Spatial spillovers; Labor market access; São Paulo Metropolitan Area.

Resumo

Pereira, Andre Nascimento Alcantara; Verdier, Thierry Andre Louis. **Habitação informal, spillovers espaciais e acesso ao mercado de trabalho no Brasil**. Rio de Janeiro, 2022. 47p. Dissertação de Mestrado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

Neste trabalho, estudo a oferta e demanda por habitação na Região Metropolitana de São Paulo, uma das principais cidades do Brasil. Usando dados detalhados de deslocamento ao trabalho, eu estimo um modelo quantitativo espacial, no qual agentes tomam decisões sobre local de residência e trabalho com base em aluguéis, salários, custos de deslocamento e amenidades. Proponho uma extensão do arcabouço usual com um setor formal de oferta de moradia para incluir também um setor informal em competição, uma importante característica institucional presente em diversos países em desenvolvimento. Eu quantifico os spillovers espaciais associados ao setor informal e investigo seu papel em prover residentes com melhor acesso ao mercado de trabalho local.

Palavras-chave

Economia urbana; Habitação; Spillovers espaciais; Acesso ao mercado de trabalho; Região Metropolitana de São Paulo.

Table of contents

List of figures

List of tables

1 Introduction

More than half the world population currently lives in cities, but adequate and affordable housing remains out of reach for many. Over a billion people, or about one quarter of the total urban population, currently lives in slum or *slum-like* conditions (UN-Habitat, 2020), and disproportionately so in low-income countries. Yet, our understanding of the effects of these settlements within the wider urban economy remains limited (Glaeser and Henderson, 2017). As Bryan et al. (2020) point out, there is an ongoing debate on whether slums provide a path to prosperity or an economic dead end. This thesis contributes to an emerging branch of the urban and development economics literature that tries to reconcile the main proposed mechanisms for these competing effects in structural models of city structure.

There are two main questions I seek to answer about these informal settlements in my setting. First, what are their equilibrium consequences for the distribution of labor market access across the city, and particularly for the people who live in these settlements. Second, how do the general equilibrium spatial spillovers of the regions with informal housing affect the overall city structure as a whole. To tackle these questions, I extend the standard quantitative urban model framework with an informal housing sector. Then, I collect comprehensive data for the São Paulo metropolitan region, and propose an estimation method that is feasible with available data. With the estimation done, we can conduct interesting counterfactual analyses to shed light on the above topics.

The provision of dense and inexpensive housing in proximity to economic opportunities is an essential channel through which slums can serve as stepping stones for low-income people, and particularly so for recent rural migrants (Cavalcanti Ferreira et al., 2016). The recent literature on quantitative urban models (see Ahlfeldt et al., 2015; Redding and Rossi-Hansberg, 2017) presents us with a natural measure of labor market access, which is an important part of this story. It captures the way in which nearby production is important for the overall attractiveness of an area for residential use, coming from access to high wages at low commuting costs. In this work, I compare the spatial and populational distribution of this measure in an estimated extension of these models including an informal housing sector and in a counterfactual where this sector does not exist.

The extension to incorporate informal housing into the model relies on recent insights from Henderson et al. (2021). While studying the dynamic evolution of Nairobi and its built environment, the authors emphasize the existence of well developed markets for both the formal and informal housing sectors. The main friction, they argue, is in the conversion between these two forms of land use. They model the different construction sectors as possessing different technological capabilities. The formal sector explores the height margin intensively, meeting increased demand for floorspace by building taller structures. Meanwhile, the informal sector explores a cover margin. When faced with increased demand, it develops a larger share of the terrain, in a crowding process that can result in significant negative agglomeration externalities. These institutional features can be incorporated in a quantitative urban model, as in fact is also being done in parallel by Gechter and Tsivanidis (2020), in an early working paper.

A common difficulty in applying quantitative urban models to developing countries is the availability of data. Sturm et al. (2021), for example, propose alternate methods using mobile phone data for estimating commuting costs, and satellite data on building heights in place of prices of land or floorspace. In our setting, the metropolitan region of São Paulo, we have good data sources that we can use in this estimation. There is comprehensive and high-resolution data from mobility surveys on commuting patterns and, while there is no reliable data on the price of floorspace, for the city of São Paulo proper we can infer floorspace supply directly using a LIDAR-based 3D scan of the city.

The remainder of this thesis is organized as follows: In chapter 2, I present more details about the setting studied and each of the data sources used. In chapter 3, I present the theoretical framework of the quantitative spatial model that we will use. In chapter 4, I present the gravity equation coming from the model and the reduced form estimate of the commuting costs parameter obtained from it. In chapter 5, I present, how starting from the observed data and values for the model parameters, we can recover the other variables including locational characteristics compatible with an equilibrium solution of the model. In chapter 6, I present the calibration procedure that we will use, a generalized method of moments (GMM) estimation, and its results. In chapter 7, I present the procedure for obtaining counterfactual predictions as well as the counterfactual results for the model with all informal housing substituted by formal sector housing. Finally, in chapter 8, I present a discussion of the rest of this work and possible conclusions.

2
Background and Data

2.1
Background and the São Paulo Metropolitan Area

We focus on the São Paulo Metropolitan Region (RMSP). This metropolis is home to over 23 million people (IBGE estimate, 2016), the largest in Brazil and second largest in the southern hemisphere, second only to Jakarta. Of this population, about 11%—over 2 million people—live in informal settlements [1].

The metropolitan area is comprised of 39 contiguous municipalities, which can be seen from the thicker red borders in figure 2.1. The central and largest one, both by land area and population, is the city of São Paulo, the state capital with over 12 million people.

The region also has highly integrated labor markets. From our commuting data, we estimate that about 19% of the working population works in a different municipality than the one they reside in. This flow is particularly strong

[1] Estimated using 2010 census data

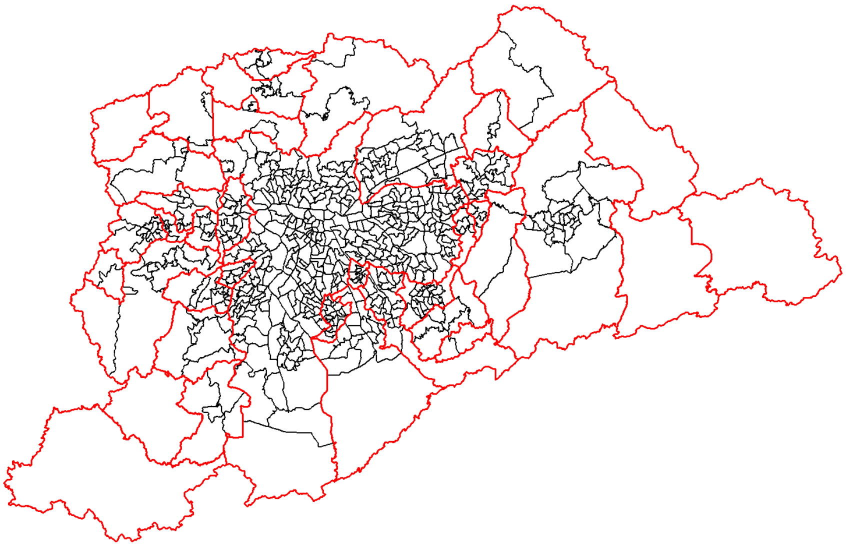


Figure 2.1: Limits of the municipalities (red) and statistical areas (black) in the RMSP.

between the central city and the other municipalities, which alone accounts for 12% of the working population, or 65% of these cross-municipality commuters. Therefore, it is important that we consider the metropolitan region as a whole in our analysis.

2.2 Census data, informality and spatial units of observation

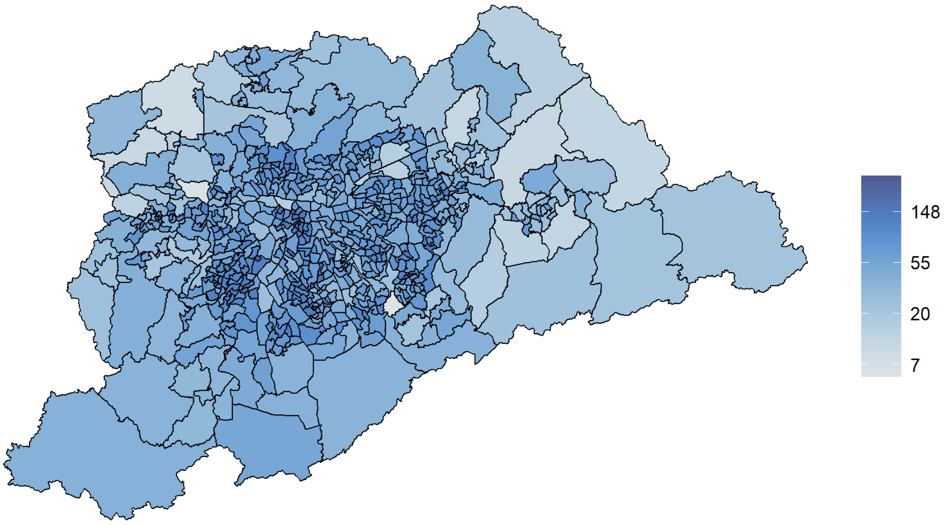
The first data I collect comes from the census. As a geographical unit of observation, I will focus on census statistical areas. For the RMSP, there are 633 of these in the 2010 version, each fully contained in a municipality, which can be seen in figure 2.1. At the present for this work, we do not use any of the census microdata, which is identified precisely at this observation level. Still, the statistical areas present both a relatively homogeneous socioeconomic spatial unit and the compelling possibility to later adapt and incorporate this information more readily, particularly when the results of the next census are released, which is why I choose to work at this level.

The other important information I obtain from the census is the locations of informal housing. I use the classification of census tracts into *aglomerados subnormais*, which are characterized by the "*irregular occupation of land owned by others—public or private—for housing purposes in urban areas and, in general, characterized by an irregular urban pattern, lack of essential public services and location in areas with restricted occupation*" (IBGE, author's translation). This data is already available since 2019 in a pre-release ahead of the 2022 (postponed from 2020) census, and since that is closer to our period of observation I choose to use those tracts.

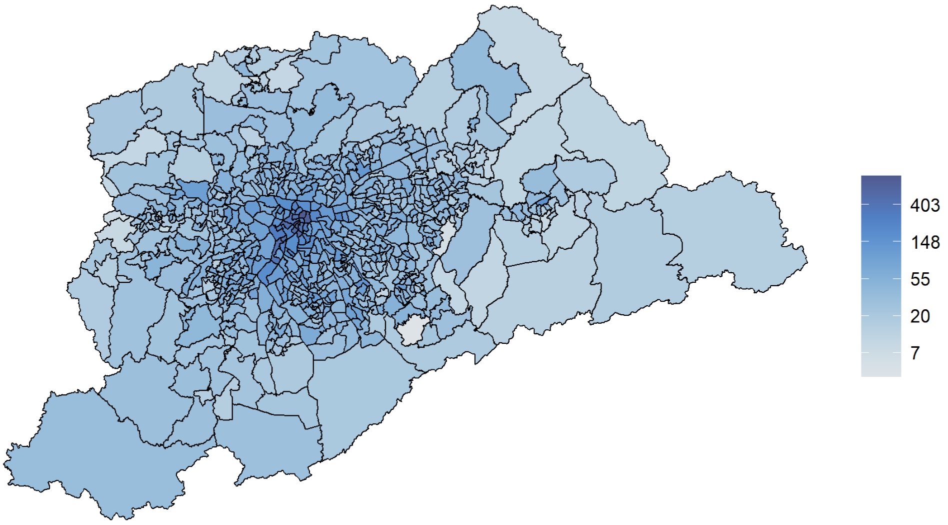
2.3 Origem-Destino survey

The most important data source I use in this work is the São Paulo metro company's *Origem-Destino* (origin-destination) survey. Conducted in partnership with government offices of the state of São Paulo, this survey is used for policymaking in mobility and commuting for the metropolitan area. The process used to ensure the sample is representative across the different areas is guided by the demographic and populational estimates of IBGE, the government agency which is among other things responsible for the census.

I use the microdata for the 2017 edition of the survey, publicly available on the metro company's website, for which a sample of 40,916 workers were interviewed. Crucially, among the collected data are geographic coordinates for their place of residence and work. Additionally, I obtain data on their commute



2.2(a): Residential working population density.



2.2(b): Employment density.

Figure 2.2: Densities (workers per built-up ha) in the RMSP's statistical areas.

time, and albeit with more limited coverage, on their wages. Also included is the statistical weight of each observation.

Combining with the census geographical data, I obtain the statistical area and formality status of residence and employment[2] for each worker. In figure 2.2, I map the resulting density of residence and employment obtained in each area. As is typical, it can be seen from this map that in São Paulo too employment is more densely concentrated in the central area than residence.

Using this data, I separate the statistical areas with sizable formal and informal sectors into two regions of observation. The precise criteria used is that there should be at least one resident and one worker in the region, and the buildable area–explained in a following subsection– should be at least 15 ha., or about 15 typical blocks, in size. This is done to avoid introducing a large number of very small regions of observation for which our data would not be reliable, particularly in the informal sector. Other regions are assigned a single sector based on which is largest (in population).

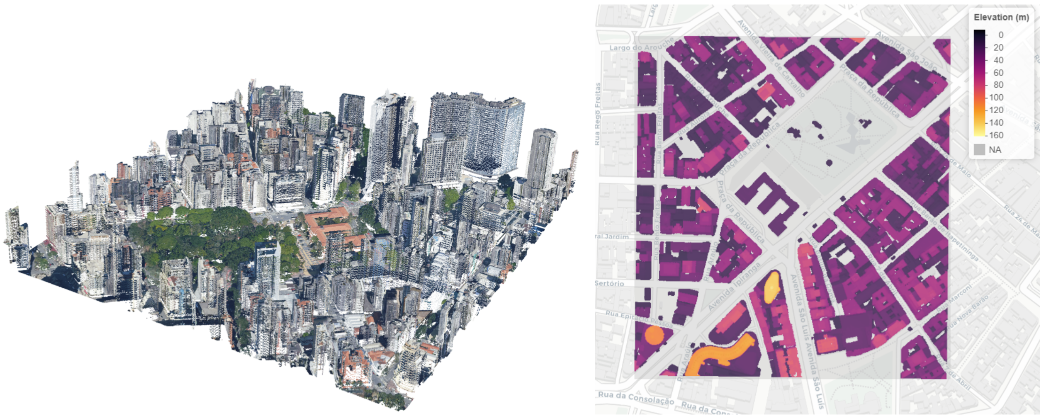
Therefore, from this data, we have 821 regions of observation, 629 of which are formal and 192 informal ones. For each, we have the number of (working) residents and the number of workers employed, which is the main type of data necessary to estimate quantitative urban models. Finally, we also have the flow between each pair of regions, which we will not use directly in the main estimate, but rather in the exercise of estimating the model's gravity equation directly, presented in section 4, which is also a useful robustness check.

2.4
LIDAR, built height, and the local supply of floorspace

Between May and July 2017, a complete LIDAR scan of the city of São Paulo was conducted at the request of the municipal government. LIDAR is a 3D-sensing technology, using timed laser pulses to infer distance in a manner similar to what radar does with sound. Since laser light is directed, it is then possible to reconstruct a high-resolution point-cloud model of the scanned surface, as exemplified in figure 2.3(a).

Since 2019, the geoinformation services portal of the city of São Paulo has made this data publicly available. Their preprocessing involves the usual classification of points in the surface of the city in a set of categories, which allow us to differentiate buildings from the rest, as well as the the calibration of a digital terrain model, which gives us height relative to the ground level instead of in absolute.

[2] In the sense of it being in a formal or informal area, not necessarily a formal or informal job.



2.3(a): Point-cloud data rendering. 2.3(b): Built height map.

Figure 2.3: Example LIDAR-derived data. Area around Praça da Republica, near the city centre.

I first process this data into a built height grid with a resolution of 1m, covering the entire city, an example of which is shown in figure 2.3(b). Then, we sum it over each area of observation to obtain the built volume. We use this as a proxy for the total supply of floorspace in the area, up to an unimportant constant factor representing the average height of a floor.

2.5 WorldCover and buildable area

An essential input in the model for housing supply will be the available buildable area in each geographical unit. A problem with using just the area of the statistical area as a whole is that, particularly as one goes further away from the city's core, there are large areas covered by water, protected forests, or even cropland. To avoid this problem, we use the European Space Agency's WorldCover dataset (Zanaga et al., 2021).

This is publicly available and based on imagery from the Sentinel-1 and Sentinel-2 satellites, taken in 2020, this presents a worldwide classification of *land use* into categories such as grassland, open water, cropland and built-up, at a 10 meters resolution. The area we assign to each region then will be the area of the part of the region that is classified by this data as built-up.

2.6 Transport network and travel times

Finally, we obtain data on the street and transport network, from OpenStreetMaps. We increment it with GTFS data on public transit lines and schedules from SPTrans, the São Paulo city company responsible for administering the municipal public transit system, and EMTU, the São Paulo state company responsible for doing the same at the metropolitan area level.

Using this, we calculate the travel times matrix between the mean locations of residence and employment of each spatial unit. This is done using the r5r package (Pereira et al., 2021), an interface between the R programming language and the realistic multimodal routing package R5.

3 Theoretical Model

The model used is a quantitative spatial model, similar to that of Ahlfeldt et al. (2015) and Heblich et al. (2020). The main departures are the addition of an informal construction sector in the regions classified as informal, and in explicitly including the area extensive factor in the preference shocks. The informal construction sector is inspired by the one presented in Henderson et al. (2021), and somewhat similar to the one in Gechter and Tsivanidis (2020). We start its presentation with the workers' problem.

3.1 Workers

In this model, a continuum of workers with heterogeneous preferences decide on a place of residence and work. A *place* is, in this context, one of the regions discussed in the previous section.

They make this decision considering the workplace wage and floorspace prices—which determine their possibility of consuming housing floorspace and of a numeraire tradable good—as well as residential amenities, commuting costs, and their own idiosyncratic preferences. More formally, the utility of worker o living at n and working at i is

$$U_{nio} = z_{nio} \frac{B_n}{\kappa_{ni}} \left(\frac{c_{ni}}{\alpha}\right)^{\alpha} \left(\frac{h_{ni}}{1-\alpha}\right)^{1-\alpha}, \tag{3-1}$$

where c_{ni} is consumption of the tradable good, h_{ni} is consumption of floorspace, α is a Cobb-Douglas preference parameter, B_n are exogenous amenities at place of residence n, κ_{ni} are commuting costs between the place of residence and work, and z_{nio} are preference shocks independently drawn from Fréchet distributions, $F_{ni}(z) = \exp(-T_n T_i \, z^{-\varepsilon})$, where T_n is the land area of location n. We will assume that commuting costs are determined by travel times τ_{ni} as $\kappa_{ni} = \tau_{ni}^{\phi}$, $\phi > 0$, that is, they increase with a constant elasticity.

The agent is also restricted by the budget condition $c_{ni} + Q_n \, h_{ni} \leq w_i$, where the wage w_i is solely determined by the place of work, while floorspace rent prices Q_n are determined at the place of residence. Solving the Cobb-Douglas maximization problem over the consumption of the two goods, if a worker were to choose a place of residence and work he would then have an indirect utility of

$$V_{nio} = z_{nio} \frac{B_n}{\kappa_{ni}} \frac{w_i}{Q_n^{1-\alpha}}. \tag{3-2}$$

It is worth noting that for a worker given all these characteristics, there's no intrinsic difference between a place being formally or informally developed. If an informally developed place is to be less desirable, that should be represented by lower amenities, or instead arise indirectly, for example from lower productivity depressing wages or the construction technology resulting in higher rents.

Given equation (3-2), through standard manipulations of the type II extreme value distribution which are presented in the appendix A.1, we can obtain that the fraction of workers that reside in n and work at i,

$$\lambda_{ni} = \frac{L_{ni}}{L} = \frac{\Phi_{ni}}{\Phi}, \quad \text{where } \Phi_{ni} = T_n T_i \left(\frac{B_n w_i}{\kappa_{ni} Q_n^{1-\alpha}} \right)^{\varepsilon} \quad \text{and } \Phi = \sum_{m,j} \Phi_{mj}. \tag{3-3}$$

Intuitively, this relation tell us that pairs of places of residence and work "pull" workers from one another, with relative strength determined by the characteristics discussed in the first paragraph.

Here, we can also understand the specific functional form chosen for the preference shocks in terms of the regions' areas. As a thought experiment, imagine separating a small and homogeneous enough region into two, with the same locational fundamentals. A desirable property of the model would be that the predictions remain the same, with just the population of the original area divided proportionally between the new regions, and the model does satisfy this property. Had the residence area term not been included, it could still satisfy it if the residential amenity were split between the two new regions proportionally to the ε power of the share of the original area they occupy. Thinking instead of recovering amenities from observed data, this would complicate the interpretation and comparison of these amenities. Meanwhile, with the explicit inclusion of T_n, it does have a nice, "density of amenities" type of interpretation.

On the employment side, the situation would be even worse. Since we do not include workplace amenities, without the T_i factor there would be nothing to split to maintain this property. Again thinking instead of an inversion, rather than obtaining equivalent but harder to interpret results for amenities we would instead mistake high employment due to high area as being due to high wages instead. The explicit factor of T_i then allows us to consistently make the assumption that workplaces are homogeneous, represented by "constant amenities", without compromising the results even in the presence of highly variable areas.

3.1.1 Open city

It is also important that we consider what how the total number of workers L in the city is determined. As is traditional in urban economics models, I will take the point of view that this city is an *open city*. This means that we consider the city to be part of a much larger economy, with an infinite number of other workers sharing a reservation utility $\overline{U}$. Equilibrium in a migration "market" then dictates that the expected utility of moving to the city should be equal to this reservation level.

As shown in appendix A.2, the expected utility of someone living in the city is

$$\overline{V} = \Gamma\left(1 - \frac{1}{\varepsilon}\right) \Phi^{\frac{1}{\varepsilon}}. \tag{3-4}$$

This, together with the open city condition, means that

$$\Phi = \left(\frac{\overline{U}}{\Gamma\left(1 - \frac{1}{\varepsilon}\right)}\right)^{\varepsilon}, \tag{3-5}$$

that is, Φ in equation (3-3) is a constant determined only by the model's parameters of the preference shock and reservation utility.

3.1.2 Commuting market clearing

A natural identity that I will also be able to take to the data in the estimation is the commuter market clearing condition. It says the number of workers in a given area, L_i, is the sum over residence locations of the conditional probability of a worker residing at n commuting to i multiplied by the number of workers residing at n, R_n. Formally,

$$L_i = \lambda_i \cdot L = \sum_n \lambda_{ni} \cdot L = \sum_n \lambda_{ni|n}(\lambda_n \cdot L) = \sum_n \lambda_{ni|n} R_n. \tag{3-6}$$

Where L is total city population (of workers). Using (3-3), this conditional probability is

$$\lambda_{ni|n} = \frac{\lambda_{ni}}{\sum\limits_j \lambda_{nj}} = \frac{\Phi_{ni}\big/\Phi}{\sum_j \Phi_{nj}\big/\Phi} = \frac{T_i(w_i/\kappa_{ni})^{\varepsilon}}{\sum\limits_j T_j(w_j/\kappa_{nj})^{\varepsilon}}.$$

Then, substituting back into (3-6), we obtain the commuting market clearing condition,

$$L_i = \sum_n \frac{T_i(w_i/\kappa_{ni})^{\varepsilon}}{\sum\limits_j T_j(w_j/\kappa_{nj})^{\varepsilon}} R_n. \tag{3-7}$$

3.1.3 Residential population and commuting market access

The residential population (of workers) in a given area in the model can be calculated as

$$R_n = L\sum_i \lambda_{ni} = \frac{LT}{\Phi} T_n \left(\frac{B_n W_n}{Q_n^{1-\alpha}}\right)^{\varepsilon}, \qquad \text{where } W_n = \left[\sum_i \frac{T_i}{T}\left(\frac{w_i}{\kappa_{ni}}\right)^{\varepsilon}\right]^{1/\varepsilon}. \tag{3-8}$$

Here, T is total city area. We refer to W_n in equation (3-8) as *commuting market access*. As mentioned in the introduction, it captures the way in which nearby production is important for the overall attractiveness of an area for residential use, coming from access to large areas of high wages at low commuting costs. Another advantage of this measure is that it incorporates the importance of a diversity of options for workers given their heterogeneous preferences regarding different workplaces, through the dependence on the ε parameter.

3.1.4 Expenditure on floorspace

Due to the nested Cobb-Douglas structure of the utility function, each worker spends on residential floorspace a constant share $1-\alpha$ of their income. Thus, the total expenditure on floorspace at location n is given by

$$\mathbb{Q}_n^R = (1-\alpha)\sum_i L_{ni} w_i = (1-\alpha)\, R_n \sum_i \lambda_{ni|n} w_i \tag{3-9}$$

3.2 Production

I adopt a simple specification for the production sector, which is not the main focus of my analysis. I assume competitive production of the tradable good y_i at every workplace area i, with a constant returns to scale technology utilizing labor (L_i) and floorspace H_i^P, and with an exogenous local productivity parameter A_i as

$$y_i = A_i\, L_i^{\beta} (H_i^P)^{(1-\beta)}. \tag{3-10}$$

From the functional form of the production function, I derive the relationship between the production/commercial part of the expenditure on floorspace and the wage bill of local firms,

$$\mathbb{Q}_i^P = Q_i H_i^P = \frac{1-\beta}{\beta} w_i\, L_i. \tag{3-11}$$

Then, from the zero profits implication of competitive production, we have the relationship between local productivity and production cost characteristics,

$$A_i = \left(\frac{w_i}{\beta}\right)^{\beta} \left(\frac{Q_i}{1-\beta}\right)^{1-\beta}. \tag{3-12}$$

3.3 Floorspace

I include two construction sectors, one formal (F) and one informal (I). Both are taken as competitive in their own areas, but only the formal sector can build in formal areas and similarly for the informal. Each sector has its own technology, with the formal sector exploring the height margin for dense construction while the informal one only does so on coverage.

3.3.1 Formal sector

In the formal sector, construction firms employ capital to convert land into more built space. Their technology is represented by the production function

$$H_n(T_n, K_n) = \hat{h}_F\, T_n^{1-\hat{\mu}}\, K_n^{\hat{\mu}} = \hat{h}_F\, T_n\, k_n^{\hat{\mu}}, \qquad \hat{\mu} \in (0,1), \tag{3-13}$$

where K_n is capital, and equivalently $k_n = K_n/T_n$ capital intensity, or capital per unit land. We can see from this way to rewrite this function that the total amount of housing supplied is extensive in the amount of land and concave in the local capital intensity chosen. We can then write the profit function of the representative local construction firm as

$$\pi_n = T_n \left(h_F\, Q_n\, k_n^{\hat{\mu}} - r k_n - q_n \right), \tag{3-14}$$

where r is the exogenous price of capital, equalized across the city, and q_n the rental price of a unit of land. From profit maximization on the capital intensity,

$$k_n = \left(\frac{\hat{\mu}\, \hat{h}_F}{r} Q_n \right)^{\frac{1}{1-\hat{\mu}}}, \tag{3-15}$$

and then from the zero profits condition that the firm will be indifferent about the amount of land, a share $\hat{\mu}$ of the revenue from this sector will be payed to capital, while $1 - \hat{\mu}$ will go to land.

Further assuming that land owners will choose to supply all the available land in their area T_n for construction, the supply of floorspace will then be an isoelastic function of local rent prices as

$$H_n(Q_n) = h_F\, T_n\, Q_n^{\mu}, \qquad \text{where } h_F = \left(\frac{\hat{\mu}}{r}\right)^{\frac{\hat{\mu}}{1-\hat{\mu}}} \hat{h}_F^{\frac{1}{1-\hat{\mu}}}, \quad \mu = \frac{\mu}{1-\hat{\mu}} > 1. \qquad (3\text{-}16)$$

For calculational ease, from now on I will only refer to these equivalent, transformed parameters.

3.3.2 Informal sector

In the informal sector, land is converted directly into floorspace without additional capital costs or the possibility of intensified use, as

$$H_n(T_n) = h_I\, T_n. \qquad (3\text{-}17)$$

Note this results in a marginal cost of floorspace of q_n/h_I, constant for any given location. From the zero profit condition, all revenue from this activity is passed on to land owners, with the informal construction sector indifferent about the operation scale and the land owners choosing again to supply all their available land. The resulting supply of floorspace is, therefore, independent of rent price

$$H_n(Q_n) = h_I\, T_n. \qquad (3\text{-}18)$$

The rationale for this model is that the informal sector utilizes cheaper construction materials and techniques to build housing possibly at a lower cost, but those are not appropriate for attempting to build taller structures. The capital expenditures of the formal sector would also be harder to justify in many cases, as the frequently insecure legal tenure regime of the land make long-term investments too risky. Meanwhile, the formal sector refrains from using the same technology even if it would be financially advantageous due to factors such as regulatory requirements or expectations about quality from its usually higher-income consumers.

To make the model's intuition more precise, we could introduce an additional construction materials component to this production function, as a perfect complement to land with exogenously set price. Then, if the rent price was high enough, the same optimal behaviour would ensue, just with the cost of the materials diverted from the land owner's share of profits. Otherwise, if it weren't, the land would simply not be developed. As we focus our analysis on urban informal areas that have been developed, this extension would lead to the same results, with an upper bound rather than estimate for the unit costs of these materials. As such, and also since we neither observe nor are particularly interested in the revenue of land owners, I choose this simpler specification instead. It is still worth mentioning this, though, as it would have been an important inclusion if our setting was instead a newer, expanding metropolitan region, as those in Sub-Saharan Africa or South Asia. With most informal developments at its borders, the profitability rent lower-bound coming from the material cost would be an important determining factor of what land gets developed or not.

4
Gravity

The first empirical test we will do is related to the commuting flows probabilities between residences and workplaces as given by equation (3-3), compared to the ones in the survey data. In log form, and considering the fact that Φ is a constant, as given by equation (3-5), we obtain the following *gravity equation* to be estimated using the data

$$\log(\lambda_{ni}) = -\nu \log(\tau_{ni}) + \psi_n^R + \psi_i^W + u_{ni}. \tag{4-1}$$

Where the composite parameter $\nu = \phi\varepsilon$. The place fixed effects $\psi_i^W = -\log(\Phi) + \log(T_i) + \varepsilon \log(w_i)$ for employment and $\psi_n^R = \log(T_n) + \varepsilon \left(\log(B_n) - [1-\alpha]\log(Q_n)\right)$ for residence are like "masses" determining the "gravitational pull" that each residence or workplace has on workers. We include an error term, u_{ni}, to account for the limited-size sample available in the survey data from which we obtain the λ_{ni} for our estimate.

Since Eaton and Kortum (2002), models generating predictions similar to the one above have been widely studied in the international trade literature. An important consideration is that due to the large number of pairs of locations, it is usual for the observed matrix of flows to be sparse, that is, many of its entries are zero. Direct regression methods such as ordinary least squares (OLS) on equation (4-1) in its presented logarithmic form have to ignore these observations if they are to avoid the problem of mathematical undefinedness. Since these excluded pairs are more likely to have both higher travel times and smaller real flows, this leads to a form of selection bias. Intuitively, by this logic the OLS estimate should typically underestimate how fast λ decreases with travel time. A common method for dealing with it, is Poisson pseudo-maximum estimator (PPML) of Silva and Tenreyro (2006). Originally formulated to deal with another source of bias, related to Jensen's inequality in the presence of heteroskedasticity, it has been shown to be robust to the presence of this sort of sparsity (Silva and Tenreyro, 2011).

In table 4.1, I present results for both the OLS and PPLM estimation. In our sample, there are just over 17 thousand non-zero pairs out of a total of about 850 thousand, which is under 2% of the total. Thus, this problem is present and my preferred method of estimation is the PPLM.

Table 4.1: Gravity estimation

	$\log(\lambda)$	λ
Estimation	OLS	PPML
	(1)	(2)
$\log(\tau)$	-0.6160***	-1.884***
	(0.0101)	(0.0094)
R^2	0.58616	
Observations	17,019	853,776
Residence fixed effects	✓	✓
Workplace fixed effects	✓	✓

Estimation results of the gravity model (4-1). OLS denotes ordinary least squares. PPLM is Poisson pseudo-maximum likelihood. Standard errors, in parenthesis, are heteroskedasticity robust: * denotes statistical significance at the 10 percent level; ** denotes statistical significance at the 5 percent level; *** denotes statistical significance at the 1 percent level.

The obtained value of $\nu_{\text{pplm}} = 1.88(0.01)$ differs significantly from the value of $\nu_{\text{ols}} = 0.62(0.01)$. We will later compare this estimated value to the one obtained from the full model estimation procedure, which does not rely on the flow data directly, and only on the populations at each residence and workplace. We will see that indeed the results obtained from the PPML method match that estimate well.

5
Recovering Local Variables

In this section, I will show how, given values for the model's parameters, one can invert from observed data on residential and workplace populations and recover the rest of the local variables, including wages, rents, and amenities. I will highlight what is needed at each step, since the incremental nature of the method is an interesting feature to keep in mind, particularly when thinking about what assumptions are important for each part of the obtained results.

5.1
Population distribution

Given a value for the composite parameter $\nu = \phi\,\varepsilon$, we can calculate $\kappa^{\varepsilon}_{ni} = \tau^{\nu}_{ni}$ Then, since we observe the number of residents and workers at each location, the commuting market clearing conditions (3-7) become a set of equations on the composite local variable $\omega_i = T_i\,(w_i/w_0)^{\varepsilon}$, as

$$L_i = \sum_n \frac{\omega_i/\kappa^{\nu}_{ni}}{\sum\limits_j \omega_j/\kappa^{\nu}_{nj}} R_n \tag{5-1}$$

As shown by Ahlfeldt et al. (2015), this system of equations can be quickly numerically inverted to obtain the value of this composite vector using a fixed-point type algorithm. Note that the system of equations sums to $L = L$, indicating that in general one equation is redundant, as a linear combination of the others. The extra degree of freedom is the overall scale of the ω_i, which we can see by noting that the system of equations is homogeneous of degree zero in them, which is why we include the w_0 factor in the transformation from ω_i to w_i.

Having obtained ω_i, we can readily recover the population distribution, that is, the probability of a worker residing at n and working at i, as

$$\lambda_{ni} = R_n\,\lambda_{ni|n} = R_n\,\frac{\omega_i/\kappa^{\varepsilon}_{ni}}{\sum_j \omega_j/\kappa^{\varepsilon}_{nj}}. \tag{5-2}$$

5.2 Wages

Given the value of ε and the observed area of each region, we can then recover the wages as $w_i = w_0 \left(\omega_i/T_i\right)^{1/\varepsilon}$. Though it has no real impact on our estimates, capturing only the overall price level, to more easily interpret the results we calibrate w_0 so that the (number of workers weighted) mean wage is the same as the mean wage in the survey data.

5.3 Floorspace

Using the wages recovered above, and with the values of the parameters α and β, we can calculate for each region the total expenditure on floorspace. Combining the residential and production expenditures on floorspace, we have that

$$\mathbb{Q}_n = (1-\alpha)\, R_n \sum_i \lambda_{ni|n} w_i + \frac{1-\beta}{\beta} w_n L_n \tag{5-3}$$

Then, we use our assumption about the construction technology to derive from this the value of floorspace rents.

For the formal sector, we multiply the equation for the elasticity of floorspace supply with respect to rents, (3-16), by the rent value and invert the resulting relation to obtain

$$Q_n = \left(\frac{\mathbb{Q}_n}{h_F T_n}\right)^{\frac{1}{1+\mu}} \qquad \text{and} \qquad H_n = (h_F T_n)^{\frac{1}{1+\mu}} \mathbb{Q}_n^{\frac{\mu}{1+\mu}}. \tag{5-4}$$

Thus, given a value for μ, we can obtain floorspace rent and the supply of floorspace in each formal area, up to an overall multiplicative constant related to h_F.

For the informal sector, we can use a similar process by multiplying equation (3-18) by local rent. The process is equivalent to the one for the formal sector with $\mu = 0$. As such, in the informal sector

$$Q_n = \frac{\mathbb{Q}_n}{h_I T_n} \qquad \text{and} \qquad H_n = h_I T_n. \tag{5-5}$$

Finally, I then calibrate h_F and h_I by matching the geometric mean of the predicted supply of floorspace to the one observed using the LIDAR data on the city of São Paulo proper, separately for the formal and informal sectors.

5.4 Productivity and amenities

Finally, we recover productivity and amenities, again up to an overall normalization. Productivity can be readily obtained from wages and rents by using equation (3-12) directly.

To obtain amenities, we need to invert the residential population equation (3-8), arriving at, up to a normalization factor that in this case is equivalent to the overall normalization of the utilities,

$$B_n = \frac{Q_n^{1-\alpha}}{W_n} \left(\frac{R_n}{T_n} \right)^{\frac{1}{\varepsilon}} . \tag{5-6}$$

6 Estimation

Some of our parameters are precalibrated from the literature. Specifically, we use $1 - \alpha = 0.75$ for the share of housing on worker's expenditures and $1 - \beta = 0.2$ for the share of floorspace on firms expenditures, both similar to Ahlfeldt et al. (2015). The rest are estimated using the generalized method of moments with the moment conditions presented below.

6.1 Commuting costs

As we've seen in in 5.1, given a value for the commuting costs composite parameter ν, we can calculate the population distribution. With this distribution, we are then able to calculate the share of the population that commutes at most 30 minutes,

$$\sum_{n,i} \mathbb{1}_{\tau_{ni} \leq 30\text{min}} \; \lambda_{ni}, \tag{6-1}$$

where $\mathbb{1}$ is an indicator function for the travel time condition. This quantity can be directly compared to its counterpart calculated on the survey data, denoted ψ. In order to build an estimator from this equivalence, we rewrite it as the following moment condition

$$g_{\nu,i} = \frac{\psi}{N} - \sum_{n} \mathbb{1}_{\tau_{ni} \leq 30\text{min}} \; \lambda_{ni}, \qquad \mathbb{E}(g_\nu) = 0. \tag{6-2}$$

Here, the factor of N, the number of regions, is used to keep the sample mean rather than sum definition standard in the method of moments.

The intuition for why this should work is that the ν control, given a set transport network, how willing commuters are to travel long distances. For a given ν and observed number of residents and workers at each location, there is a unique joint distribution of the population flowing between these locations. This estimation procedure calibrates ν to the value required to explain the proportion of people in the city that travel up to 30 minutes compared to those who don't.

With this procedure, together with the later moments, we obtain an estimate of $\nu = 1.68(0.05)$. It is worth noting that this method makes use only of the L_i and R_n but not the observed survey flows directly, and as such it

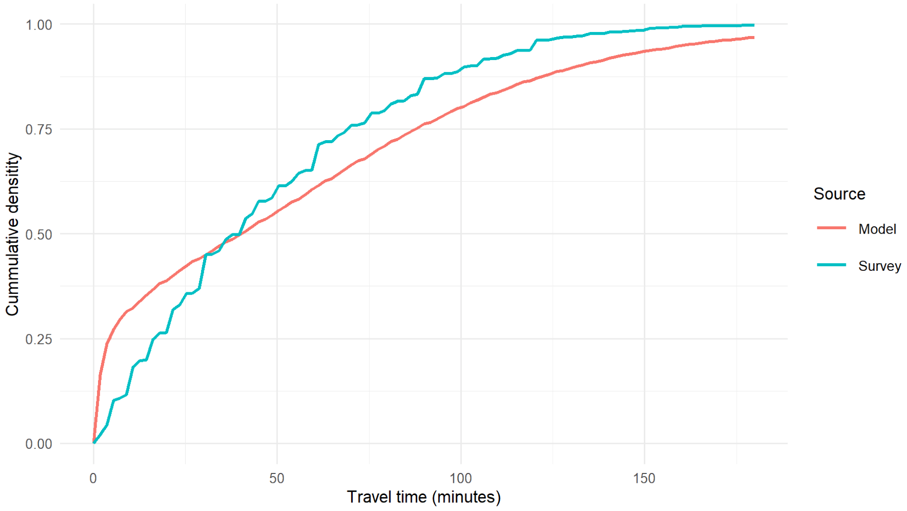


Figure 6.1: Cumulative distributions of travel times, as empirically observed in the survey data and predicted by the calibrated model.

does not suffer from the same problems mentioned in 4. The fact that this value and the one obtained then from the PPLM regression are similar while coming from very different estimation methods gives increased confidence in their validity.

In figure 6.1, we plot the empirical distribution of reported travel times in the survey data against the predictions of the model at the estimated value of ν. We see that it seems to capture well the commuting behaviour of the population of the São Paulo Metropolitan Region.

6.2
Preference shocks

The composite ω_i can be interpreted, from its role in both the commuting market clearing equation (3-7) or the population distribution equation (3-3), as a measure of how much people favor workplace i in their locational choices. From the same equations, we see that an interpretation of the ε parameter is how much differences in locational characteristics are amplified to differences in locational choices. Meaning, depending on ε, these observed differences in favorability can be coming from more or less extreme differences in underlying characteristics. Our estimation moment for ε will be guided by this intuition.

I will use a moment condition on the variance of log wages. First, from the survey data, we calculate σ_w^2, the variance of survey log wages across locations. Then, I define the moment condition

$$g_{\varepsilon,i} = \sigma_w^2 - \frac{N}{N-1}\left(\log(w_i) - \overline{\log(w)}\right)^2, \qquad \mathbb{E}(g_\varepsilon) = 0. \qquad (6\text{-}3)$$

Here, the prefactor just performs the Bessel correction to obtain an unbiased

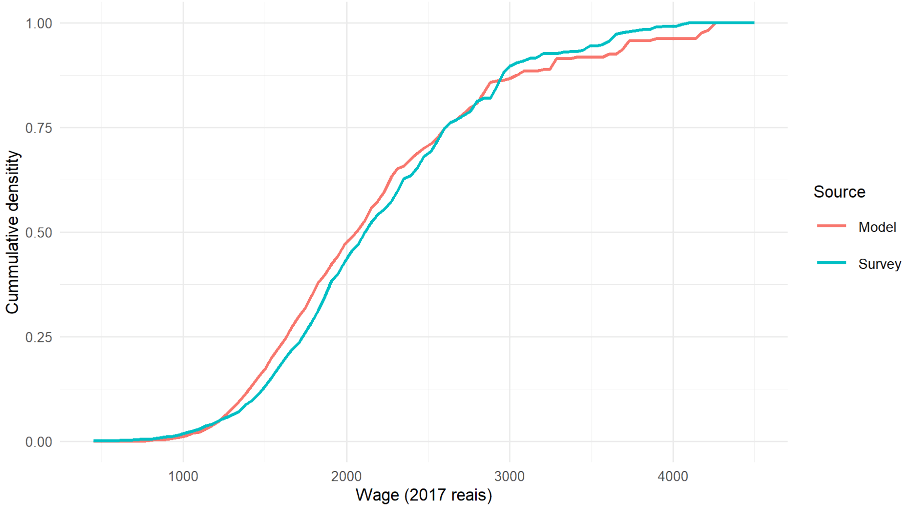


Figure 6.2: Cumulative distributions of wages.

variance estimator. Note that since the wage scale factor is an additive constant in the log, it is cancelled in the deviation regardless of its value, and as such does not impact the estimate.

From the estimate, we obtain the value of $\varepsilon = 3.15(0.65)$. While slightly on the lower side of the usual estimates for this parameter, I believe this result is reasonable. Thinking of ε as amplifying differences in locational characteristics, or equivalently making idiosyncratic factors less important, it makes sense that it would be lower in a metropolitan region of the size of São Paulo, for example as people are not as willing to from one side of the megacity to the other in order to enjoy a particularly good combination of high wages and low rents. Further, given the high inequality in human capital attainment, a factor we are not including in the model, they could not even be able capitalize on that opportunity anyway.

In figure 6.2, we plot the empirical distribution of wages as against the model's predictions with this calibration. From the model, we take the cumulative distribution function of workplace wages weighted by workplace employment. As there is no within workplace variation in the model, for the comparison we also use the average wage per workplace in the survey, weighted by that workplace's number of workers. These two distributions in particular match remarkably well, which is evidence in favor of the argument in the last paragraph, and of the proposed estimation procedure in general. In figure 6.3, we map the obtained wage values in both sectors.

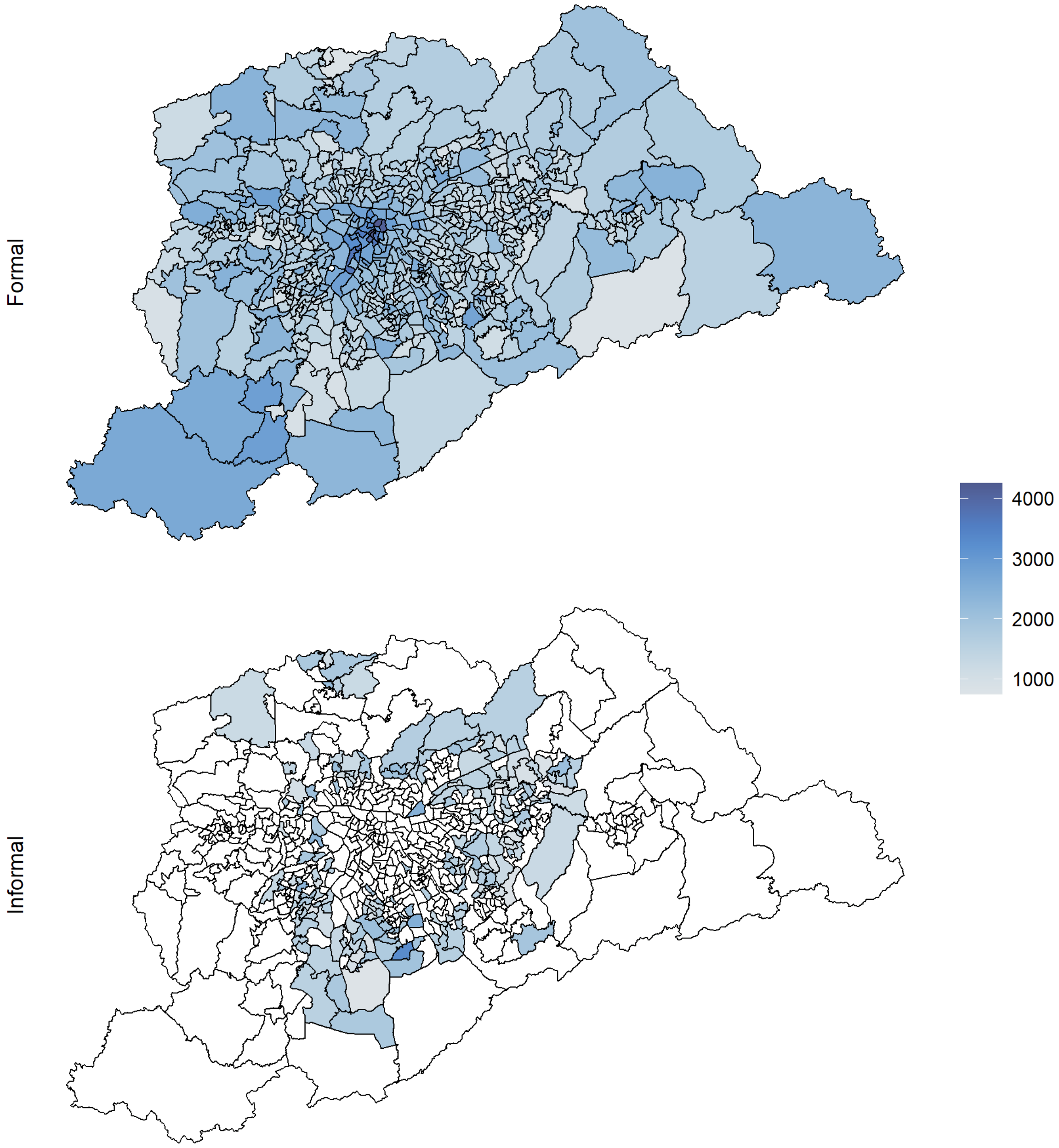


Figure 6.3: Wages in formal and informal workplaces, obtained by the inversion procedure with the estimated parameters.

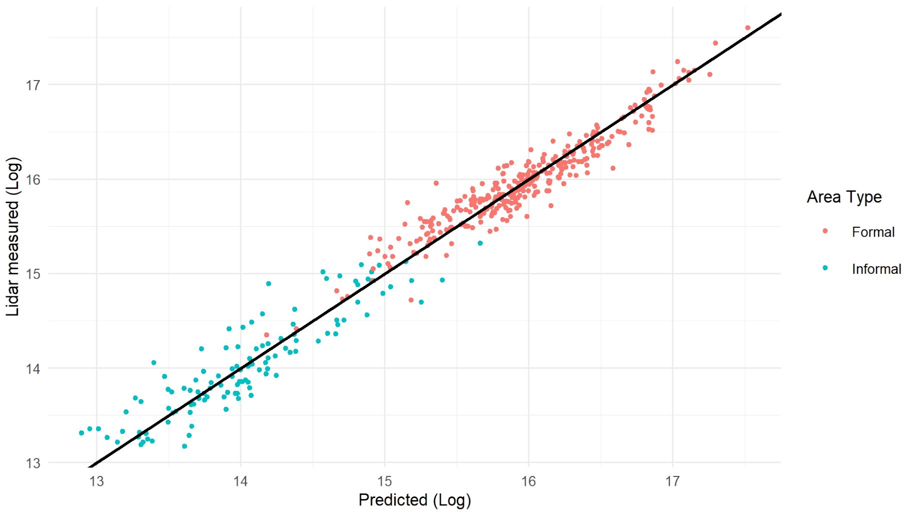


Figure 6.4: Predicted and observed floorspace in the city of São Paulo, with 45° line for reference.

6.3 Formal floorspace

In the formal sector for floorspace, the supply elasticity parameter μ controls how changes on the demand side's expenditure per unit area get translated into changes in the supply of housing per unit area. Therefore, just like with wages and the parameter ε, we will estimate μ by matching the predicted and observed variance in the supply of floorspace per unit area.

Remember that we only have floorspace observations for this comparison for the city of São Paulo, and so we should limit the estimation variables to that region as well. Thus, our moment condition for μ is

$$g_{\mu,i} = \mathbb{1}_{i\in\text{SP, F}} \left[\sigma_F^2 - \frac{M_F}{M_F-1}\left(\log(H_i/T_i) - \overline{\log(H/T)}\right)^2\right], \; \mathbb{E}(g_\mu) = 0, \quad (6\text{-}4)$$

where the indicator function for the formal areas of the city of São Paulo $\mathbb{1}_{i\in\text{SP, F}}$ makes the moment zero when evaluated outside of this scope, and the Bessel factor is taken with M_F, the number of formal areas in the city of São Paulo.

The obtained estimated value for the elasticity parameter is then $\mu = 1.05(0.25)$. For comparison, Henderson et al. (2021) obtain for the city of Nairobi a value of $\mu = 1.42$ [1] , indicating that the supply of housing there is more responsive to prices, or equivalently that the costs of building higher are more steep in our setting.

With this estimate, I plot the predicted versus LIDAR observed supply of floorspace in the city of São Paulo in figure 6.4, again obtaining a good match.

[1] More technically, their estimated parameter is actually related to ours as $\gamma = 1+\frac{1}{\mu} \iff \mu = \frac{1}{\gamma-1}$, and they estimate a value of 1.703.

It is worth emphasizing that the only quantity derived from the LIDAR data used in the estimation and consequently influencing these predictions are the already mentioned variance and mean, so it is reassuring that there is such a good match on a region by region basis. Note also that the predictions for the informal sector also match well.

7
Counterfactual

7.1
Generation procedure

Before presenting counterfactual results, it is useful to discuss the method to obtain them. Focusing on variations $\hat{x} = x'/x$ of variables from their initial state x to their state in the new equilibrium x', we look at set of equations in order.

1. Productivity

   When conducting these counterfactuals, we assume that the exogenous productivity of each location remains unchanged, that is, $\hat{A}_i = 1$. Then, from the ratio of the productivity equation (3-12) before and after

   $$1 = \hat{w}_i^{\beta}\, \hat{Q}_i^{1-\beta}, \tag{7-1}$$

   that is, we establish a relation between the variations of wages and rents, $\hat{w}_i = \hat{Q}_i^{1-\frac{1}{\beta}}$.

2. Population distribution

   Then, from the population distribution equation (3-3), substituting in the above relation, and with exogenous amenities such that $\hat{B}_n = 1$,

   $$\hat{\lambda}_{ni} = \left(\frac{\hat{w}_i}{\hat{Q}_n^{1-\alpha}}\right)^{\varepsilon} = \left(\frac{\hat{Q}_i^{1-\frac{1}{\beta}}}{\hat{Q}_n^{1-\alpha}}\right)^{\varepsilon} \tag{7-2}$$

3. Outside utility

   From the equation relating Φ to the utility of the outside economy, we know that $\hat{\Phi} = 1$. Imposing this condition is functionally equivalent to requiring that the final population distribution remains normalized,

   $$\hat{\Phi} = \frac{\sum_{ni} \hat{\lambda}_{ni} \Phi_{ni}}{\Phi} = \sum_{ni} \hat{\lambda}_{ni} \lambda_{ni} = 1. \tag{7-3}$$

   Note that, like each $\hat{\lambda}_{ni}$, this equation is homogeneous of degree $\varepsilon\,(\alpha - 1/\beta)$ in the $\hat{Q}_n$. Therefore, this equation is functionally a normalization condition on the rent changes.

4. Demand for floorspace

   We can then write the variation of the expenditure on floorspace in terms of our variations in rents and total city population (together with base period quantities) as

$$\hat{\mathbb{Q}}_n \mathbb{Q}_n = \hat{L}\, L \left[(1-\alpha) \sum_i \hat{w}_i w_i \hat{\lambda}_{ni} \lambda_{ni} + \frac{1-\beta}{\beta} \hat{w}_n w_n \sum_m \hat{\lambda}_{mn} \lambda_{mn}. \right] \tag{7-4}$$

   This gives us the demand side of our final equilibrium condition.

5. Supply of floorspace

   Separating in the three case that are interesting for our analysis,

   – If a region remains formal,

$$\hat{H}_n = \hat{Q}_n^{\mu} \qquad \Longrightarrow \qquad \hat{\mathbb{Q}}_n = \hat{Q}_n^{1+\mu} \tag{7-5}$$

   – If it becomes formal,

$$\hat{H}_n = \frac{h_F}{h_I} \hat{Q}_n^{\mu} Q_n^{\mu} \qquad \Longrightarrow \qquad \hat{\mathbb{Q}}_n = \left(\frac{h_F Q_n^{\mu}}{h_I} \right) \hat{Q}_n^{1+\mu} \tag{7-6}$$

   – If it remains informal

$$\hat{H}_n = 1 \qquad \Longrightarrow \qquad \hat{\mathbb{Q}}_n = \hat{Q}_n \tag{7-7}$$

   This gives us the supply side of the equilibrium condition.

Altogether, by imposing market clearing by equating the supply and demand equations above, we have a set of N equations for N independent variables: N rent changes, one of which is a function of the others by the normalization condition, and the population change.

I solve this by an iterative numerical procedure, which I will briefly describe. First, we start with a guess, $\hat{Q}_n^0$, which we force to obey the normalization condition given by (7-3) by otherwise dividing it by the would-be value of $\hat{\Phi}^{\frac{1}{\varepsilon(\alpha-1/\beta)}}$. Then we calculate the associated value of $\hat{L}$ that makes the floorspace market clearing condition hold on average, as that must also be a property of any full solution. Now, since the average value of the equation is zero, unless we are in the full correct solution there must be some n for which the demand side term is greater and some for which the supply one is. I adjust my guess first by

$$\hat{Q}_n^1 = \left(\frac{1}{2} + \frac{1}{2} \frac{\text{demand term}}{\text{supply term}} \right) \hat{Q}_n^0. \tag{7-8}$$

The simple intuition is that if, at a given rent price, demand would outpace supply, then the price should be adjusted upward to make markets clear again, and similarly for the opposite case. Note that by the previous observation at least some prices get adjusted upwards and some downwards. Finally, I reimpose the normalization condition on this adjusted guess, which might work against some of the adjustments, but not the most severely needed at this step. By iteratively applying this procedure to update a guess, calculating the sum of squares of the residual value of the equation at each step, I quickly find a solution within numerical tolerance. While I do not yet prove that it works in general, the associated intuition and the fact that it does work in reaching solutions in our case is enough to justify its presentation at this point.

7.2 Labor market access without informal housing

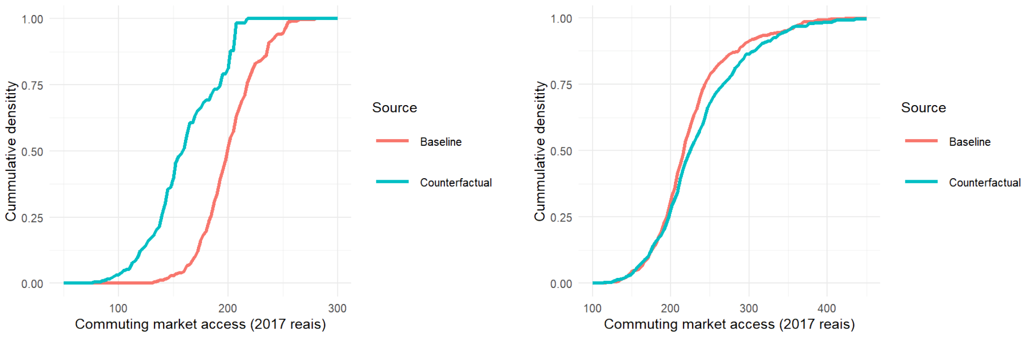
I proceed to use the above method to obtain the counterfactual where all regions of informal housing are instead converted into formal housing ones, focusing on the matters of labor market access and spatial spillovers that are our final objects of study.

Our main result concerns the distribution of labor market access, which we plot in figure 7.1, along with the baseline one for comparison. First, focus on the results for the previously informal regions, formalized in the counterfactual, presented in subfigure 7.1(a). We see that labor market access for the population of these regions is greatly decreased. This quantity goes from a mean value of 200 reais to just over 140, dropping by about 30%. The same happens to the median as well. This helps us see that indeed the informal sector is fulfilling a role in providing its population with increased access to the city's labor market.

The same cannot be said, though, of the formal sector, with results presented in subfigure 7.1(a). While the gains are modest, with a change of the mean value from 226 to 236, or about 4%, the population of these regions does see a slight increase in labor market access.

We can understand the mechanisms for these changes as follows: close to the city centre, there a few important regions where land is being used inefficiently. Without the formal sector technology, construction does not respond elastically to the high desirability of these areas, leading to lower than ideal supply of floorspace and higher rents. Substituting this for the formal sector leads to the correction of these problems there. This in turn increases the wages payed and number of workers employed by the productive sector in these regions. As they are central and accessible to a large share of the city's population, these small spillovers in labor market access compound to large gains for the population of the formal sector regions as a whole.

Meanwhile, most of the population of these informal settlements does



7.1(a): Originally informal

7.1(b): Originally formal

Figure 7.1: Baseline and counterfactual populational distribution of commuting market access, for regions with originally formal and informal classification.

not live in one of these few central areas. Rather, most informal settlements are at least a moderate distance from the city centre, with moderate local housing demand. The change forces the substitution of the cheap, informal housing technology with an expensive formal one that increases rents and depresses wages locally. As these regions are not as intensely affected by the gains of the inner city core, their remaining populations see a decrease in labor market access. This phenomenon is illustrated by the map in figure 7.2, of the counterfactual wages. We can also see that the change in wages directly in the formal sector areas is modest.

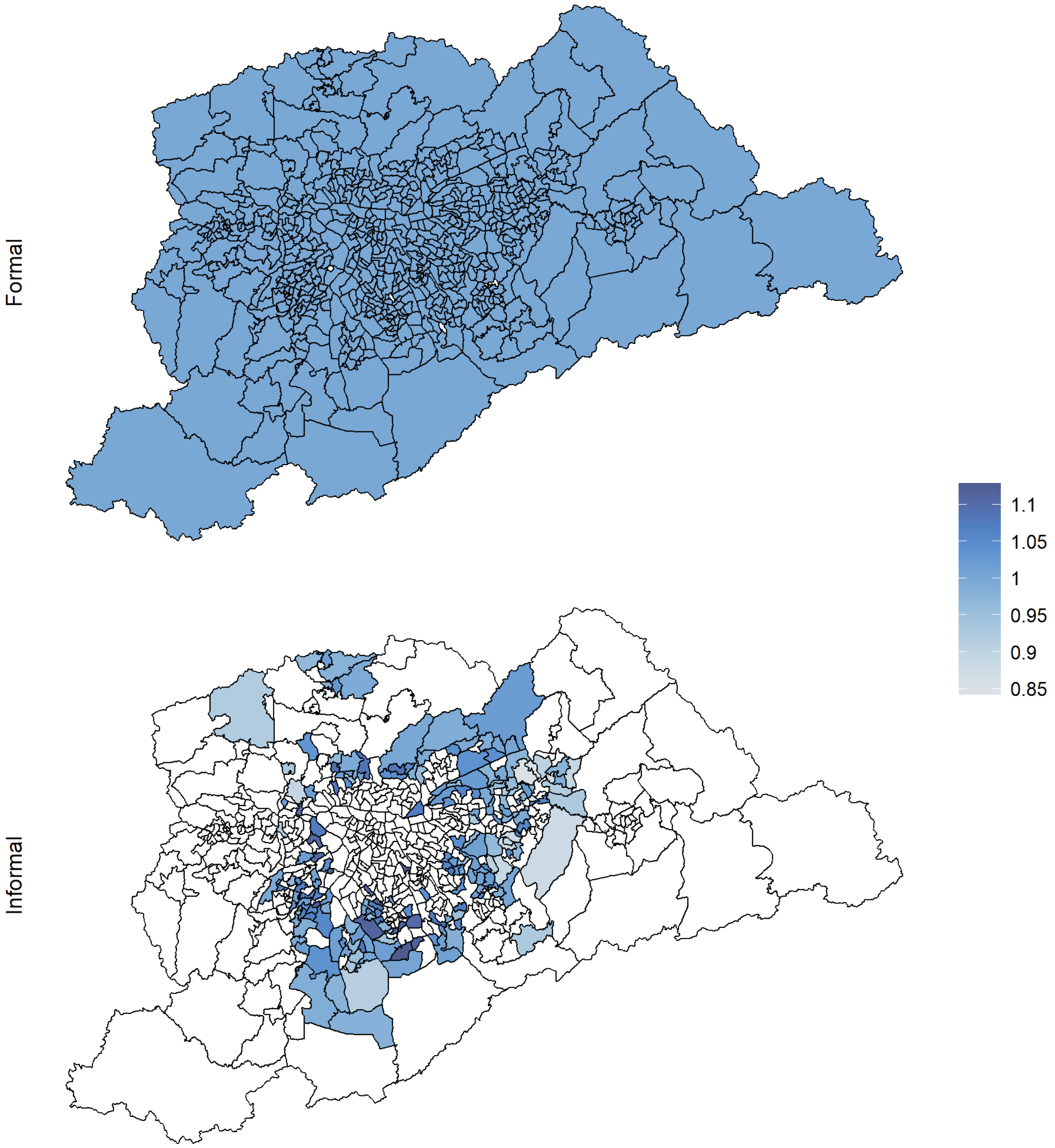


Figure 7.2: Proportional change in counterfactual wages in regions, by original classification.

8
Discussion

We have seen how the theoretical framework of quantitative urban models can be extended to incorporate informal housing, and also the importance of proper area scaling in its interpretability and internal consistency. We have also seen how in that framework we can estimate the model and recover local variables from few observed quantities. About the commuting costs part of the model and the gravity equation, we have seen how the Poisson pseudo-maximum likelihood estimator outperforms the usual ordinary least squares estimator on sparse commuting data.

I also presented compelling evidence that the estimated model captures well the characteristics of the São Paulo metropolitan area, by showing comparisons of the distributions of commuting times and wages from data and the model, and of the predicted supply of floorspace with the observations coming from the 3D scanning of the city at each location. In the counterfactual section, we get to leverage this estimated model to evaluate the effects of the informal housing sector on the city. As in many situations related to the urban environment, we see a situation of concentrated gains (from their existence) and dispersed losses.

The informal sector does seem to perform a role of providing its inhabitants with cheaper housing and access to the city's labor market. It is particularly well-suited at doing so at locations far from the city centre, where its advantage in building cheaper are put to best use. Meanwhile, near the city centre, our estimate suggests that the institutional frictions that prevent the conversion of areas of informal housing to formal have sizable negative impacts on the rest of the city. While this inefficiency suggests there exists a role for policy in alleviating these frictions, the counterfactual estimate cautions us that it is important to consider the possible distributive consequences of any such intervention. At the very least, good regional targeting seems to be required for maximum benefits to be attained. These concerns are even more important when considering that populations of these areas are already usually poorer and less politically influential.

It still important to discuss some of the limitations of the approach taken this work, which also point towards important directions in which this research agenda can evolve in the future. First, the only a priori heterogeneity between agents is in preferences between location pairs: differences in education levels, formality of employment, or other sociodemographic dimensions that could be relevant are not considered. Then, the informal construction sector is assumed to have no margin of response to increased demand. While it is certainly less responsive than the formal one, it is not fully so, and incorporating the denser coverage margin emphasized by Henderson et al. (2021) would be an important improvement. In particular, it would make some of the obtained results less mechanical, in the sense of reliant on the modelling assumptions about the informal sector, and more empirically grounded. Also in that direction, another important goal is to model endogenously the choice between the formal and informal sector sector technologies, including the conversion frictions in such a way that they can be quantified. Finally, the recent urban economics literature, and especially the quantitative urban models one, strongly emphasizes the role of agglomeration economies in the making of cities as they are. Ultimately, we would like to understand how these differ between the two sectors and include that feature as well.

Both by eventually incorporating some or all of these proposed improvements, and by jointly considering evidence coming from work with complementary advantages (and limitations), it is my hope and belief that this work could be used to inform policymaking about this enormously relevant topic.

Bibliography

UN-HABITAT. **World Cities Report 2020: The Value of Sustainable Urbanization**. United Nations Human Settlement Programme (UN-Habitat), 2020.

GLAESER, E.; HENDERSON, J. V.. **Urban economics for the developing world: An introduction**. Journal of Urban Economics, 98:1–5, 2017.

BRYAN, G.; GLAESER, E. ; TSIVANIDIS, N.. **Cities in the developing world**. Annual Review of Economics, 12:273–297, 2020.

CAVALCANTI FERREIRA, P.; MONGE-NARANJO, A. ; TORRES DE MELLO PEREIRA, L.. **Of cities and slums**. Federal Reserve Bank of St. Louis Working Paper Series, 2016.

AHLFELDT, G. M.; REDDING, S. J.; STURM, D. M. ; WOLF, N.. **The economics of density: Evidence from the berlin wall**. Econometrica, 83(6):2127–2189, 2015.

REDDING, S. J.; ROSSI-HANSBERG, E.. **Quantitative spatial economics**. Annual Review of Economics, 9:21–58, 2017.

HENDERSON, J. V.; REGAN, T. ; VENABLES, A. J.. **Building the city: from slums to a modern metropolis**. The Review of Economic Studies, 88(3):1157–1192, 2021.

GECHTER, M.; TSIVANIDIS, N.. **Spatial spillovers from urban renewal: evidence from the mumbai mills redevelopment**. Penn State University (mimeo), 2020.

STURM, D.; VENABLES, A. J. ; TAKEDA, K.. **Applying the quantitative urban model to cities in developing countries**. Working paper, 2021.

ZANAGA, D.; VAN DE KERCHOVE, R.; DE KEERSMAECKER, W.; SOUVERIJNS, N.; BROCKMANN, C.; QUAST, R.; WEVERS, J.; GROSU, A.; PACCINI, A.; VERGNAUD, S.; CARTUS, O.; SANTORO, M.; FRITZ, S.; GEORGIEVA, I.; LESIV, M.; CARTER, S.; HEROLD, M.; LI, L.; TSENDBAZAR, N.; RAMOINO, F. ; ARINO, O.. **ESA WorldCover 10 m 2020 v100**. Zenodo, 2021.

PEREIRA, R. H. M.; SARAIVA, M.; HERSZENHUT, D.; BRAGA, C. K. V. ; CONWAY, M. W.. **r5r: Rapid realistic routing on multimodal transport networks with $\mathbf{R^5}$ in r**. Transport Findings, Mar. 2021.

HEBLICH, S.; REDDING, S. J. ; STURM, D. M.. **The making of the modern metropolis: evidence from london**. The Quarterly Journal of Economics, 135(4):2059–2133, 2020.

EATON, J.; KORTUM, S.. **Technology, geography, and trade**. Econometrica, 70(5):1741–1779, 2002.

SILVA, J. S.; TENREYRO, S.. **The log of gravity**. The Review of Economics and statistics, 88(4):641–658, 2006.

SILVA, J. S.; TENREYRO, S.. **Further simulation evidence on the performance of the poisson pseudo-maximum likelihood estimator**. Economics Letters, 112(2):220–222, 2011.

A
Appendix

A.1
Population distribution

From our indirect utility equation (3-2), the distribution of utility is given by

$$\Pr\left[V_{nio} \leq V\right] = \Pr\left[z_{nio} \leq \frac{\kappa_{ni} Q_n^{1-\alpha}}{B_n w_i} V\right] = \exp(-\Phi_{ni} V^{-\varepsilon}). \tag{A-1}$$

As we consider a continuum of workers in the city, the fraction of workers in a particular residence-workplace pair is the same as the individual probability of a workers choosing this pair. As the preference shocks for each pair are independent, using the expression above, we can calculate this probability as

$$\begin{aligned}
\lambda_{ni} &= \Pr\left[V_{nio} \geq \max_{m,j \neq n,i} V_{mjo}\right] \overset{\text{(LIE)}}{=} \mathbb{E}\left(\Pr\left[V_{nio} \geq \max_{m,j \neq n,i} V_{mjo} \,\middle|\, V_{nio}\right]\right) \qquad \text{(A-2)} \\
&\overset{\text{(iid)}}{=} \mathbb{E}\left(\prod_{m \neq n, j \neq i} \Pr\left[\, V_{mj} \leq V_{ni} \,|\, V_{ni} \,\right]\right) \overset{\text{(A-1)}}{=} \mathbb{E}\left(\prod_{m \neq n, j \neq i} e^{-\Phi_{mj} V_{ni}^{-\varepsilon}}\right) \\
&\overset{\text{(A-1)}}{=} \int_0^\infty \left(\Phi_{ni}\, \varepsilon\, V_{ni}^{-\varepsilon - 1}\right) \prod_{m,j} e^{-\Phi_{mj} V_{ni}^{-\varepsilon}} \; \mathrm{d}V_{ni} \\
&= \Phi_{ni} \int_0^\infty e^{-\left(\sum\limits_{m,j} \Phi_{mj}\right) V^{-\varepsilon}} \; \mathrm{d}(V^{-\varepsilon}) = \frac{\Phi_{ni}}{\sum\limits_{m,j} \Phi_{mj}}.
\end{aligned}$$

A.2
Expected utility

With a calculation essentially similar to the first steps of the one above, we have that

$$\Pr\left[\max_{n,i} V_{nio} \leq V\right] = \prod_{n,i} e^{-\Phi_{n,i} V^{-\varepsilon}} = e^{-\Phi V^{-\varepsilon}}. \tag{A-3}$$

Given this, the expected value of V is

$$\begin{aligned}\overline{V} &= \int\limits_0^\infty V \, \frac{\mathrm{d}}{\mathrm{d}V}\left(e^{-\Phi V^{-\varepsilon}}\right) \, \mathrm{d}\,V = \Phi^{\frac{1}{\varepsilon}} \int\limits_0^\infty (\Phi V^{-\varepsilon})^{-\frac{1}{\varepsilon}} \, e^{-\Phi V^{-\varepsilon}} \, \frac{\mathrm{d}}{\mathrm{d}V}\left(-\Phi V^{-\varepsilon}\right) \, \mathrm{d}V = \\ &\overset{y=\Phi V^{-\varepsilon}}{=} \Phi^{\frac{1}{\varepsilon}} \int\limits_0^\infty y^{-\frac{1}{\varepsilon}} \, e^{-y} \, \mathrm{d}y = \Phi^{\frac{1}{\varepsilon}} = \Phi^{\frac{1}{\varepsilon}} \, \Gamma\left(1 - \frac{1}{\varepsilon}\right), \qquad \text{(A-4)}\end{aligned}$$

where in the last step we just used the integral definition of the Γ function.