Commuting and Productivity: Quantifying Urban Economic Activity using Cell Phone Data

Gabriel Kreindler and Yuhei Miyauchi

March, 2015

LIRNE*asia*
info@lirneasia.net | www.lirneasia.net

LIRNE*asia* is a pro-poor, pro-market think tank whose mission is *Catalyzing policy change through research to improve people's lives in the emerging Asia Pacific by facilitating their use of hard and soft infrastructures through the use of knowledge, information and technology.*

Contact: 12 Balcombe Place, Colombo 00800, Sri Lanka. +94 11 267 1160.
info@lirneasia.net
www.lirneasia.net

This work was carried out with the aid of a grant from the International Development Research Centre (IDRC), Canada and the Department for International Development (DFID), UK.

Commuting and Productivity: Quantifying Urban Economic Activity using Cell Phone Data[1]

Gabriel Kreindler and Yuhei Miyauchi[2]

1 Introduction

Human mobility and economic activity are intertwined. International migration flows are responsive to economic activity in the destination country, as well as to the distance between the origin and destination countries (Mayda, 2010; Peri, 2012). The elasticity of migration with respect to economic activity is generally of similar magnitude, or smaller, than the elasticity with respect to distance, which points to a tradeoff between distance and higher economic opportunity. The responses of regional and seasonal migration to economic opportunity and economic shocks have similar features (Blanchard & Katz, 1992; Hare, 1999).

These results suggest that how people move and the intensity of economic activity at different locations are related, even if we zoom in further in space (i.e. within cities) and time (i.e. at daily level). In particular, one may reasonably suspect that this relationship is stronger for commuting within urban areas, because commuting costs recur daily, while migration costs are one-off or occasional. However, to date this relationship has not been studied quantitatively.

In this project, we use a revealed preference approach to study the link between commuting flows and urban economic activity, and quantify this relationship using cell phone data. First, we set up a theoretical model (following (Ahlfeldt, Redding, Sturm, & Wolf, 2014)) that links individual home- and job-choice decisions, commuting flows and economic productivity, and show that it can be implemented empirically through a gravity model. Second, we use cell

[1] Acknowledgements: The authors are grateful to the LIRNEasia organization for providing access to cell phone data and an excellent working environment, and especially to Sriganesh Lokanathan, research manager at LIRNEasia, whose dedication and relentless efforts made this project possible. We acknowledge funding from the International Development Research Centre (IDRC). We sincerely thank Alex Bartik, Arnaud Costinot, Seema Jayachandran, Sriganesh Lokanathan, Danaja Maldeniya, Ben Olken, the LIRNEasia BD4D team (Dedunu Dhananjaya, Kaushalya Madhawa, and Nisansa de Silva), and seminar participants at MIT and at LIRNEasia for constructive comments and feedback. We also thank Laleema Senanayake and Thushan Dodanwala for assistance processing GIS and census data.
[2] Contact: gek@mit.edu and miyauchi@mit.edu.

phone data from Sri Lanka to construct a measure of commuting with fine spatial and temporal variation. Third, we estimate to model using the commuting data, and then calculate a measure of productivity at each location; we also use the estimated model to construct a measure of output (economic activity), and a measure of residential income, both at each location. Finally, we validate the model's predictive power. Specifically, we assess the correlation between the residential income measure computed from the model and a separate measure of residential income, given by a new and precise data source of nighttime lights.

Our use of cell phone data to measure mobility follows a recent literature that has argued that this type of passively collected "big data" is a valuable resource, especially in countries and contexts where traditional data collection is less developed. This literature has also tried to quantify and address the biases inherent in cell phone data. Our focus in this paper is on (loosely defined) urban areas,[3] for two reasons. First, cell tower density is significantly higher in urban, developed areas, which gives us higher resolution. Secondly, commuting and our theoretical setup are most relevant in an urban context. Using commuting flows derived from cell phone data leads to a unique tradeoff. The very large sample size and temporal variation afforded by this approach is a significant improvement over survey-based data; however, this comes at the cost of loss of precision in terms of detailed data, for example in terms of being able to zoom in on a certain demographic group (e.g. male prime-aged workers).

Ideally, we would like to validate the productivity and output measures derived from the estimated model, using independent data with a high geographic and temporal granularity. Unfortunately, currently this type of data is only available at much higher levels of geographic and temporal aggregation in our setting. Instead, we report a validation exercise using nighttime light data captured from satellites. Nighttime light data is a good measure of residential income (Chen & Nordhaus, 2011; Henderson, Storeygard, & Weil, 2010; Mellander, Stolarick, Matheson, & Lobo, 2013), which we use in an indirect test of the model. Currently, our validation exercise is restricted to geographic variation.

There are several directions in which we are planning to advance this work in the future. For example, we hope to further exploit the time variation available in the commuting data, in order to look at changes in the spatial distribution of economic activity over short and medium time horizons. We can investigate the effects of short term disturbances, such as the transportation restrictions imposed by national events. We can also study whether, in the medium term, economic activity becomes more concentrated around city centers, or whether it actually moves towards city outskirts. Another direction is obtaining data that would allow a more direct validation of the model. More detailed data, such as census statistics on education attainment within small administrative units can be used to calibrate a version of

[3] We do not limit our analysis to urban areas as defined by administrative boundaries, because commuting flows may often traverse these boundaries.

the model that allows skill heterogeneity. This would allow us to explore how the commuting patterns of workers of different skill levels differ.

We hope that our project will contribute to the understanding of urban structure, specifically on the relationship between productivity, commuting costs and work choices. From a practical perspective, our method has the potential to deliver an indicator of economic productivity with high spatial and temporal resolution. Such detailed data can help improve economic policy and urban planning in many dimensions. First, our method can be applied to investigate medium term suburbanization patterns. In the Colombo area, between 2001 and 2012, there has been a shift in residential population away from the central areas towards the suburbs (Ministry of Transport, 2013). Quantifying the concurrent shifts in residential and working population, and the associated changes in commuting patterns, is of great importance for transportation planning. Second, our method can be used to evaluate the economic impact of policies or natural events that change mobility patterns. Examples include fuel price or subsidy changes, road closures (such as the *Oborodh* in Bangladesh in January 2015), and travel restrictions (such as the *cordon sanitaire* in West Africa due to the Ebola outbreak). The impacts of such events on commuting and economic activity are generally only superficially understood, and cell phone data and our method would help provide more detailed evidence. Third, our method will serve as a first step toward a platform to monitor urban economic activity in real or near-real time.[4] This step would be most useful for policy-makers, because it holds the promise of switching from historical studies (which generate knowledge) to data on events as they unfold (which enables immediate response).

The organization of the paper is as follows. In section 2 we briefly describe the argument (model, data and empirics) in our paper. Section 3 reviews the relevant literature. Section 4 sets up the model, derives the gravity equation and measures of interest, and discusses alternative model assumptions and extensions. The data sources and the estimation strategy are described in section 5. Section 6 reports a validation exercise using nighttime light data, and section 8 concludes. Some model derivations, data and empirical method details are relegated to appendices.

2 The Argument at a Glance

2.1 The Job Choice Model and Gravity Equation

The central piece of our theoretical model, which relies greatly on (Ahlfeldt et al., 2014), is the job choice tradeoff between high job wage (or job quality) and commuting distance. We believe this is, at least implicitly, a first order feature of the job choice process.

[4] This point is closely related to recent literature on real-time urban monitoring (Calabrese, Colonna, et al., 2011; Calabrese et al., 2014; P. Wang et al., 2012). Of course, the actual implementation of such a real-time monitoring system requires the development of technology and infrastructure that is far beyond the scope of this paper.

Formally, a worker ω who lives at location i evaluates a potential work location j according to the income $y_{ij\omega}$ she would receive by working at j. Assuming that each worker inelastically supplies a unit measure of labor, income is given by the formula

$$y_{ij\omega} = \frac{w_j z_{ij\omega}}{d_{ij}}, \tag{1}$$

where w_j denotes the wage at work location j, d_{ij} is a measure of the distance between i and j, and $z_{ij\omega}$ is an idiosyncratic shock that is specific to the worker ω and the origin-destination pair (i, j). The $z_{ij\omega}$ shocks ensure that there is a positive probability that workers commute between any i and j. We assume, following (Ahlfeldt et al., 2014; Eaton & Kortum, 2002), that $z_{ij\omega}$ follows a Fréchet distribution with shape parameter ϵ; we will later show that the scale parameter can be normalized to 1.

Equation (1) implies that the probability π_{ij} that a worker residing in origin i commutes to destination j is given by

$$\pi_{ij} = \frac{\left(w_j / d_{ij}\right)^{\epsilon}}{\sum_s (w_s / d_{is})^{\epsilon}}. \tag{2}$$

This identity describes a gravity equation.

Before exploring it further, we pause to discuss alternative assumptions that can be accommodated by the framework described above. We have implicitly assumed that workers choose their home and work locations sequentially. (Ahlfeldt et al., 2014) show that if agents choose their home and work location *simultaneously*, and the idiosyncratic shocks are realized at the level of origin-destination pairs (i, j), equation (2) continues to hold. Intuitively, this is because π_{ij} is the probability to commute to j, *conditional* on having chosen home location i. Another issue is that equation (1) implies that the only aggregate shifter of the income attainable at destination j is distance. In reality, there may be other, possibly origin-destination specific factors that are important. In section 4.1 and in the empirical application, we incorporate aggregate origin-destination specific heterogeneity; to gain empirical traction over estimating this, we assume it varies smoothly in space, which allows us to estimate it using local weighted smoothing techniques described in section 5.3. Finally, note that equation (1) implies that distance and the shock z_{ij} affect income (through labor supply or productivity). An alternative assumption is that they only affect the agent's choice utility; we plan to use the model to explore this issue empirically in future work.

To implement equation (2) empirically, we work with the logarithmic version

$$\log(\pi_{ij}) = \psi_j + \beta \log(d_{ij}) - \mu_i + \varepsilon_{ij}, \tag{3}$$

where ψ_j is a log transformation of the wage at j, μ_i captures origin-specific factors, and ε_{ij} is measurement error. Intuitively, ψ_j captures a destination's attractiveness, after controlling for distances to all origin locations, and their respective sizes. Equation (3) is a type of gravity equation, which is an empirical model that has been used extensively in transportation economic, international trade and migration (Anderson, 1979; Duran-fernandez & Santos, 2014; Erlander & Stewart, 1990; Sohn, 2005).

The benefit of using an explicit model of workers' decisions is that it guides us on how to compute the economic measures we are interested in, such as total output and mean residential income at a particular location.

2.2 Data and Estimation

We map the model to the real world using commuting flows extracted from cell phone data. The Call Detail Record (CDR) data contains information on cell phone transactions in Sri Lanka for a period of over a year. For our purposes, each transaction contains a timestamp, user identifiers and geographical identifiers (at the cell phone tower level). We identify towers (or their Voronoi cell) with locations from the model. Section 5.1 discusses the data in more detail.

We use a simple algorithm to construct daily commuting flows from CDR data, building on the literature on this topic (Calabrese, Di Lorenzo, Liu, & Ratti, 2011; P. Wang, Hunter, Bayen, Schechtner, & González, 2012). Essentially, for each user and each day with available data, we construct a commuting trip with origin given by the first location available in the 5:00 to 10:00 time interval, and the destination by the last transaction in the 10:00 to 15:00 interval.[5]

We aggregate commuting trips over the entire time period to commuting flows between pairs of an origin and a destination. In Figure 1 and Figure 2 we offer a first visual analysis of the properties of commuting flows. Figure 1 shows how the (log) volume of commuting trips between two locations varies as a function of the log distance between them. The relationship is linear and decreasing over the entire interval. Figure 2 shows an example of commuting flows for a fixed origin location (indicated by the green polygon) in the greater metropolitan Colombo area. The two panels show the raw commuting flows estimated from cell phone data, and the smoothed counterparts, using the method described in section 5.3. Here too we notice that commuting intensity and distance are inversely related. However, the figure also shows that there is significant variation in commuting flows even after accounting for distance. Furthermore, this variation appears related to the spatial distribution of activity in

[5] We choose these time intervals to approximate the pre- and post-commuting times in Sri Lanka. Clearly, if a significant share of commuting and economic activity takes place at different times, this method will be biased. We do not believe this is a major concern for two reasons. First, the daily aggregate rhythm of cell phone users is similar to those in developed countries (not shown). Secondly, our results are robust to using a completely different method to classify commuting trips (see section B.1).

the Colombo area. For example, the region in the top right, which is the destination for a high number of commuting trips (relative to surrounding areas) is an Export Processing Zone (EPZ). It is this meaningful variation that is captured by the destination fixed effects in the gravity model.

We estimate equation (3) on commuting flows between pairs of towers in Sri Lanka. Our sample covers several hundred million commuting trips from 262 non-holiday weekdays from a year's worth of data. We use a linear regression model with two sets of fixed effects (corresponding to origin and destination locations). Finally, we use weighted local smoothing (with variable kernel bandwidth) to estimate origin-destination specific factors that are smooth in space (these factors were not explicitly modelled above but are part of the model in section 4).

The final step of the estimation procedure is to use the estimated fixed effects and coefficients to construct two measures that are important for our analysis. The first measure is total output X_j at a (destination) location j, which measures the output created by agents who commute to (and thus work at) j. The second is residential mean income Φ_i at an (origin) location i. Intuitively, Φ_i captures the average wage brought "home" by workers who live at i and who work in various other destinations. (In practice, we compute both measures for all towers.)

2.3 The Validation Exercise – Comparison with Nighttime Lights

Having estimated economic productivity (wages) and economic activity (output), exclusively based on mobility flows derived from cell phone data, we would ideally proceed to validate these measure using independent wage and output measures. Unfortunately, in our context this type of data is only available aggregated at province level.

Instead, we present results from a validation exercise using the mean residential income measure Φ_i. We want to find out whether it contains information beyond simpler statistics derived from cell phone data (such as the local tower density), and to achieve this we compare it with nighttime lights, a recognized measure of residential income (Chen & Nordhaus, 2011; Henderson et al., 2010; Mellander et al., 2013). We use a new version of night lights, VIIRS, curated by the Earth Observatory Group (EOG) at National Oceanic and Atmospheric Administration's (NOAA). The VIIRS data has higher spatial and temporal resolution than the previously used data, and it does not have a saturation point. The last point is important, given that we focus on urban areas where the previously used nighttime light data is often censored at its highest value.

Figure 3 shows a graphical comparison: nighttime light VIIRS in the top panel, and the $\log(\Phi_i)$ measure (at the tower cell level) in the bottom panel. The correspondence is generally good, yet there are clear points where the model can be improved (e.g. patterns along roads). In Table we show using linear regression analysis that the income measure is informative

regarding nightlights after controlling for population density (interpolated from the census), tower cell size, and various other simple indicators derived from cell phone data.

3 Literature Review

The relationship between mobility and economic activity is an implicit feature of classic urban economic models. The classic monocentric urban model assumes all production is concentrated in the central business district (CBD), whither all workers commute (Mills, 1967). Later models allow production to occur at various locations, potentially leading to mixed land use patterns (Fujita & Ogawa, 1982; Lucas & Rossi-Hansberg, 2002; Wheaton, 2004). However, it is difficult to take these models to the commuting data, for several reasons. To achieve tractability, they assume space is one-dimensional or impose radial symmetry; also, commuting patterns are deterministic, with all workers who live in a given location commuting to the same destination. In these models all agents have identical preferences, which implies that in equilibrium there is no cross-commuting, which leads the model to significantly understate the total level of commuting (Anas et al 1998 p 1444).[6]

(Ahlfeldt et al., 2014) introduce an urban economic model which predicts more realistic mobility patterns. The agents' joint choice of home and work locations takes center stage in their model. Agents care about residential housing rents and amenities, wages and commuting time; crucially, random idiosyncratic factors break the determinacy present in previous model and ensure positive commuting flows for every pair of origin and destination. This implies that commuting flows are described by a gravity equation.[7] They show that if amenities and underlying productivity are exogenously given, there is a unique equilibrium in terms of residential and work assignments, wages and rents. They then prove that, fixing the fundamental model parameters, there is a one-to-one mapping between the vector of residential and work populations, on one hand, and the vector of wages and rents, on the other. In other words, conditional on model parameters, commuting *flows* are not necessary to identify wages and rents; residential and employment populations suffice. (Ahlfeldt et al., 2014) then go on to estimate the fundamental parameters from long-term temporal variation introduced by the existence of the Berlin Wall.

In this paper, we focus on estimating productivity from commuting flows, and hence take as our starting point the subset of the model in (Ahlfeldt et al., 2014) that studies work location choice. The advantage of using the commuting flows is that this allows us to control for a non-

[6] Spatial distribution aside, there is clear evidence that job seekers trade off wages and commuting costs (Ommeren & Fosgerau, 2009; van Ommeren, J. van den Berg, & Gorter, 2000).

[7] (Anas, 1983) previously showed that discrete choice modeling with random utility a la (McFadden, 1973) leads to a gravity equation for commuting flows. Furthermore, if agents choose destinations (origin-destination pairs), the resulting gravity equation is unconstrained (constrained). Compared to (Ahlfeldt et al., 2014), (Anas, 1983) does not give the precise interpretation of commuting choice as the tradeoff between higher wages and the higher cost of commuting to more distant working places (including monetary, time and pure utility cost of commuting).

parametric smooth function of origin-destination pairs. (Ahlfeldt et al., 2014) also run a gravity equation (our Equation (3), which does not include the smooth origin-destination function) using microdata on commuting, aggregated into 144 flows between pairs of districts in Berlin. In this report, we solely report the results obtained from the gravity model with the smooth origin-destination function. We leave a more thorough contrast between these two methods for future research.

Separately from the urban economics literature, empirical studies in regional and transportation economics have developed and estimated gravity models of travel patterns.[8] This literature has tried to give micro-foundations of gravity models. There are mainly two approaches: one is to assume an entropy-maximizing principle (Wilson, 1967), while the other is to assume utility-maximizing behavior (Anas, 1983; McFadden, 1973; Niedercorn & Bechdolt, 1969). In particular, (Anas, 1983) shows that discrete choice modeling with random utility a la (McFadden, 1973) leads to the gravity equation of commuting. Although this literature does not interpret the estimated destination attractiveness as a wage, it should be noted that our empirical approach closely follows this literature.

Research using cell phone datasets has grown steadily since these datasets have become available. Here, we briefly summarize the literatures that are directly relevant to our analysis.

Firstly, researchers have explored the use of cell phone data for understanding mobility patterns (Blumenstock, 2012; Csáji et al., 2013; Deville et al., 2014; Simini, González, Maritan, & Barabási, 2012; A. P. Wesolowski & Eagle, n.d.; A. Wesolowski et al., 2012). (Blumenstock, 2012) uses cell phone data from Rwanda to infer patterns of internal and seasonal migration. (Csáji et al., 2013) use cell phone data from Portugal to explore users' mobility from several angles. Relevant for this paper, they categorize Home and Work locations based on the frequency with which users connect to towers. For users for whom they identify a unique Home and a unique Work location, they find that the distribution of the distance travelled between Home and Work is approximatively log-normal distribution. (See Figure 1 on page 31 for our version of this figure).

Secondly, some papers have developed methods to construct commuting flows (or more general origin-destination matrices) from cell phone location data (Calabrese, Di Lorenzo, et al., 2011; Iqbal, Choudhury, Wang, & González, 2014; P. Wang et al., 2012). These papers cover a variety of countries, contexts and methods, which we believe lends support to the corresponding step of constructing commuting flows in our paper.

Lastly, there is a related literature that studies how important urban indicators can be measured with cell phone data (Calabrese, Colonna, Lovisolo, Parata, & Ratti, 2011; Calabrese, Ferrari, & Blondel, 2014; Reades, Calabrese, Sevtsuk, & Ratti, 2007). This literature puts

[8] Recent examples include (Duran-fernandez & Santos, 2014; McArthur, Kleppe, Thorsen, & Ubøe, 2011; F. Wang, 2001).

emphasis on the high-frequency and real-time nature of cell phone data, which is also the strength of our method.

4 A Model of Mobility, Work Choice and Economic Activity

We now set up a model that links mobility flows to the spatial distribution of productivity. In our model, economic activity at a certain location in space can be decomposed into the number of workers who work at that location, and their average productivity. While worker commuting patterns are in principle directly observable from the cell phone data, productivity is not, and hence needs to be inferred.

The setup is as follows. Space is partitioned into a finite set L of locations, which may serve, at the same time, as residential locations and work locations. In our application, these locations will correspond to Voronoi cells of cell antenna towers. We refer to *residential* and *work location*, and denote them by i and j respectively, or, interchangeably, as *origin* and *destination* (of a commuting trip).

We assume workers are homogenous, and we focus on their work location choice problem, conditional on their home location. For ease of exposition, we do not explicitly model the home location choice (see (Ahlfeldt et al., 2014) for a complete model). Workers who live at a given location choose where to work based on the wages offered at different locations, the commuting distances involved, other aggregate bilateral factors, and idiosyncratic preference or productivity shocks. The idiosyncratic shocks leads to a gravity equation for the commuting probabilities. We derive formulas for the total output at a work location, and the mean income at a residential location, as functions of observable factors (distance) and quantities estimated from the gravity equation. (In the case of total output, an unestimted model parameter is also necessary.)

Our setup is very close to the model in (Ahlfeldt et al., 2014). The relation between the two models is discussed more in section 4.4.

4.1 The Worker's Job Choice Problem

We assume that any location $i \in L$ either has no workers who reside at i, or a continuum of workers. Each worker is endowed with a unit of labor, and supplies it inelastically. We focus on the work location choice faced by each worker. A worker ω residing at i can choose to work at any destination location $j \in L$ that offers employment. The income $y_{ij\omega}$ earned by ω if she chooses destination j is given by:

$$y_{ij\omega} = \frac{w_j k_{ij} z_{ij\omega}}{d_{ij}}.$$

In the above formula, w_j is the wage offered at location j; we thus assume that all firms operating at location j offer the same wage, and that all workers at j are paid the same.

In reality, for a variety of reasons, workers at j from different origins may have different productivities. For example, high skilled workers who work at j may come disproportionately from a certain origin i. This would show up in the data as a systematic departure from the model. Modelling this heterogeneity explicitly would take us beyond the scope of this paper; instead, we introduce the origin-destination-specific productivity factor k_{ij}. Note that there are as many k_{ij} terms as there are (potential) commuting flows. It follows that we cannot estimate all k_{ij}'s without any restriction. In practice, at the estimation stage we impose that k_{ij} varies smoothly as a function of i and j's geographic positions.

The distance between i and j is captured by the measure d_{ij}, which we assume takes the form $d_{ij} = {D_{ij}}^{\tau}$ where D_{ij} is the geodesic distance between i and j in kilometers.[9] Finally, the term $z_{ij\omega}$ is an idiosyncratic shifter of effective labor supply. In other words, worker ω from location i contemplating to work in location j has, for idiosyncratic reasons, effective labor supply equal to $z_{ij\omega}$. We assume that the random variable $z_{ij\omega}$ is i.i.d. following the Fréchet distribution, with scale parameter T and shape parameter ϵ.[10] Standard results on the Fréchet distribution (reviewed in Appendix A.1) imply that $y_{ij\omega}$ is also a Fréchet-distributed random variable, with shape ϵ and scale $T_{ij} = T w_j^{\epsilon} k_{ij}^{\epsilon} d_{ij}^{-\epsilon}$.

We assume that each worker chooses the work location j where $y_{ij\omega}$ is maximized. Then, the probability that a worker ω residing in i commutes to j is given by:[11]

$$\pi_{ij} = \frac{\left(\frac{w_j k_{ij}}{d_{ij}}\right)^{\epsilon}}{\sum_s \left(\frac{w_s k_{is}}{d_{is}}\right)^{\epsilon}}. \qquad (4)$$

In the absence of random shocks, all workers from i would choose the same work location j. The $z_{ij\omega}$ shocks leads to a non-degenerate distribution of work location choices, and the variance of $z_{ij\omega}$ controls the dispersion of the choice probabilities. In other words, a higher

[9] The log distance functional form seems to be a good fit for the data, as indicated in Figure 1. In practice, in order to avoid the problem of dividing by zero when $i = j$, we let D_{ij}^{τ} be the distance plus Δ, where $\Delta > 0$ is a small number; this practice is common when using the log functional form.

[10] The Fréchet distribution with shape parameter ϵ and scale parameter T has cumulative distribution function $\Pr(z_{ij\omega} \leq z) = e^{-Tz^{-\epsilon}}$.

[11] Note that the scale parameter T disappears from the expression. This implies that T is not identified just from the probabilities. This allows us to assume, without loss of generality, that $T = 1$.

variance of $z_{ij\omega}$ decreases the relative importance of distance, wage and bilateral productivity.

4.2 Mobility Flows follow a Gravity Equation

We now show that equation (4) describes a gravity equation for commuting probabilities. By taking the logarithm, we obtain the following relationship (recall $d_{ij} = {D_{ij}}^{\tau}$ where D_{ij} is distance in kilometers plus a small number Δ):

$$\log(\pi_{ij}) = \epsilon \log(w_j) + \epsilon \log(k_{ij}) - \epsilon\tau \log(D_{ij}) - \log\left(\sum_s \left(\frac{w_s k_{is}}{D_{is}^{\tau}}\right)^{\epsilon}\right)$$

We propose to estimate this equation through the following empirical equation:

$$\log(\pi_{ij}) = \psi_j + \kappa_{ij} + \beta \log(D_{ij}) - \mu_i + \varepsilon_{ij} \quad (5)$$

In equation (5), $\psi_j = \epsilon \log(w_j)$ is a destination fixed effect, κ_{ij} is a smooth non-parametric function of i and j's geographic locations,[12] $\mu_i = \log\left(\sum_s \left(\frac{w_s k_{is}}{D_{is}^{\tau}}\right)^{\epsilon}\right)$ is an origin fixed effect, $\beta = -\epsilon\tau$ and ε_{ij} is a random error term that accounts for measurement error.

Gravity models such as the one in (5) have been widely used as empirical models for international trade (Anderson, 1979), transportation (Erlander & Stewart, 1990) and commuting behavior (Duran-fernandez & Santos, 2014; Sohn, 2005). One distinction is that in (5) we allow (with some constraints) for pair-specific shifters κ_{ij}; this is not typically the case because most empirical gravity models focus on the effect of distance, whereas the objects of interest for us are the destination fixed effects ψ_j. Another distinction is that most applications of gravity models to transportation or commuting flows use some form of constrained gravity model. For example, the double constrained gravity model is guaranteed to match the aggregate outflows and inflows at each location; these models are mostly useful for estimating the effect of distance.

Note that equation (4) implies a constraint between the coefficients in (5). Namely, the origin fixed effects should satisfy the following relationship:

$$\mu_i = \log\left(\sum_s \exp(\psi_s + \kappa_{ij} + \beta \log(D_{is}))\right). \quad (6)$$

[12] In the empirical exercise, we implement a kernel smoothing estimator that depends on i and j. See equation (7) in section 5.3 for more details.

In our empirical application we will not incorporate this constraint in the estimation, mainly due to computational constraints. There are still some ways to efficiently impose such constraints upon estimation (such as MPEC; Mathematical Program with Equilibrium Constraints[13]). We leave this for future work.

Another assumption we make is $Cov\big(\kappa_{ij}, \log(D_{ij})\big) = 0$. With this assumption, the estimation of the gravity equation with nonparametric term κ_{ij} is greatly simplified; first we estimate the gravity equation without the term κ_{ij} to obtain the coefficient β, and then we smooth the residual term nonparametrically. In principle, we could relax this assumption by estimating the partial linear model developed by (Robinson, 1988). We also leave this for future work.[14]

4.3 Mapping the Gravity Equation into Economic Indicators

The gravity equation derived from an explicit modeling of the decision problem of workers allows us to map each term of gravity equation into economically meaningful objects. In this subsection, we explain them one by one.

4.3.1 Wage

The first measure that we can identify, in relative terms, is the wage, or more precisely the productivity per unit of effective labor supplied. Recall that from the gravity equation we directly estimate $\hat{\psi}_j = \epsilon \ln\big(w_j\big)$; we cannot identify the level of the wage because we cannot identify ϵ. Essentially, this is larger if the probability of workers who commute to j is higher than other places, conditional on distance.

The following decomposition illustrates more precisely, albeit in a recursive way, exactly what $\hat{\psi}_j$ captures. Let R_i be the number of residents in location i. Then, by adding $\log(R_i)$ to both sides of equation (5), we obtain

$$\log\big(y_{ij}\big) = \psi_j + \kappa_{ij} + \beta \log\big(D_{ij}\big) + \tilde{\mu}_i + \varepsilon_{ij}$$

where y_{ij} is the number of workers who commute from i to j and

$$\tilde{\mu}_i = -\mu_i + \log R_i = \log\left(\frac{R_i}{\sum_s \exp\left(\psi_s + \kappa_{ij} + \beta \log(D_{is})\right)}\right).$$

Summing over i and re-arranging yields

[13] See (Dube, Fox, & Su, 2012; Su & Judd, 2012) for the use of MPEC in different context.

[14] We need to be careful about the nonparametric specification of κ_{ij}. If we assume it to be fully nonparametric with respect to i and j, then D_{ij} is completely explained by κ_{ij} and β is not identified. We opt to constrain κ_{ij} to vary smoothly as a function of the locations of i and j.

$$\psi_j = \frac{1}{N}\sum_i \log(y_{ij}) - \frac{1}{N}\sum_i \log\left(\frac{R_i \exp \kappa_{ij} \left(D_{ij}\right)^{\beta}}{\sum_s \exp\left(\psi_s + \kappa_{is}\right)(D_{is})^{\beta}}\right) - \frac{1}{N}\sum_i \varepsilon_{ij}$$

Hence the wage at destination j can be decomposed into three parts: the aggregate log-inflow, an "accessibility" measure, and the deviation from the gravity equation due to measurement error. The last term is likely to be small by the law of large numbers. In fact, if we use the regression residual $\hat{\varepsilon}$ for ε, $\frac{1}{N}\sum_i \hat{\varepsilon}_{ij} = 0$ by construction due to the destination fixed effects. The intuition of the accessibility measure is as follows: a higher number of people commuting to j may be due to the high wage at j, or it may be due to many people living at locations i close to j. Furthermore, if the employment opportunities ψ_s available to people living at such a location i worsen, while holding ψ_j fixed, this will also increase the flow towards j; this effect is captured by the denominator in the accessibility measure.

4.3.2 Residential Worker Income

What is the mean income y_i^* of an agent who lives at a certain location i? The model outlined so far allows us to express this as a function of quantities estimated from equation (5).

Recall that we assume that k_{ij}, d_{ij} and $z_{ij\omega}$ directly affect productivity, so a worker's income is $y_{i\omega} = y_{ij\omega} | j \in \arg\max_s y_{is\omega}$. This random variable is also Fréchet with scale $T_i = \sum_s T_{is} = \sum_s w_s^{\epsilon} k_{is}^{\epsilon} d_{is}^{-\epsilon}$, and its distribution is independent of the destination j (this is a non-generic property specific to the Fréchet distribution). Hence the mean income y_i^* is Fréchet distributed with scale T_i. Using standard properties (see Appendix A.1) we obtain that both $\log E(y_{i\omega})$ and $E(\log(y_{i\omega}))$ are proportional to $\log(T_i)$. This leads to the following measure based on the estimated parameters in equation (5):

$$\widehat{\log(\Phi_i)} = \log\left(\sum_s \exp\left(\hat{\psi}_s + \hat{\kappa}_{is} + \hat{\beta}\log\left(D_{ij}\right)\right)\right)$$

This will be our estimated measure of mean (residential) income.

4.3.3 Output

In order to infer output at a particular location, some assumptions on the production function are necessary. (Ahlfeldt et al., 2014) assume a Cobb-Douglas production function with labor and land as inputs, and with a constant share parameter and competitive input markets. For our purpose, it is sufficient to assume that the production function is Cobb-Douglas with respect to labor input with constant share parameter and competitive input market; the rest of the inputs can be unspecified. By this assumption, the output is proportional to the labor expenditure, which we can directly map to gravity equation.

Recall from the previous section that the labor expenditure for a worker from i working in j, namely $y_{ij\omega} | j \in \arg\max_s y_{is\omega}$, is Fréchet distributed with scale $T_i = \sum_s T_{is}$. The average labor

expenditure is thus the mean of this random variable, which is given by $T_i^{1/\epsilon}\Gamma(1-1/\epsilon)$ where $\Gamma(\cdot)$ is the gamma function.

Output X_j at a location j can be expressed as the sum over all origin locations i of the commuting inflow from i, namely $R_i\pi_{ij}$, multiplied by the labor expenditure per worker from i. Hence,

$$\begin{aligned} X_j &= \sum_i E\left(y_{ij\omega}\middle| j \in \arg\max_s y_{is\omega}\right) R_i\pi_{ij} \\ &= \Gamma(1-1/\epsilon)\sum_i \left(\sum_s w_s^\epsilon k_{is}^\epsilon d_{is}^{-\epsilon}\right)^{1/\epsilon} R_i \frac{w_j^\epsilon k_{ij}^\epsilon d_{ij}^{-\epsilon}}{\sum_s w_s^\epsilon k_{is}^\epsilon d_{is}^{-\epsilon}} \\ &= \Gamma(1-1/\epsilon)\sum_i R_i w_j^\epsilon k_{ij}^\epsilon d_{ij}^{-\epsilon}\left(\sum_s w_s^\epsilon k_{is}^\epsilon d_{is}^{-\epsilon}\right)^{1/\epsilon-1}. \end{aligned}$$

Here we used the expression for π_{ij} from equation (4). Taking logs and using parameters estimated from equation (5) yields our working measure for output. Note that we need to estimate or consider an assumed value for ϵ in order to construct this measure.

$$\widehat{\log(X_j)} = \hat{\psi}_j + \log\left(\sum_j R_i e^{\hat{\kappa}_{ij}+\hat{\beta}\log(D_{ij})}\left(\sum_s e^{\hat{\psi}_s+\hat{\kappa}_{is}+\hat{\beta}\log(D_{is})}\right)^{1/\epsilon-1}\right).$$

4.4 Relation to (Ahlfeldt et al., 2014)

For the most part, the model in (Ahlfeldt et al., 2014) is strictly more general than the model presented here. However, the estimation technique are different in the two papers. (Ahlfeldt et al., 2014) remarkably show that flow data is not necessary in order to estimate wages and rents. They show that if amenities and underlying productivity are exogenously given, there is a unique equilibrium in terms of residential and work assignments wages and rents. Using this, they then prove that fixing the fundamental model parameters, there is a one-to-one mapping between the vector of residential and work populations, on one hand, and the vector of wages, on the other. This system of equations is exactly identified, and they use an iterative procedure to find the vector of wages.

In this paper, we take the more straightforward approach of using commuting flows to estimate wages. Given that now the model is highly over identified, we allow for more flexibility in terms of pairwise factors. Even so, we can evaluate the fit of the model to the data.

5 Measuring Mobility, Productivity and Economic Activity using Mobile Phone Data

We now show how to empirically implement the framework. We describe the data and the method to construct commuting flows, and we then describe the empirical specification that we use. Finally, we review the model and construct the output and income measures using the estimated parameters.

5.1 Data Sources

We use two main data sources: transaction-level cell phone data and VIIRS satellite captured nighttime light data. We describe these in turn, as well as other administrative and geographic data.

5.1.1 Transaction-Level Cell Phone Data

We use call detail record (CDR) data from multiple operators in Sri Lanka for a period of over a year. CDR data includes an observation for each transaction, such as making or receiving a voice call, sending or receiving a text message, or initiating an internet connection through GPRS.[15] For our purposes, each observation contains a timestamp, the user identifiers of the participants, and the cell antennas to which they are connected. The data is anonymized at the user level, which allows us to observe all transactions associated with a given user throughout the period.[16]

The geographic locations of the cell antennas are described in an auxiliary dataset. In general there are multiple cell antennas associated with the same cell tower. For this analysis, we group cell antennas at the tower level. Towers are not evenly distributed in space – they are much denser in urban and developed areas.[17]

5.1.2 VIIRS Satellite Nightlight Intensity Data

Researchers have recently started using nighttime lights data captured from satellites as a proxy for income in developing countries (Chen & Nordhaus, 2011; Henderson et al., 2010; Mellander et al., 2013). This data, which is available at fine geographic resolution for the entire world, responds to the problem of unreliable or inexistent economic activity data in developing countries, especially at subnational levels. Following the literature, we investigate the correlation between the income measures derived from our method and the nighttime lights.

[15] The dataset also contains information on recharge (phone credit top-up); we do not use this information.
[16] More precisely, each user ID corresponds to a unique telephone number. Thus, in principle the same person can have multiple cellphones, or they can change their cellphone number.
[17] We are not able to report descriptive statistics about the spatial distribution of cell towers due to a non-disclosure agreement.

In our main specification in section 6 we use a new dataset of nighttime lights, captured by the Visible Infrared Imaging Radiometer Suite (VIIRS), which recently became available. In what follows, we describe this data in more detail. (We also present results using the older nighttime lights data based on the United States Air Force Defense Meteorological Satellite Program (DMSP), which were also used in (Chen & Nordhaus, 2011; Henderson et al., 2010; Mellander et al., 2013))

The Visible Infrared Imaging Radiometer Suite (VIIRS) is a component of the Suomi National Polar-orbiting Partnership (S-NPP) satellite, launched in October 2011.[18] Its primary purpose is to collect measurements of clouds, aerosols, ocean color, surface temperature, fires, and albedo. The Earth Observation Group (EOG) curates global monthly composite cloud-free images, using the day/night band of the VIIRS. This band captures low light images on a nightly basis. The raw data is restricted to nights with zero moonlight, and to areas without clouds (as detected using another band of the VIIRS).[19] In this project we use one monthly cloud-free composite.

Figure 3 shows the DMSP data source used in (Chen & Nordhaus, 2011; Henderson et al., 2010; Mellander et al., 2013) and the new VIIRS data for the region of Colombo. Several advantages of the VIIRS data are visible. The resolution is higher, at 15 arc second grids compared to 30 arc second grids for the DMSP. The VIIRS data seems to be more "in focus." Importantly for our (urban) application, the VIIRS data is not saturated in highly developed and urban areas, whereas the DMSP data is saturated at its highest value.

5.1.3 Geographic Information and Census Population

We use population counts from Sri Lanka's 2011 census, at the Grama Niladhari (GN) level, the lowest administrative level in Sri Lanka. There are 14,021 GN's in Sri Lanka in our data.

We obtained geographic administrative GN boundaries from the Survey Department of Sri Lanka, which we combine with spatial data on cell towers. We compute the geodesic distance in kilometers between every pair of towers.[20]

For each cell tower, we interpolate the population based on the information from the census. Specifically, we assume that population is uniformly distributed within each GN. We partition every cell tower's Voronoi cell into subareas corresponding to different GNs, and calculate the population of each subarea, based on its land surface relative to the entire GN it belongs to. Our estimate of the cell tower's population is obtained by summing over all subareas.

[18] Source: http://viirsland.gsfc.nasa.gov/ and http://npp.gsfc.nasa.gov/viirs.html.
[19] This data is available at http://ngdc.noaa.gov/eog/viirs/download_monthly.html
[20] We use the `geodist` add-on STATA command, written by Robert Picard.

5.2 Computing Mobility Flows from Cell Phone Data

Obtaining detailed commuting data within cities has traditionally proved challenging. This is typically achieved using infrequent and expensive large-scale transportations surveys. For example, the Chicago Metropolitan Agency for Planning (CMAP) organized transportations surveys in 1990 and 2007-2008; the sample size for the latter was 10,552 households, who provided detailed travel inventories for each member of their household.[21] Another example is the US Census *Long Form*, which is administered to around one in six households, and includes information on the workplace location. (The 2010 Census Long Form was replaced by the American Community Survey, which is an annual rotating survey.[22])[23]

Recently, researchers have proposed using location information from cell phone data to estimate users' home and work locations, and commuting flows. (Csáji et al., 2013) use cell phone data from Portugal to explore users' mobility from several angles. Relevant for this paper, they categorize "home" and "work" locations based on the frequency with which users connect to towers. The key idea is to use data from an extended period of time for the same user to identify the most common location where s/he connects during a night time interval (for example 21:00 to 06:00) and during a working time interval. We use this method as a robustness check (the exact procedure is described in appendix B.1).

Other papers have developed methods to construct *transient* mobility flows from cell phone location data (Calabrese, Di Lorenzo, et al., 2011; Iqbal et al., 2014; P. Wang et al., 2012). For example, (Calabrese, Di Lorenzo, et al., 2011) first group together locations of a given user that are close in time and space into virtual locations, and then define trips as paths between consecutive virtual locations. Trip constructed in this manner have a start and end time, and can be aggregated into (time-dependent) origin-destination matrices.

The algorithm that we use to estimate commuting trips is a simplified version of the transient trips algorithm.[24] We define a commuting trip as a pair of locations (identified at the cell tower level) of the same user that occur on the same day, such that the origin location corresponds to the first transaction in the 5:00 to 10:00 time interval, and the destination location corresponds to the last transaction in the 10:00 to 15:00 interval. We choose to focus on the morning commute in light of evidence that shopping and other types of trips are more likely in the afternoon interval (Frank & Murtha, 2010). By definition, a user can have at most one trip per day, and some users will not have trips on certain days, if they do not have at least

[21] See http://www.cmap.illinois.gov/data/transportation/travel-tracker-survey.
[22] See https://www.census.gov/history/pdf/ACSHistory.pdf.
[23] Similar transportation surveys exist in developing countries as well. For example, in Sri Lanka the CoMTrans Survey, organized in 2012-2013, interviewed around 31,000 households in the Western Province. Unfortunately, this data is currently not available for comparison with the measures we construct.
[24] The authors thank Danaja Maldeniya for providing an early version of the code used to process *trips* from the raw cell phone data.

one transaction in each of the morning and afternoon time intervals.[25] We next aggregate all trips on a certain day to obtain an origin-destination (OD) matrix of commuting flows between pairs of cell towers.

We now provide some statistics on the commuting flow data. Our full sample covers more than a year of data and nine hundred million individual commuting trips. Of the total number of tower pairs with positive flows, approximately 3.7% have on average at least one commuting trip per day.

For the results in Section 6, we aggregate the OD matrix over 262 non-holiday weekdays. The total number of trips is approximately 30% of the theoretical maximum if we observed each user on every day in the sample. We restrict the sample to trips between different towers, which are no more than 50km apart. We do not include within tower commuting flows, which account for 49% of the total number of trips, because of the possibility that they capture non-commuting behavior.

For our working sample, approximately 8.9% of tower pairs correspond to commuting flows of at least one commuting trip per day on average. In other words, most commuting flows in the sample are small. This is a consequence of the large number of towers and the long time period; we expect any individual commuting flow to be a noisy measure of the true underlying commuting flow between the origin and the destination areas. Taken together, however, these commuting flows hold a considerable amount of information.

One concern that may arise is the representativeness of the commuting flows from the cell phone data. We conduct two robustness checks to address this issue. First, we use an alternative measure of commuting flows as a robustness exercise. The key idea is to use the entire history of locations for a user to extract a "home" and a "work" location (Csáji et al., 2013). The procedure is described in Appendix B.1. Second, we compare the number of cell phone users whose home are categorized in a particular Divisional Secretariats (DS) with the population extracted from the population census.[26] Figure 4 illustrates the comparison. The fact that an increase in DS population is associated with a proportional increase of cell phone users with a factor of 1 implies that there is no differential sample selection of cell phone users from the entire population of Sri Lanka, at the level of DS.

5.3 Gravity Equation and Smoothing

In this section we describe how we estimate the gravity model described by equation (5). Recall the equation:

[25] In ongoing work we plan to apply weights that are the inverse of the probability to observe a trip, given the number of data points (calls, text messages or internet connections) on that particular day. We can also vary the home and work time intervals to assess robustness.
[26] Divisional Secretariat is an administrative units, and there are 313 DS in Sri Lanka.

$$\log(\pi_{ij}) = \psi_j + \kappa_{ij} + \beta \log(D_{ij}) - \mu_i + \varepsilon_{ij}$$

Here π_{ij} measures the commuting probability from i to j, as a fraction of the aggregate outbound flow from i.

We compute these probabilities using the commuting flows described in the previous section. The equation includes destination fixed effects ψ_j and origin fixed effects $-\mu_i$. The variable D_{ij} measures the direct flight distance (in kilometers) between the coordinates of towers i and j (plus a small number). To account for the possibility of users connecting to nearby towers even in the absence of movement, we also include a dummy for whether the Voronoi cells of towers i and j share an edge.

The sample for the gravity equation is composed of all ordered pairs of distinct towers with positive aggregate commuting flow (over the entire time period), located at most 50km away.

We estimate the gravity equation in two steps. In the first step, we estimate a linear regression model using a special procedure to account for the computational burden imposed by the two-way fixed effects (Guimarães and Portugal 2009). [27] We report the results in Table 1. As expected from Figure 1, the distance between the origin and destination is strongly and negatively associated with the commuting flow. Conversely, neighboring towers have a higher flow, after controlling for distance. In the second step, we perform local linear smoothing of several parameters estimated from the gravity model. We smooth the destination fixed effects using a quartic kernel in geographic coordinates. In addition, we use a variable bandwidth to account for the large variations in tower density across space (Fan & Gijbels, 1992). Specifically, the smoothed destination fixed effects are given by:

$$\hat{\tilde{\psi}}_j \equiv \frac{\sum_s \frac{1}{h_s} K\left(\frac{d(j,s)}{h_s}\right) \hat{\psi}_s}{\sum_s \frac{1}{h_s} K\left(\frac{d(j,s)}{h_s}\right)}$$

In words, the smoothed version of the destination fixed effect at j is a weighted average of other fixed effects, with higher weight on geographically nearby towers. In the expression above $K(x) = (1-x^2)^2 \mathbb{I}\{x^2 \leq 1\}$ is a quartic kernel, $d(j,s)$ denotes the geodesic distance between towers j and s, and h_s is a tower specific bandwidth. Specifically, we choose $h_s = hR_s/\bar{R}$ where $h = 0.05$ degrees (approximately 5.5 km) and $R_s/\bar{R}$ is the radius of the equivalent area disc of tower s, divided by the average radius (over all towers s').[28] In other words, we choose a bandwidth of approximately 5.5 km for the tower with average

[27] This procedure is implemented in STATA using the add-on command `gpreg`, developed by Johannes F. Schmieder. (Head & Mayer, 2013) also recommend using this procedure (p 27).
[28] The equivalent area disc of a polygon is the disc with the same area as the polygon.

(equivalent area disc) radius, and adjusted it for each other tower proportionally to the radius of that tower.

We also smooth the residuals from the gravity equation $\hat{\varepsilon}_{ij}$ to obtain our estimates of κ_{ij}. The idea is the same as for destination fixed effects, except that the residuals depend on both origin and destination tower. The formula is

$$\widehat{\boldsymbol{\kappa}}_{\boldsymbol{ij}} \equiv \frac{\sum_{i',j'} \frac{\mathbf{1}}{\boldsymbol{h}_{i'}\boldsymbol{h}_{j'}} \boldsymbol{K}\left(\frac{\boldsymbol{d}(\boldsymbol{i},\boldsymbol{i}')}{\boldsymbol{h}_{i'}}\right) \boldsymbol{K}\left(\frac{\boldsymbol{d}(\boldsymbol{j},\boldsymbol{j}')}{\boldsymbol{h}_{j'}}\right) \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{i'j'}}}{\sum_{i',j'} \frac{\mathbf{1}}{\boldsymbol{h}_{i'}\boldsymbol{h}_{j'}} \boldsymbol{K}\left(\frac{\boldsymbol{d}(\boldsymbol{i},\boldsymbol{i}')}{\boldsymbol{h}_{i'}}\right) \boldsymbol{K}\left(\frac{\boldsymbol{d}(\boldsymbol{j},\boldsymbol{j}')}{\boldsymbol{h}_{j'}}\right)} \tag{7}$$

In words, the smoothed residual from the gravity equation corresponding to the origin-destination pair (i,j) is a weighted average of other residuals, with higher weight given to residuals with origins close to i and destinations close to j.

5.4 Output and Income Measures

We now use the formulas for residential (origin) income $\widehat{\log(\Phi_i)}$ and (destination) output $\widehat{\log(X_j)}$ introduced in sections 4.3.2 and 4.3.3. Both measures are constructed using the distance matrix $\left(D_{ij}\right)_{ij}$, residential population R_i, and the quantities estimated from the gravity or equation or derived thereof: $\hat{\psi}_j$, $\hat{\kappa}_{ij}$, and $\hat{\beta}$. In addition, we need to make an assumption on ϵ in order to construct the output measure. Given that we do not actually use the output measure at this stage, we defer exploring how the output measure is sensitive to the ϵ parameter. We compute both measures on the same sample as we used to estimate the gravity equation. For example for the income measure at origin i, this means that we sum over all destinations $j \neq i$ that are within 50km of i and for which there is a positive flow between i and j in the data. As mentioned in section 5.3, we exclude within-tower flows due to the possibility that they capture non-commuting behavior; however, this decision is not unequivocal, and we plan to explore it in more detail in future work.

6 Empirical Results – A First Comparison: Residential Income and Nightlights

We proceed to compare of residential average income predicted by our method and the nighttime lights. As we have already discussed in Section 5.1.2, nighttime lights are the best possible currently available data to validate our model. It has been pointed out that nighttime lights are a good indicator of residential income with fine spatial resolution (Chen & Nordhaus, 2011; Henderson et al., 2010; Mellander et al., 2013), while other sources of information about income or economic activity are only available at a coarse spatial aggregation level (provincial level).

Following the discussion above, we compare the nighttime lights with our measure of average residential income predicted by the model. Figure 3 provides a visual comparison between VIIRS nightlights and the model-predicted income $\log(\Phi_i)$ in the Colombo area. On average, the two measures are correlated and both capture the intuitive pattern of residential development in the Colombo area. One notable difference between the two measures is an oil refinery in the top center area of the figure, which is bright in the VIIRS nighttime lights and much less prominent as measured by the income measure. This indicates that there is room for improvement of the VIIRS nighttime lights as a measure of income.

While Figure 3 suggests that our residential income measure is a good proxy for VIIRS nighttime lights, it is not clear whether the measure has additional information over more easily available proxy for income. To answer this question, we need to investigate whether our income measure is a significant predictor of VIIRS nightlights even conditional on other proxies for income. Table addresses this point. Each column in this table reports results from a regression of mean VIIRS nighttime lights on the model-predicted log mean income and other covariates. Column (1) reports results from a regression without any covariates. The log income measure is significantly positive and the R-squared is 0.71, confirming that the measure is a good predictor of VIIRS nighttime lights. Column (2) controls for the size of the Voronoi cell that includes the cell phone tower, population density extracted from the population census, and the distance to Colombo Fort (the central business district of Colombo). Note that these three variables are usually easier to obtain than the cell phone data. The results show that while these three variables are also a significant predictor of VIIRS nighttime lights, our income measure is also significant conditional on these variables. This indicates that our measure has additional information over easily obtainable index on income.

Column (3) to (5) test whether the income measure has additional explanatory power over alternative and simpler statistics derived from cell phone data. Column (3) tests whether our income measure is significant conditional on the destination fixed effects of the gravity equation, i.e. our measure of log wage. Since our measure of residential income is a transformation of the set of destination fixed effects, it might well be that the destination fixed effects are already a good measure of nightlights, and there is no additional information from the (non-linear) transformation. The regression results confirms that it is not the case. Similar robustness check are conducted in Column (4) and (5). In Column (4), we test whether the log income measure is still significant after conditioning on the log of the ratio of inflow over outflow, while Column (5) conditions on the log of inflow. Both are simpler statistics that are derived from the commuting flows extracted from cell phone data. We find mixed results here; In both Column (4) and (5), log income is still significant.

Table replicates the results of Table further conditioning on the district fixed effects. The goal of this exercise is to verify that the results in Table are not only picking up variation due to some districts having both stronger nighttime lights and higher predicted income. The message is essentially the same; the log income measure is significant except when it is conditioned on the log inflow.

In Table 4, we replicate Table 2 using the older nighttime lights data based on United States Air Force Defense Meteorological Satellite Program (DMSP) instead of VIIRS data. DMSP nightlights data are used in (Chen & Nordhaus, 2011; Henderson et al., 2010; Mellander et al., 2013). Basic results are the same except for Column (5), where now our income measure is significant while the log inflow is not.

Table 5 is a robustness exercise for Table 3 where we estimate the entire model using an alternative measure of commuting flow. Specifically, for each user we use the entire panel of locations to determine a "home" and a "work" location for that user; the exact procedure is detailed in Appendix B.1. The results are quite similar to the baseline results reported in Table 3.

7 Extensions

7.1 Time Variation

The model in the above section is static and does not have any time component. However, in reality, both economic activity and people's mobility vary in time. In this subsection, we provide two alternative ways to incorporate time in our model.

The first way to incorporate time is to assume that workers decide working locations and wages in the first period, and in every subsequent period workers determine whether they actually show up for work (at the predetermined wage). The second way to incorporate time is to allow wages and working locations to change every period. The first scenario implies that commuting flows averaged over time satisfy a gravity equation, and that the income and economic activity measures for a given time period depend on the commuting flows realized in that period. The second scenario implies that commuting flows in every period satisfy a gravity equation.

Which interpretation of the two is suitable depends on the context and the duration of the reference period. If periods are short, workers cannot adjust their working location and the first scenario is more natural, while if it they are long the second scenario is more natural. Note that we can formulate the second scenario using an alternative interpretation for "wages" w_j as a return from a general economic activity (not necessarily the production of a final good). For example, w_j may measure the value of shopping or other economic transaction occurring at j. Note that in this case the interpretation of economic activity as output, and that of the residential income measure, will probably change. We leave this for future work.

One can also consider a type of dynamic model where workers do not have complete information about the available jobs and workers search for jobs in each period. Such models are considered in (Rouwendal, 1999), for example. In such a model, revelation preference argument of workers' working location choice does not work nicely. Hence, we do not consider such a model in this paper.

7.2 Sensitivity of the exercise to the level of aggregation

In the model, we assume that all firms at a given location offer the same wage and the individual productivity shock $z_{ij\omega}$ does not depend on the firm within destination j. However, in reality, the number of firms and wages offers in a given destination location may vary. In particular, in our data, the finest achievable geographic aggregation level is determined by the density of cell towers, and in particular it is correlated with economic activity (See Table 2).

In Appendix A.2, we describe how the geographic aggregation of origin and destination locations matter. Here, we summarize the main results. First, the aggregation level of origins do not matter, except that it is desirable to weight the gravity equation by the number of people who reside in the origin upon estimation. Secondly, the aggregation level of destinations affects the interpretation of w_j; if location j is in fact composed of several sublocations, w_j represents a CES aggregate of the wages at the sublocations within j. Importantly, the level of aggregation does not affect the expressions for the output and income measures derived in the previous section.

7.3 Worker Skill Heterogeneity

So far we have assumed that all workers are homogenous and face the same potential wage profile. In reality, there are different types of workers facing different wage profiles for the same destination locations. If we observe commuting flows for different skill types of people, we can separately estimate the gravity equations for different skill types. However, it is often the case that we only observe the aggregate flows, as in our case with cell phone data.

In Appendix A.3, we show how we can extend our method given data on the skill composition of workers in each origin. Such information, for example on educational attainment, is usually available from the population census or other surveys. Technically, such a model does not lead to a linear gravity equation, and the estimation of parameters requires nonlinear estimation methods such as maximum likelihood.

7.4 Alternative Interpretation on k_{ij}, $z_{ij\omega}$ and d_{ij}

In Section 4, we assumed that k_{ij} and $z_{ij\omega}$ affect production and d_{ij} affects the income each worker can bring back home. However, an alternative interpretation of these terms is that they just affect workers' decisions of where to commute, but they do not affect production or income. The gravity equation (5) holds irrespective of these considerations. However, the mapping from terms of the gravity equation to economic indicators (output and income) will change. Which interpretation is better is mostly an empirical question, and we leave the investigation for future work.

8 Conclusion

In this paper we used commuting flows derived from cell phone data and a job choice model to estimate the spatial distribution of productivity and economic activity. Our exploratory analysis strongly suggests that this approach is justified. Cell phone data contains ample information on mobility, which in turn is informative about economic activity. We briefly discuss what we believe are the most promising extensions and applications for future research.

In our model, following most urban economic models (Ahlfeldt et al., 2014), we assume all workers are homogenous. However, cities are home to workers of very different skill levels, especially in developing and middle-income countries. Hence, it would be interesting to study how mobility patterns differ and interact for various worker skill levels (for example, as measured by educational attainment). Section 7.3 discussed how the model and estimation can be adapted to achieve this. In terms of data, census or socio-economic household survey data with relatively precise geographic indicators can be used to calibrate the skill mix at each residential location. The joint analysis of the spatial distribution of worker skill levels and their mobility patterns may help lead to a better understanding of urban patterns, and thus seems to be a promising avenue for future research.

Unlike surveys, which offer snapshots at certain moments in time, cell phone data is available continuously. We hope to explore this important feature in future work. First, our method can be applied to investigate suburbanization patterns. In the Colombo area, between 2001 and 2012, there has been a shift in population away from the central areas towards the suburbs (Ministry of Transport, 2013). Our method can complement this type of analysis by measuring the change in mobility and economic activity patterns within Colombo.

Ultimately, we hope researchers will be able to use our method to evaluate the economic impact of policies or natural events that change mobility patterns. Examples include fuel price changes, road closures (such as the *Oborodh* in Bangladesh in the beginning of 2015), and travel restrictions (such as the *cordon sanitaire* in West Africa due to Ebola outbreaks). Such studies have the promise to deliver both a deeper understanding of micro-behavior related to these shocks, as well as useful insights for policy makers who seek to quantify and alleviate the negative consequences of these shocks.

References

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., & Wolf, N. (2014). The Economics of Density: Evidence from the Berlin Wall. *NBER Working Paper Series*, (20354).

Anas, A. (1983). Discrete choice theory, information theory and the multinomial logit and gravity models. *Transportation Research Part B: Methodological*, *17*(I), 13–23. doi:10.1016/0191-2615(83)90023-1

Anas, A., & Arnott, R. (1998). Urban Spatial Structure. *Journal of Economic Literature*, *36*(3), 1426–1464.

Anderson, J. E. (1979). A theoretical foundation for the gravity equation. *The American Economic Review*, *69*(1), 106–116.

Blanchard, O. J., & Katz, F. (1992). Regional Evolutions. *Brookings Papers on Economic Activity*, *1*, 1–75.

Blumenstock, J. E. (2012). Information Technology for Development Inferring patterns of internal migration from mobile phone call records : evidence from Rwanda, *18*(2), 37–41.

Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., & Ratti, C. (2011). Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome. *IEEE Transactions on Intelligent Transportation Systems*, *12*(1), 141–151. doi:10.1109/TITS.2010.2074196

Calabrese, F., Di Lorenzo, G., Liu, L., & Ratti, C. (2011). Estimating Origin-Destination Flows Using Mobile Phone Location Data. *IEEE Pervasive Computing*, *10*(4), 36–44. doi:10.1109/MPRV.2011.41

Calabrese, F., Ferrari, L., & Blondel, V. D. (2014). Urban Sensing Using Mobile Phone Network Data: A Survey of Research. *ACM Computing Surveys*, *47*(2), 1–20. doi:10.1145/2655691

Chen, X., & Nordhaus, W. D. (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(21), 8589–8594. doi:10.1073/pnas.1017031108

Csáji, B. C., Browet, A., Traag, V. a., Delvenne, J. C., Huens, E., Van Dooren, P., … Blondel, V. D. (2013). Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and Its Applications*, *392*(6), 1459–1473. doi:10.1016/j.physa.2012.11.040

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, a. E., … Tatem, a. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 1–6. doi:10.1073/pnas.1408439111

Dube, J.-P., Fox, J. T., & Su, C. (2012). Improving the Numerical Performance of Static and Dynamic Aggregate Discrete Choice Random Coefficients Demand Estimation. *Econometrica*, *80*(5), 2231–2267. doi:10.3982/ECTA8585

Duran-fernandez, R., & Santos, G. (2014). Gravity , distance , and traffic flows in Mexico. *Research in Transportation Economics*, *46*, 30–35. doi:10.1016/j.retrec.2014.09.003

Eaton, J., & Kortum, S. (2002). Technology, Geography, and Trade. *Econometrica*, *70*(5), 1741–1779.

Erlander, S., & Stewart, N. F. (1990). *The gravity model in transportation analysis: theory and extensions*.

Fan, J., & Gijbels, I. (1992). Variable Bandwidth and Local Linear Regression Smoothers. *The Annals of Statistics*, *20*, 2008–2036. doi:10.1214/aos/1176348900

Frank, P., & Murtha, T. (2010). Trips Underway by Time of Day by Travel Mode and Trip Purpose for Metropolitan Chicago.

Fujita, M., & Ogawa, H. (1982). Multiple Equilibria and Structural Transition of Non-Monocentric Urban Configurations. *Regional Science and Urban Economics*, *12*, 161–196.

Guimarães, P., & Portugal, P. (2009). A Simple Feasible Alternative Procedure to Estimate Models with High-Dimensional Fixed Effects. *IZA Discussion Paper Series*, (3935).

Hare, D. (1999). “Push” versus “pull” factors in migration outflows and returns: Determinants of migration status and spell duration among China’s rural population. *Journal of Development Studies*, *35*(3), 45–72. doi:10.1080/00220389908422573

Head, K., & Mayer, T. (2013). Gravity Equations : Workhorse , Toolkit , and Cookbook. *Handbook of International Economics*. doi:10.1016/B978-0-444-54314-1.00003-3

Henderson, J. V., Storeygard, A., & Weil, D. N. (2010). Measuring Economic Growth from Outer Space. *American Economic Review*, *102*(2), 994–1028.

Iqbal, M. S., Choudhury, C. F., Wang, P., & González, M. C. (2014). Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, *40*, 63–74. doi:10.1016/j.trc.2014.01.002

Lucas, R. E., & Rossi-Hansberg, E. (2002). ON THE INTERNAL STRUCTURE OF CITIES. *Econometrica*, *70*(4), 1445–1476.

Mayda, A. M. (2010). International migration: A panel data analysis of the determinants of bilateral flows. *Journal of Population Economics*, *23*, 1249–1274.

McArthur, D. P., Kleppe, G., Thorsen, I., & Ubøe, J. (2011). The spatial transferability of parameters in a gravity model of commuting flows. *Journal of Transport Geography*, *19*(4), 596–605. doi:10.1016/j.jtrangeo.2010.06.014

McFadden, D. (1973). Conditional Logit Analysis of Qualitative Choice Behavior.pdf. In *Frontiers in Econometrics*. New York: Academic Press.

Mellander, C., Stolarick, K., Matheson, Z., & Lobo, J. (2013). Night-Time Light Data : A Good Proxy Measure for Economic Activity ?, (315), 33.

Mills, E. S. (1967). An Aggregative Model of Resource Allocation in a Metropolitan Area. *The American Economic Review Papers and Proceedings of the Seventy -Ninth Annual Meeting of the American Economic Association*, *57*(2), 197–210.

Ministry of Transport. (2013). *Draft Urban Transport Master Plan for Colombo Metropolitan Region and Suburbs*.

Niedercorn, J. H., & Bechdolt, B. V. (1969). An Economic Derivation of the "Gravity Law" of Spatial Interaction. *Journal of Regional Science*, *9*(3), 273–282. doi:10.1111/j.1467-9787.1969.tb01340.x

Ommeren, J. Van, & Fosgerau, M. (2009). Workers ' marginal costs of commuting. *Journal of Urban Economics*, *65*(1), 38–47. doi:10.1016/j.jue.2008.08.001

Peri, G. (2012). The effect of immigration on productivity: Evidence from US states. *Review of Economics and Statistics*, *94*(1), 348–358.

Reades, J., Calabrese, F., Sevtsuk, A., & Ratti, C. (2007). Cellular Census: Explorations in Urban Data Collection. *IEEE Pervasive Computing*, *6*(3), 30–38. doi:10.1109/MPRV.2007.53

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica*, *56*(4), 931–954.

Rouwendal, J. (1999). Spatial job search and commuting distances. *Regional Science and Urban Economics*, *29*, 491–517.

Simini, F., González, M. C., Maritan, A., & Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, *484*, 96–100. doi:10.1038/nature10856

Sohn, J. (2005). Are commuting patterns a good indicator of urban spatial structure? *Journal of Transport Geography*, *13*, 306–317. doi:10.1016/j.jtrangeo.2004.07.005

Su, C., & Judd, K. L. (2012). Constrained Optimization Approaches to Estimation of Structural Models. *Econometrica*, *80*(5), 2213–2230. doi:10.3982/ECTA7925

Van Ommeren, J., J. van den Berg, G., & Gorter, C. (2000). ESTIMATING THE MARGINAL WILLINGNESS TO PAY FOR COMMUTING. *Journal of Regional Science*, *40*(3), 541–563.

Wang, F. (2001). Explaining intraurban variations of commuting by job proximity and workers' characteristics. *Environment and Planning B: Planning and Design*, *28*(1993), 169–182. doi:10.1068/b2710

Wang, P., Hunter, T., Bayen, A. M., Schechtner, K., & González, M. C. (2012). Understanding road usage patterns in urban areas. *Scientific Reports*, *2*, 1001. doi:10.1038/srep01001

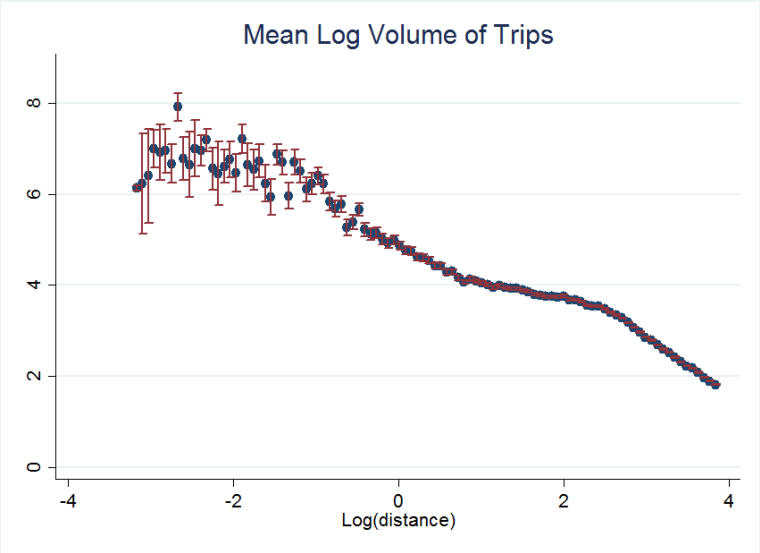
Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science (New York, N.Y.)*, *338*(6104), 267–70. doi:10.1126/science.1223467

Wesolowski, A. P., & Eagle, N. (n.d.). Parameterizing the Dynamics of Slums. *AAAI Spring Symposium: Artificial Intelligence for Development*.

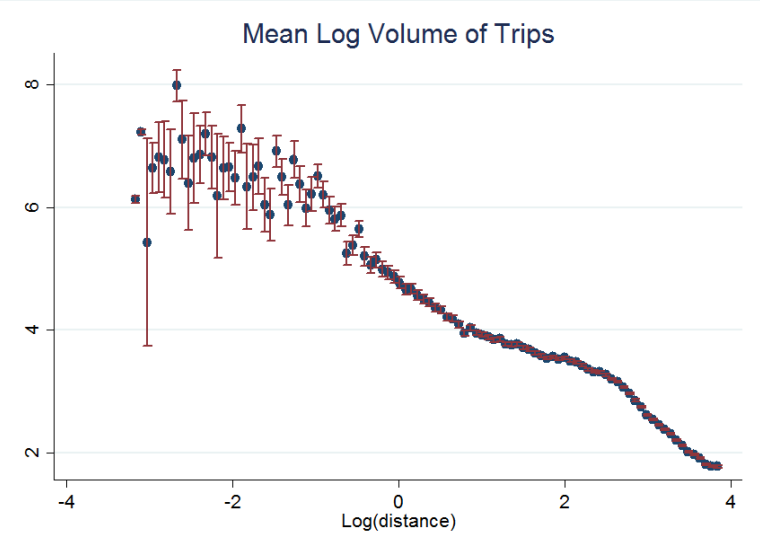
Wheaton, W. C. (2004). Commuting, congestion, and employment dispersal in cities with mixed land use. *Journal of Urban Economics*, *55*, 417–438. doi:10.1016/j.jue.2003.12.004

Wilson, A. (1967). A statistical theory of spatial distribution models. *Transportation Research*, *1*(3), 253–269. doi:10.1016/0041-1647(67)90035-4

Figures



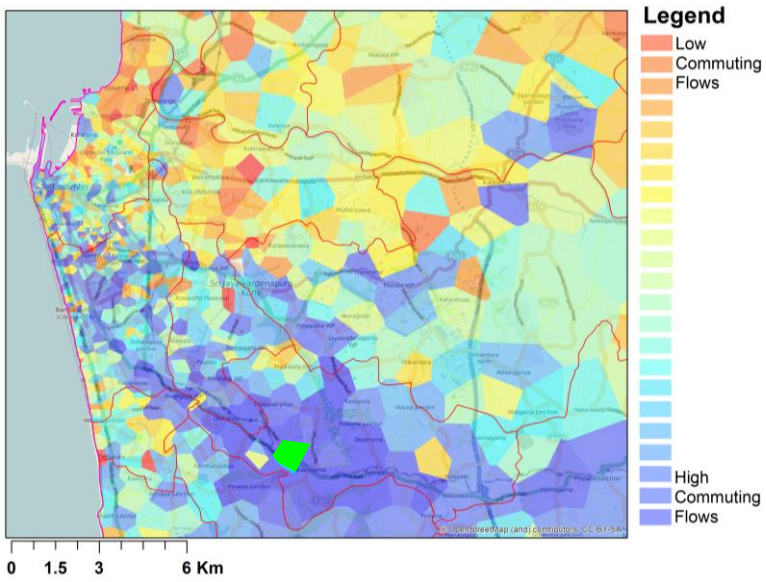
(A) Sri Lanka



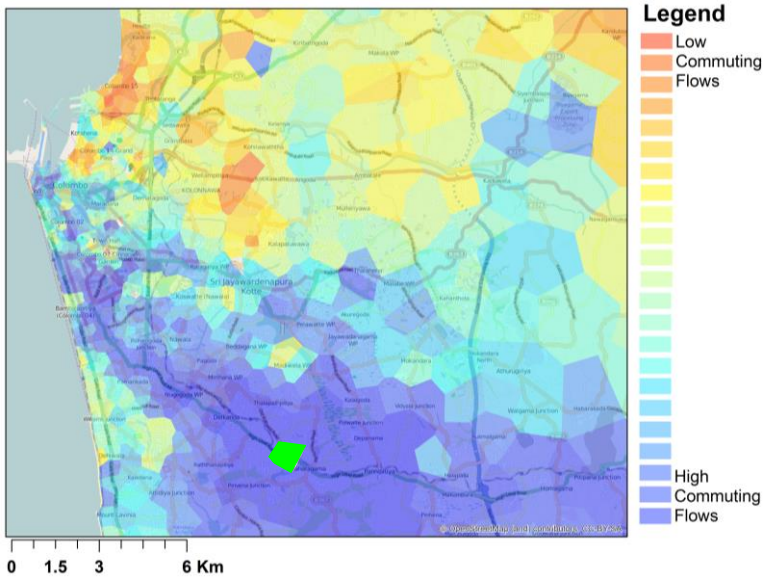
(B) Western Province only

Figure 1. Commuting Trips and Distance.

Notes: These graphs show average log commuting volume between origin and destination tower pairs, as a function of the log distance between them. Panel (A) uses all pairs of towers in Sri Lanka, while panel (B) restricts to the Western Province (which is the most populated and includes the capital Colombo). Log distance is binned into 100 bins; 95% confidence intervals are computed using $\mathbf{1.96}$ times the bin-level standard error of the estimated average flow.



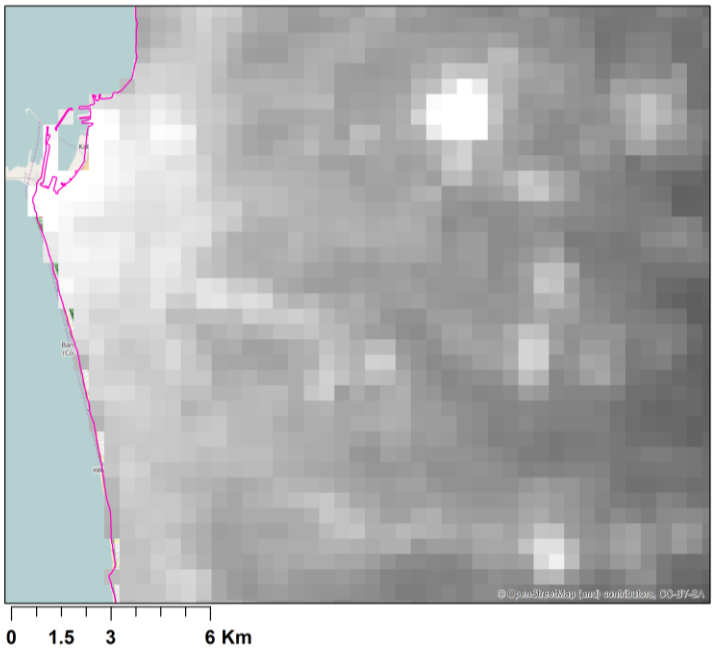
(A) Raw Log Commuting Flows



(B) Smooth Log Commuting Flows

Figure 2. Commuting flows from one origin.

Notes: Panel (A) shows raw and panel (B) shows smoothed commuting flows from a given origin tower cell (indicated by the green cell) to all possible other destination tower cells. Red lines indicate sub-district (divisional secretariat) boundaries.



(A) VIIRS Nighttime Lights

(B) Mean Residential Income Measure $\log(\Phi_i)$

Figure 3. Mean Residential Income and Nighttime Lights.

Note: the bright spot (top-center) in panel (A) corresponds to an oil refinery.

Tables

	Log(Commuting Flow)
$\log(D_{ij})$	-1.41
	(0.001)
Neighboring i and j	1.31
	(0.008)
Origin FE μ_i	Yes
Destination FE ψ_j	Yes
Observations	$\sim 10^6$
R-squared	0.52
Adj R-squared	0.52

Table 1. Gravity Model Estimation

Notes: This table reports the results from a regression of log commuting flows between towers i and j on the log distance between i and j, a dummy equal to 1 if i and j have neighboring Voronoi cells, and a set of origin (i) fixed effects, and a set of destination (j) fixed effects. The sample is all ordered pairs of distinct towers with positive commuting flow, and which are no more than 50km apart. The OLS regression is implemented in STATA using the `gpreg` command (Guimarães & Portugal, 2009).

	(1)	(2)	(3)	(4)	(5)
	Log nighttime lights (VIIRS)				
Log(mean income)	0.928***	0.116***	0.105***	0.108***	0.092***
	(0.056)	(0.024)	(0.024)	(0.024)	(0.023)
Log(tower cell size)		-0.298***	-0.302***	-0.300***	-0.319***
		(0.024)	(0.023)	(0.022)	(0.026)
Log(population density)		0.279***	0.275***	0.279***	0.251***
		(0.039)	(0.039)	(0.035)	(0.040)
$\mathrm{Log}(d_{\mathrm{capital}})$		0.001**	0.001***	0.001**	0.001*
		(0.001)	(0.001)	(0.001)	(0.001)
Destination FE			-0.032		
			(0.033)		
Log(inflow/outflow)				0.012	
				(0.106)	
Log(inflow)					0.153***
					(0.049)
Observations	$\sim 10^3$ - 10^4	$\sim 10^3$ - 10^4	$\sim 10^3$ - 10^4	$\sim 10^3$ - 10^4	$\sim 10^3$ - 10^4
R-squared	0.71	0.92	0.92	0.92	0.92

Table 2 Model-predicted Residential Income and VIIRS Nighttime Lights

Notes: Each column in this table reports the results from a regression of mean nighttime lights (VIIRS) in the Voronoi cell of a cell phone tower on log mean income ($\log \Phi_i$) in that cell tower (defined in section 4.3.2), and other regressors. The other regressors are constructed as follows: the tower cell size is the area of the Voronoi cell around the tower, population density is interpolated from the population census (see section 5.1.3), d_{capital} measures distance to the district capital, the destination fixed effect $\hat{\psi}_i$ is estimated using the gravity equation (see section 5.3), the inflow into a tower is the number of commuting trips that have that tower as their destination, and the outflow is the number of commuting trips originating from a tower. The sample consists of all cell phone towers in Sri Lanka. Standard errors are clustered at district level. *** p<0.01, ** p<0.05, * p<0.1.

	(1)	(2)	(3)	(4)
Dependent Var.	Log nighttime lights (VIIRS)			
Log(mean income)	0.049*	0.050*	0.047*	0.016
	(0.026)	(0.025)	(0.027)	(0.023)
Log(tower cell size)	-0.297***	-0.294***	-0.268***	-0.323***
	(0.012)	(0.011)	(0.022)	(0.010)
Log(population density)	0.206***	0.193***	0.226***	0.155***
	(0.037)	(0.032)	(0.034)	(0.025)
$\mathrm{Log}(d_{\mathrm{capital}})$	-0.001	-0.002	-0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Destination FE		0.194***		
		(0.031)		
Log(inflow/outflow)			0.232***	
			(0.074)	
Log(inflow)				0.297***
				(0.033)
District Fixed Effects	Yes	Yes	Yes	Yes
Observations	$\sim 10^3$ - 10^4	$\sim 10^3$ - 10^4	$\sim 10^3$ - 10^4	$\sim 10^3$ - 10^4
R-squared	0.95	0.96	0.95	0.96

Table 3 Model-predicted Residential Income and VIIRS Nightlights, Conditional on District Fixed Effects

Notes: Each column in this table reports the results from a regression of mean nighttime lights (VIIRS) in the Voronoi cell of a cell phone tower on log mean income ($\log \Phi_i$) in that cell tower (defined in section 4.3.2), and other regressors, and district fixed effects. See Table 2 for other notes. Standard errors are clustered at district level. *** p<0.01, ** p<0.05, * p<0.1.

Appendix A. Model Derivations

A.1. Some Facts about the Fréchet Distribution

We review some basic properties of the Fréchet distribution. The cumulative distribution function of a Fréchet random variable with scale parameter T and shape parameter ϵ is $F(z) = \exp\{-Tz^{-\epsilon}\}$. Consider a sequence of independent Fréchet random variables z_k with scale T_k and the same shape ϵ, for $k = \{1, \dots, K\}$. The probability that the maximum is achieved by the j^{th} variable, with $j \in \{1, \dots, K\}$, is given by

$$\Pr(j \in \operatorname{argmax}_k x_k) = \frac{\exp(T_j)}{\sum_k \exp(T_k)}.$$

The class of Fréchet random variables is closed with respect to the max operator. The random variable $\max_k z_k$ is Fréchet distributed with scale $T = \sum_k T_k$ and shape ϵ. Moreover, the conditional maxima, namely $z_j | j \in \operatorname{argmax}_k x_k$, have exactly the same distribution as the unconditional maximum.

The mean of a Fréchet distributed variable is $E(z) = T^{1/\epsilon}\Gamma(1 - 1/\epsilon)$ where $\Gamma(\cdot)$ is the gamma function. It follows that

$$\ln\big(E(z)\big) = \frac{\ln(T)}{\epsilon} + \ln\big(\Gamma(1 - 1/\epsilon)\big).$$

We also have that for some absolute constant K:

$$E(\ln(z)) = \frac{\ln(T)}{\epsilon} - \frac{K}{\epsilon}.$$

A.2. Sensitivity to Different Geographic Aggregation Level

In our main model, we assume that all firms at a given location offer the same wage and the individual productivity shock $z_{ij\omega}$ does not depend on the firm within destination j. Here we show how different levels of geographic aggregation of the origin and destination locations matter for our analysis.

First, we discuss how the aggregation level of origins matters. Suppose that origin i has outflow y_{ij} to destination j. Suppose that we divide origin i into two different origins, i' and i''. Also assume that i' and i'' are close so that for all j, $d_{i'j} \approx d_{i''j}$. Workers living at i' and i'' face approximately the same job choice problem, hence $\pi_{ij} \approx \pi_{i'j} \approx \pi_{i''j}$ for all j. This implies that the gravity equation holds for the same parameters, in particular, $\mu_i \approx \mu_{i'} \approx \mu_{i''}$. Note that, depending on the structure of the measurement errors, weighting the regression

by the number of workers in the origin leads to a more efficient estimator. Intuitively, this is because as the number of workers at i grows, π_{ij} is more precisely estimated.

Next, we discuss how the aggregation level of destinations matters. In the main model we assume that for each origin i destination j and worker ω the scale of the shock $z_{ij\omega}$ is constant. Consider instead the case when each destination j contains multiple sub-locations $\ell_j \in \{1, \dots, n_j\}$, and workers experience i.i.d shocks $z_{i\ell_j\omega}$ at the level of the sub-locations ℓ_j, and these shocks have constant scale. (For convenience, in what follows we drop the j subscript in ℓ_j.) Assume for simplicity that $k_{i\ell} = 1$ for all i and ℓ; the argument goes through without this assumption. Hence $z_{i\ell\omega} \sim F(z) = \exp\{-Tz^{-\epsilon}\}$. Assume sub-location ℓ has wage w_ℓ, and assume for simplicity that all sub-locations are located approximately at the same geographic location (so $d_{i\ell_j} = d_{ij}$ for all ℓ_j). It follows that income from working at ℓ_j is given by

$$v_{i\ell\omega} = \frac{z_{i\ell\omega} w_\ell}{d_{ij}} \sim G_{i\ell}(v) = \exp\{-Tv^{-\epsilon} w_\ell^\epsilon d_{ij}^{-\epsilon}\}$$

A worker at location i first chooses the best sub-location at j, which will give income: $v_{ij} = \max_\ell z_{i\ell\omega} w_\ell / d_{ij}$. The maximum of Fréchet-distributed variables with scales $T_\ell = T w_\ell^\epsilon d_{ij}^{-\epsilon}$ is still Fréchet-distributed, with scale $T_j = \sum_\ell T_\ell = T d_{ij}^\epsilon \sum_\ell w_\ell^\epsilon$. If we use the notation $w_j \equiv (\sum_\ell w_\ell^\epsilon)^{1/\epsilon}$, then the scale of the max at j becomes $T_j = T w_j^\epsilon d_{ij}^{-\epsilon}$. Thus we can write $v_{ij} = \max_\ell \frac{z_{i\ell\omega} w_\ell}{d_{ij}} \sim G_{ij}(v) = \exp\{-Tv^\epsilon\} w_j^{-\epsilon} d_{ij}^{-\epsilon}$. The constructed wage at location j, w_j, is a CES aggregate of the real wages at sub-locations of j. Once we have done this at every destination j, the probability of workers from i going to j is the same as before, except that w_j now has a different interpretation, i.e.

$$\pi_{ij} = \sum_\ell \Pr\left(\ell \in \arg\max_s y_{is\omega}\right) = \frac{\left(\frac{w_j}{d_{ij}}\right)^\epsilon}{\sum_s \left(\frac{w_s}{d_{is}}\right)^\epsilon}.$$

This again leads to a gravity equation, with a different interpretation of w_j.

We next show that the formulas for the output and income measures do not change under the above definition of w_j. For output (denote $K = \Gamma(1 - 1/\epsilon)$)

$$X_j = \sum_\ell \sum_i E\left(y_{i\ell\omega} \middle| \ell \in \arg\max_s y_{is\omega}\right) R_i \Pr\left(\ell \in \arg\max_s y_{is\omega}\right)$$

$$= \sum_\ell \sum_i K \left(\sum_s \frac{w_s{}^\epsilon}{d_{is}^\epsilon}\right)^{1/\epsilon} R_i \Pr\left(\ell \in \arg\max_s y_{is\omega}\right)$$

$$= K \sum_i \left(\sum_s \frac{{w_s}^\epsilon}{d_{is}^\epsilon} \right)^{1/\epsilon} R_i \sum_\ell \Pr\left(\ell \in \arg\max_s y_{is\omega} \right)$$
$$= K \sum_i R_i \left(\frac{w_j}{d_{ij}} \right)^\epsilon \left(\sum_s \frac{{w_s}^\epsilon}{d_{is}^\epsilon} \right)^{1/\epsilon - 1}$$

This is the same as the expression in Section 4.3.3, with $k_{ij} = 1$. For income,

$$E(y_i^*) = \sum_j \sum_k E\left(y_{ijk\omega} \middle| j \in \arg\max_s y_{ijs\omega} \right) \Pr\left(jk \in \arg\max_s y_{ijs\omega} \right)$$
$$= E\left(Fréchet\left(\sum_s \frac{{w_s}^\epsilon}{d_{is}^\epsilon} \right) \right),$$

which is the same as the expression in Section 4.3.2.

A.3. Worker Heterogeneity

In this sub-section we explain how we can relax the assumption that workers are homogenous. For simplicity, assume that there are just two types of workers, H and L. Assume that the wages for type H and L are different in each destination place j. We denote the wages w_j^H and w_j^L, correspondingly.

If we observe the commuting flow of type H and L separately, we can simply estimate the two gravity equations to back out w_j^H and w_j^L separately. However, we often do not separately observe the commuting flows of different types of workers, which is also the case in our cell phone data. Instead, the proportion of types of workers in each origin i is sometimes available from population census or surveys (e.g. education level at each census grid). We show that we can still back out w_j^H and w_j^L using this information.

For simplicity, assume that $k_{ij} = 1$ for all i and j; the argument goes through straightforwardly without this assumption. Denote the number of type L and H workers residing in i by y_i^L and y_i^H, correspondingly. Then, the commuting flow from i to j can be written as

$$y_{ij} = \frac{\left(w_j^L / d_{ij}\right)^{\epsilon^L}}{\sum_s (w_s^L / d_{is})^{\epsilon^L}} y_i^L + \frac{\left(w_j^H / d_{ij}\right)^{\epsilon^H}}{\sum_s (w_s^H / d_{is})^{\epsilon^H}} y_i^H.$$

Note that this expression does not lead to linear model of a gravity equation by taking logarithm any more. Still, the parameters in this model can be estimated with nonlinear

estimation methods such as maximum likelihood. Here, we heuristically argue that the identification of the parameters is also achieved.

Consider two adjacent origins i and i' such that $d_{ij} \approx d_{i'j}$ for all j. Using the above expression,

$$\frac{y_{ij} - y_{i'j}\frac{y_i^L}{y_{i'}^L}}{y_i^H - y_{i'}^H\frac{y_i^L}{y_{i'}^L}} \approx \frac{\left(w_j^H/d_{ij}\right)^{\epsilon^H}}{\sum_s (w_s^H/d_{is})^{\epsilon^H}}.$$

The expression on the right hand side is exactly the same as the one in the gravity equation after taking logarithm. Hence, $\{w_j^H\}$ and the distance coefficient for type H are identified. Similar argument goes for type L.

Appendix B. Data and Measurement

B.1. Commuting Flows using Home-Work Categorization

In this section we describe the procedure to construct commuting flows using the home-work categorization. To identify the "home" location of a particular user, we consider all locations between hours 21:00 and 05:00. For each such night interval, we record the towers that appear at least once that night. The "home" location is the tower that appears on most nights. For the "work" location we restrict to non-holiday weekdays, and consider all locations between hours 10:00 and 15:00. We then proceed similarly: for each day we record the towers that appear at least once, and the "work" location is the tower that appears on most days.

To compute commuting flows, we restrict to users for whom we have identified both a "home" and a "work" location, who account for approximately 93% of all users.

	(1)	(2)	(3)	(4)	(5)
	Nighttime lights (DMSP)				
log(mean income)	16.152***	2.529***	2.302***	2.020***	2.584***
	(1.218)	(0.756)	(0.747)	(0.761)	(0.795)
log(tower cell size)		-5.704***	-5.769***	-7.002***	-5.551***
		(0.434)	(0.439)	(0.564)	(0.460)
log(population density)		3.272***	3.201***	2.316***	3.580***
		(0.482)	(0.479)	(0.485)	(0.572)
$\log(d_{\text{capital}})$		-0.013	-0.009	-0.012	-0.011
		(0.011)	(0.011)	(0.011)	(0.011)
Destination FE			-1.111		
			(0.922)		
log(inflow/outflow)				-9.707**	
				(4.491)	
log(inflow)					-1.707
					(1.063)
Observations	$\sim 10^3$ - 10^4	$\sim 10^3$ - 10^4	$\sim 10^3$ - 10^4	$\sim 10^3$ - 10^4	$\sim 10^3$ - 10^4
R-squared	0.670	0.845	0.846	0.849	0.847

Table 4 Model-predicted Residential Income and DMSP Nighttime Lights

Notes: See notes for Table 3. The dependent variable is mean DMSP nighttime lights as in (Chen & Nordhaus, 2011; Henderson et al., 2010; Mellander et al., 2013); see section 5.1.2 for details.

	(1)	(2)	(3)	(4)
	Log nighttime lights (VIIRS)			
log(mean income)	0.110***	0.094***	0.114***	0.077***
	(0.027)	(0.025)	(0.025)	(0.022)
log(tower cell size)	-0.328***	-0.326***	-0.294***	-0.358***
	(0.010)	(0.012)	(0.023)	(0.008)
log(population density)	0.174***	0.175***	0.199***	0.135***
	(0.029)	(0.024)	(0.025)	(0.023)
$\log(d_{\text{capital}})$	-0.000	-0.001	-0.000	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Destination FE		0.342***		
		(0.095)		
log(inflow/outflow)			0.205***	
			(0.048)	
log(inflow)				0.242***
				(0.040)
District Fixed Effects	Yes	Yes	Yes	Yes
Observations	$\sim 10^3$ - 10^4	$\sim 10^3$ - 10^4	$\sim 10^3$ - 10^4	$\sim 10^3$ - 10^4
R-squared	0.95	0.96	0.95	0.96

Table 5 Model-predicted Residential Income,
using the Home-Work Categorization, and Nighttime Lights

Notes: See notes for Table 3. The income measure here is constructed by estimating the model using commuting flows constructed using the Home-Work categorization; see appendix B.1.

B.2. Representativeness of Cell Phone Data

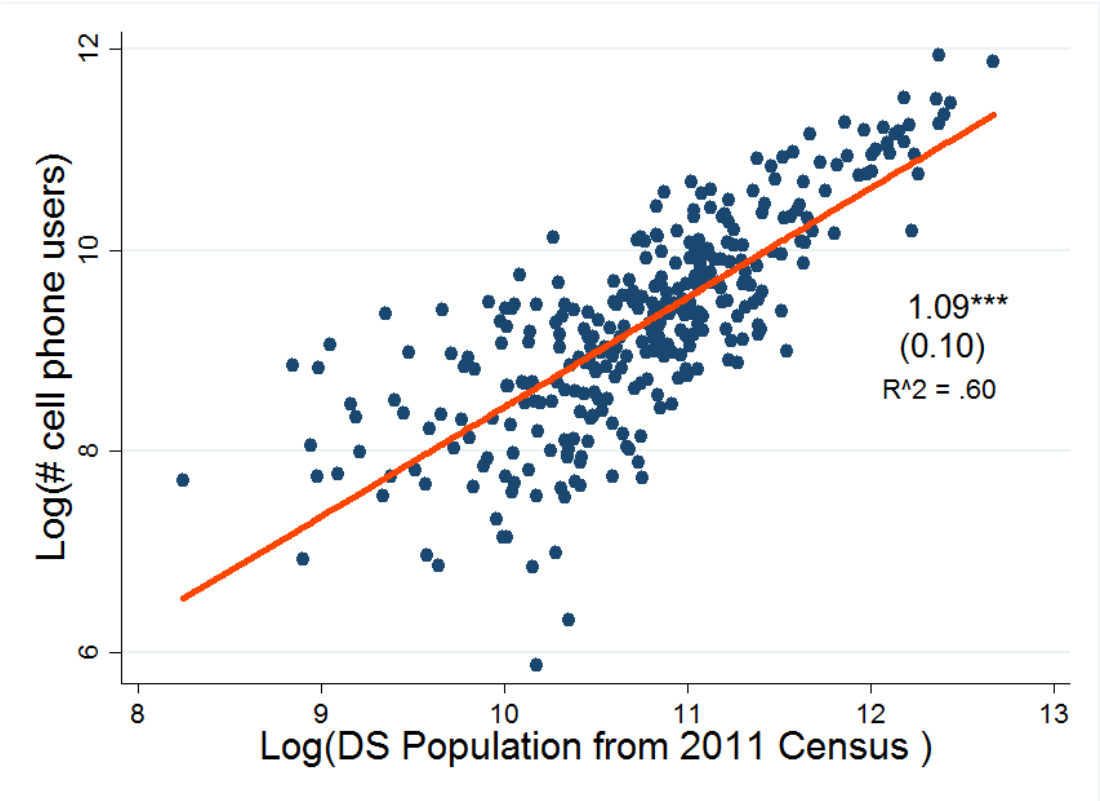


Figure 4. Number of Cell Phone Users and Census Population

Note: This graph plots the number of cell phone users whose home is categorized in each Divisional Secretariats (DS) against the census population of that DS. Appendix B.1 explains the home categorization in more detail.