Measuring Commuting and Economic Activity inside Cities with Cell Phone Records*

Gabriel E. Kreindler†

Yuhei Miyauchi‡

February 21, 2019

JEL Codes: C55, E24, R14

Abstract

We show that commuting flows constructed from cell phone transaction data predict the spatial distribution of wages and income in cities. In a simple workplace choice model, commuting flows follow a gravity equation whose destination fixed effects correspond to wages. We use cell phone data from Dhaka and Colombo, covering hundreds of millions of commuter-day observations, to invert this relationship. Model-predicted income at the workplace level predicts self-reported survey workplace income, and model-predicted residential income predicts nighttime lights. In an application, we estimate that predicted commuter income is 4-5% lower on days with hartals (transportation strikes) in Dhaka.

*The authors are grateful to the LIRNEasia organization for providing access to Sri Lanka cell phone data, and especially to Sriganesh Lokanathan, Senior Research Manager at LIRNEasia. The authors area also grateful to Ryosuke Shibasaki for navigating us through the cell phone data in Bangladesh, to Anisur Rahman and Takashi Hiramatsu for the access to the DHUTS survey data, and International Growth Center (IGC) Bangladesh for hartals data. The cell phone data for Bangladesh is prepared by the Asian Development Bank for the project (A-8074REG: "Applying Remote Sensing Technology in River Basin Management"), a joint initiative between ADB and the University of Tokyo. We are grateful to Akira Matsushita and Lauren Li, who provided excellent research assistance. We sincerely thank David Atkin, Alexander Bartik, Abhijit Banerjee, Sam Bazzi, Arnaud Costinot, Dave Donaldson, Esther Duflo, Gilles Duranton, Ed Glaeser, Seema Jayachandran, Sriganesh Lokanathan, Danaja Maldeniya, Ben Olken, Steve Redding, members of the LIRNEasia BD4D team, and seminar participants at MIT, LIRNEasia, NEUDC 2016, the Harvard Urban Development Mini-Conference, and the ADB Urban Development and Economics Conference, for constructive comments and feedback. We thank Dedunu Dhananjaya, Danaja Maldeniya, Laleema Senanayake, Nisansa de Silva, and Thushan Dodanwala for help with Hadoop code and GIS data in Sri Lanka. We gratefully acknowledge funding from the International Development Research Centre (IDRC) and The Weiss Fund for the analysis of Sri Lanka data, and from the International Growth Center (IGC) for the analysis of Bangladesh data. We also acknowledge Darin Christensen and Thiemo Fetzer's R code to compute Conley standard errors (`http://www.trfetzer.com/using-r-to-estimate-spatial-hac-errors-per-conley/`), on which we built our code.

†University of Chicago. Email: gekr@uchicago.edu

‡Stanford University, Asia-Pacific Research Center. Email: miyauchi@stanford.edu

1 Introduction

Measures of urban economic activity at fine temporal and spatial scales are important yet scarce. Such data is necessary to understand how cities respond to localized shocks such as changes in transportation infrastructure or floods, and to help governments target scarce public resources. These issues are especially salient in large cities in developing countries, which are growing fast yet are least covered by conventional data sources. For example, less than 10% of the urban population in sub-saharan African countries is covered by a census of firms with wage data.[1] At the same time, comprehensive new data sources on urban behavior, especially individual mobility, are becoming available across the world.

In this paper, we provide a theory-based method to predict the spatial distribution of urban economic activity from commuting choices.[2] The revealed-preference logic of our approach is simple. A core function of cities is to connect workers and jobs. While many factors enter into workplace choice decisions, areas with high wages should disproportionately attract workers, keeping distance and home locations fixed. We propose inverting this reasoning to infer the relative average wage at a location based on how "attractive" it is as a commuting destination. We use tools from urban and trade models to formalize this intuition. In the model, work location decisions aggregate up to a gravity equation on commuting flows, and destination fixed effects are proportional to log wages. This property holds for a general class of urban models developed to evaluate urban policies and transport infrastructure (e.g., Ahlfeldt et al. 2015; Heblich et al. 2018; Tsivanidis 2018; Severen 2019).

We implement our approach using call detail record (CDR) data from two large metropolises: Colombo, Sri Lanka and Dhaka, Bangladesh. CDR data is a prototypical example of "big data" available in developing countries, and it contains phone user location for every transaction (phone call or text message). We construct individual commuting trips by observing a user's location at different times during a single day, which leads to almost half a billion days with commuting information. We show that commuting flows constructed this way correlate strongly with commuting flows from a transportation survey from Dhaka, while additionally offering very fine geographic resolution and daily time variation. We pool the data over time and estimate the gravity equation implied by the model. We return to the high-frequency temporal aspect of the data in an application below.

Estimated wages by location are derived solely from observed commuting decisions and

[1] Authors' calculation (Appendix A.7). Moreover, government statistics may sometimes be unreliable. For example, Nigeria and Japan recently made consequential revisions to jobs and output data (Financial Times, 2014; Japan Times, 2019).

[2] Several papers use big data sources to empirically predict economic indicators (Blumenstock et al. 2015; Jean et al. 2016; Glaeser et al. 2017). Our approach is to use economic theory to guide how we interpret the data.

data on travel times, without any model training with actual wage data. We next assess how well this simple measure captures real differences in wages, using two income proxy data sources. First, we compare average model-predicted *workplace* income with commuter income data from a large transportation survey covering 1% of the population of Dhaka. Second, we compare in each city model-predicted average *residential* income with high resolution satellite nighttime lights, an intuitive proxy for residential income.

Model workplace income is significantly positively correlated with survey workplace income, including after controlling for location employment density and distance to the central business district (CBD). These results hold robustly when repeating the exercise with residual log income, after projecting out demographic and occupation covariates from the survey. Hence, commuting choices encode valuable information on which urban areas are productive.

At the same time, model-predicted income only explains around a quarter of the spatial variation in survey income. These results are related to a contemporaneous conversation on the ability of structural urban models to fit urban data. Interestingly, Severen (2019) finds that model wages estimated using census tract commuting flows barely predict tract-level wages in Los Angeles; this result may be related to the small size of the spatial unit, as well as non-wage factors determining commuting choices. Tsivanidis (2018) calibrates and estimates a general equilibrium model in Bogotá, and finds that model-predicted wages across 19 urban areas predict survey wages. By comparison, our focus in this paper is using a parsimonious model together with detailed commuting data to predict relative income.

In the second validation exercise, in both cities, model-predicted residential income is a robust predictor of satellite nighttime lights. This relationship remains stable within sub-districts and after controlling for residential density and distance to the CBD. The explanatory power is significantly higher (R^2 over 0.8), in part due to the inclusion of peri-urban areas in the analysis.

A key advantage of the model is that we can compute how income "moves" around the city. In both validation exercises, we perform a horse-race between residential- and workplace-income. While the two measures are highly correlated, we find suggestive evidence that model *workplace* income better correlates with *workplace* survey income data, and model *residential* income better correlates with nighttime lights.

The ideal application of our income-prediction method and of the high-frequency commuting data is to trace out the spatial and temporal impact of urban events and policies. To illustrate this potential, we estimate the economic cost of *hartals*, a type of strike intended to disrupt transportation and economic activity in Bangladesh. The daily commuting data exhibits visible differences between hartal and workdays, and model-predicted income falls by around 4-5% on hartal days. This effect is driven mostly by the extensive margin (fewer trips), and,

to a lower extent, by larger effects for high-income commuters. While precisely estimated, these changes are relatively small; these results are in line with previous studies of hartal.

In the conclusion, we revisit how our project fits into a larger approach of using economic theory together with the rich *choice* information made available by "big data."

2 Cell-Phone Data and Commuting Flows

2.1 Data Sources

Cell phone transaction data. We use call detail record (CDR) data from large operators in Sri Lanka and Bangladesh to compute detailed commuting matrices. CDR data includes an observation for each transaction, such as outgoing or incoming voice call and text messages, or GPRS internet connections. Each observation has a timestamp, the anonymized participant user identifiers, and their cell tower locations. Towers are unevenly distributed in space; they are denser in urban and developed areas. We focus on the greater metropolitan areas around the capital cities of Colombo and Dhaka. The data covers a little over a year in Sri Lanka and four months in Bangladesh in the early 2010's.[3]

We construct commuting trips by observing a phone-user connect to towers at different times of the day. On a given day, we define a user's *origin* as the location of the first transaction between 5am to 10am, and the user's *destination* as the location of the last transaction between 10am and 3pm.[4] By definition, a user has at most one commuting trip per day. If the origin and destination correspond to the same cell tower, we say that the user was stationary, otherwise the user made a proper commuting trip. If transaction data is missing in either time interval, commuting behavior is not observed for that user-day.[5] We then aggregate over users and non-holiday weekdays to obtain an origin-destination (OD) matrix of commuting flows between every pair of cell towers.

One potential concern is that cell phone data is not representative of urban commuters. Cell phone ownership is high in both countries, but the ownership may vary systematically by demographic characteristics. A separate concern is that calling behavior may be correlated with mobility. To address these issues, in section 2.2 we compare the commuting flows constructed here with a representative transportation survey.

[3]In Bangladesh, the data only covers outgoing voice calls. Our sample covers the Western Province in Sri Lanka, and the Dhaka, Narayanganj, and Gazipur Districts in Bangladesh.

[4]We focus on the morning commute, as other types of trips (e.g., shopping) are more likely in the evening (Frank and Murtha, 2010).

[5]Commuting data (including stationary trips) is available for 16% (in Dhaka) and 29% (in Colombo) of the theoretical maximum number of user-days, that is, if we observed each user on every day in the sample (Table C.1).

For robustness, we also construct commuting flows using individual-level "home" and "work" locations, defined as the modal locations during weekday nighttime and daytime periods, respectively, using all data for an individual. The flows using the two methods are strongly correlated and all our validation results are robust to this choice (Appendix Figure B.3, panel B).

Google Maps travel time. As a proxy for travel costs, we obtain estimated typical driving travel times between pairs of cell towers using the Google Maps API. Because of the large number of bilateral pairs, in each city we obtain Google data for 90,000 randomly selected pairs of towers, and interpolate to pairs with nearby origin and nearby destination.[6]

Household transportation survey. We use individual survey data from the 2009 Dhaka Urban Transport Network Development Study or DHUTS (JICA 2010). The survey covers 16,394 randomly selected households in the Dhaka City Corporation (DCC), Dhaka's urban core, as well as a sample of 1,716 households outside the DCC. Home and work locations are recorded at the level of 108 "survey areas." Our main analysis sample consists of 12,510 commuters who live and work within the 90 survey areas inside the DCC and who report positive income from work, excluding students, homemakers, and the unemployed. In the main analysis, we exclude households outside of DCC, because the 18 corresponding survey areas are significantly coarser and detailed information on sampling is not available. Our results are robust to including all commuters who live and work inside the DCC (Appendix Table C.5).

2.2 Cell Phone Data Captures Aggregate Commuting Behavior Accurately

In Dhaka, commuting flows derived from cell phone data are strongly related to those from the DHUTS commuting survey. To show this, we aggregate the cell phone commuting data up to the level of survey areas. The sample consists of 7,915 pairs of distinct survey areas with positive cell phone commuting flows, with a total of 12,510 trips in the DHUTS survey (this includes government workers) and over 18 million trips between 1.5 million tower pairs in the cell phone data. Commuting flows from the two data sources are strongly related, including when we control for log travel time, origin and destination survey area fixed effects (Appendix Table C.2), consistent with previous research validating cell-phone-based commuting flows (Calabrese et al., 2011; Wang et al., 2012; Iqbal et al., 2014). In addition, the decay of commuting flows with travel time is virtually identical between the two data

[6] Appendix A.6 describes the interpolation procedure in detail. We collected data for Sri Lanka in 2016 and for Bangladesh in 2017. We extracted travel time without traffic congestion (Google did not provide travel time with traffic congestion in Bangladesh in 2017).

sources (Appendix Figure B.3, Panel A). Commuting flows constructed using "home" and "work" locations also decay at the same rate with log travel time (Panel B).

Residential population density from from cell phone data is also strongly correlated with population density from the census (Appendix Table C.3). The adjusted R-squared is 0.61 in Dhaka and 0.49 in Colombo. The slope is 1.16 for both cities, hence cell phone data slightly over-represents population in denser areas.

3 Model: Commuting Flows, Gravity, and Wages

Is it possible to infer the spatial distribution of wages from commuting flows? The interaction between wages and commuting costs to determine urban structure is fundamental in classical urban economics models (Alonso, 1960; Mills, 1967; Muth, 1968). Here, we explore this insight using a new generation of models inspired from the trade literature, designed to better match spatially disaggregated urban data (Ahlfeldt et al., 2015).

In the model, commuters decide their work location taking into account wages at different potential work locations, commuting costs, and destination-specific idiosyncratic utility shock. Together with a parametric assumption on utility shocks, this implies that log bilateral commuting flows follow a linear gravity equation, with destination fixed effects capturing log wages. Furthermore, this relationship holds in equilibrium regardless of how wages are determined.

3.1 Workplace Choice Model

Space is partitioned into a finite set of locations L, which may serve as both residential locations and work locations. In our application, these correspond to Voronoi cells around cell phone towers (depicted in Appendix Figures B.1, B.2 for the two cities).

There is a unit mass of workers, and each worker ω sequentially decides where to live, and then where to work. We do not impose any restrictions on the home location choice. (Assuming joint home and work location choice leads to the same gravity equation (Ahlfeldt et al., 2015).) Given her residential location (or origin) i, the worker chooses her work location (or destination) j. The utility of worker ω residing in location i if she chooses destination j is:

$$U_{ij\omega} = \frac{W_j Z_{ij\omega}}{D_{ij}^{\tau}} \tag{1}$$

W_j is the wage per effective unit of labor supply at location j (all firms at location j offer the same wage), D_{ij} is the travel time between i and j, and $Z_{ij\omega}$ is an idiosyncratic utility shock that is i.i.d. following the Fréchet distribution, with scale parameter T and shape parameter

ϵ. In our baseline model, we assume that each worker supplies one unit of labor, and hence earns income W_j if she works in location j. In particular, we abstract from heterogeneity due to skill or other worker attributes.[7]

Each worker observes the shocks $Z_{ij\omega}$ and chooses the work location j where $U_{ij\omega}$ is maximized. The probability that a worker commutes to j conditional on residing in i is given by $\pi_{ij} = (W_j/D_{ij}^{\tau})^{\epsilon}/\sum_s (W_s/D_{is}^{\tau})^{\epsilon}$. Taking logs, and denoting log quantities by lowercase letters:

$$\log(\pi_{ij}) = \epsilon w_j - \epsilon\tau d_{ij} - \log\left(\sum_s \exp\left(\epsilon w_s - \epsilon\tau d_{is}\right)\right) \tag{2}$$

3.2 Estimating the Gravity Equation

We estimate equation (2) through the following empirical gravity model:

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

where μ_i and ψ_j are origin and destination fixed effects, and ε_{ij} accounts for measurement error or other unmodeled factors. Most importantly, ψ_j is proportional to the (relative) log wage at j with a factor of ϵ, the Fréchet dispersion parameter. (As usual in discrete choice models, we can only identify wages up to scale.) Our main goal is to recover the ψ_j's from observed commuting choices. Note that, for this purpose, it is not necessary to model explicitly how wages are determined in equilibrium. In other words, this mapping between commuting choices and wages holds in any general equilibrium model that micro-founds the gravity equation for commuting flows with a discrete commuting choice model.

We implement two approaches to estimate ψ_j. First, we simply estimate (3) by OLS. Second, we estimate the equation imposing the formula for μ_i given by the structural gravity equation (2).[8] The two approaches yield very similar results (Section 3.4).

Lacking detailed bilateral commuting flow data, some authors estimate log wages with an exactly identified procedure using residential and employment populations, and calibrated or separately estimated parameters (Ahlfeldt et al., 2015; Tsivanidis, 2018). Our approach

[7] We model and investigate empirically two extensions where labor supply varies across individuals. First, in Appendix A.2, labor supply (and hence income) depends on observable demographics. Second, in Appendix A.3, $Z_{ij\omega}$ and D_{ij} partly affect labor supply, rather than only affecting utility, as in the main analysis. We develop a method to estimate how much $Z_{ij\omega}$ and D_{ij} affect income using survey income data. The results are consistent with D_{ij} being a pure utility shock, and $Z_{ij\omega}$ partly affecting income (Appendix Table C.11).

[8] Specifically, we estimate (3) following an iterative procedure (Appendix A.1). In each iteration, we estimate (3) without origin fixed effects, after subtracting the model origin terms from the previous iteration from the left-hand side. We iterate until the vector of destination fixed effects converges. The procedure is identical to SILS (structurally iterated least squares) proposed in the trade gravity literature (Head and Mayer, 2014), except without destination fixed effect constraints. See Fally (2015) for potential bias without imposing model constraints.

using commuting flows is more robust against noise in the gravity equation (3), and allows us to test the stability of our results (Appendix A.4).

3.3 Mapping Model Locations to Geographic Areas

A key advantage of the model is that model locations can be mapped directly to two-dimensional urban data. However, since in the model productivity shocks are independent across locations, the choice of location units matters. Empirically, larger Voronoi cells may mechanically yield larger destination fixed effects.

We show that when the true model has multiple independent shocks at sub-locations within a given location, commuting flows defined at the larger location level still follow gravity equation (2) using an "effective" wage at that location. In particular, assume that location j is divided into N_j smaller areas where workers draw independent shocks, and all areas have the same "true" wage W_j^R. By standard Fréchet properties, the commuting probability to j is equivalent to a model with a single shock at j and "effective" wage $W_j = N_j^{1/\epsilon} W_j^R$.[9] From equation (3) we estimate $\psi_j = \epsilon \log W_j$ and we recover the true wage as the *area-adjusted* destination fixed effect

$$\hat{\psi}_j^R = \hat{\psi}_j - \log(N_j). \tag{4}$$

In robustness exercises, using un-adjusted destination fixed effects does not affect results, except when including distant peri-urban areas where cell phone towers are very sparse.

3.4 Estimation Results: Gravity and Wages

We estimate gravity equation (3) using cell phone commuting flows and Google Maps travel times. Our goal is to recover the destination fixed effects, which in the model are proportional to workplace log wages. The estimation sample is non-holiday weekday commuting trips between pairs of towers excluding nearby and very distant towers.[10]

Table 1 reports the results, based on almost 20 million commuting flows between $\sim 1,900$ locations in Dhaka (columns 1-2) and 130 million flows between $\sim 1,200$ locations in Colombo (columns 3-4). The gravity equation is estimated with unconstrained OLS (columns 1 and 3) and by imposing model constraints (columns 2 and 4). See footnote 8 and Appendix A.1

[9] In Appendix A.5 we prove a more general approximate invariance-to-aggregation result for destination fixed effects. Redding and Weinstein (2019) prove a related result for gravity models in trade.

[10] In Dhaka, we further exclude 31 days with transportation strikes (hartals). Tower pairs closer than 3 minutes are excluded as they may capture calls randomly connecting to different towers ("tower-bouncing") rather than real commuting. Destination fixed effects estimated including nearby tower pairs are virtually identical (Appendix Table C.4). Towers over the 99th percentile of the travel time distribution are also excluded (137 and 96 minutes in Dhaka and Colombo, respectively).

for estimation details for the latter.

Commuting probability decreases strongly with travel time. Interestingly, although the average commuting trip is 25% longer on average in Sri Lanka, once we adjust for residential locations (i.e., gravity equation with origin fixed effects), the coefficients for Dhaka and Colombo are very similar, -1.65 and -1.76. This is a substantive finding, as the two cities differ in terms of economic development, population, and urban structure (mono- vs poly-centric).

Figure 1 displays smoothed estimated wages in Dhaka and Colombo using choropleth maps. Estimated wages are higher near city centers and alongside some (but not all) major road corridors. Moreover, secondary centers are visible, especially in Dhaka. The next sections will compare these results with independent income proxies.

Destination fixed effects using different estimation methods are highly correlated, including when we add 1 to tower pairs with zero commuting flows (57% of all possible tower pairs in Bangladesh and 15% in Sri Lanka) (Appendix Table C.4).[11]

For the rest of the paper, we use the destination fixed effects from the first estimation method of the gravity equation (columns 1 and 3) as model-predicted log wages (before area adjustment), and we report robustness results in the appendix.

4 Validation using Survey Income and Nighttime Lights

The method above infers wages based on observed commuting choices. Does this approach predict real-world within-city income patterns? We now cross-validate with two alternate income data sources: self-reported income from a transportation survey in Dhaka, and high-resolution satellite nighttime lights data in Dhaka and Colombo.

4.1 Model-Predicted and Survey Workplace Income in Dhaka

Our first validation exercise compares income from the model and survey income from the DHUTS survey (Section 2.1).We compute average income at the workplace level in each survey areas in the DCC, the finest geographic location available in the DHUTS survey.

The model-predicted income measure is the area adjusted destination fixed effects $\hat{\psi}_j^R$. In the model, this equals log labor income divided by ϵ, the Fréchet shape parameter of worker's unobserved preferences. Hence, we expect a regression coefficient of around $1/\epsilon$. Since survey areas are coarser than cell phone towers, we average model income within each of the 88 survey areas with non-government workers, weighting each tower by its workplace population from the cell phone data.

[11] Implementing the Poisson pseudo-maximum-likelihood estimator from Silva and Tenreyro (2006) to deal with zero flows is difficult due to the large number of locations.

In our main exercise, we correlate survey income and model-predicted income, at the workplace survey area location level. This is a transparent way to check how our measure – which is obtained solely from commuting choices and data on travel times – lines up with real data.

We benchmark results in two ways. First, we consider two alternative variables to check that our results are not driven by simpler measures: employment density (computed from cell phone data), an established empirical predictor of income, and distance to the Central Business District (CBD), the main dimension of variation in monocentric urban models. Second, we repeat the entire exercise replacing log income with the residual after partialing out demographic and job variables from the survey (age, gender, years of education, occupation and job sector). We are interested whether predicted income from the model with ex-ante homogenous workers is confounded by observable worker characteristics.[12]

Table 2 presents the main results. Given that government jobs are typically paid less and are centrally located yet not market allocated, our estimation sample excludes government workers. Including them significantly weakens the correlation between model and survey income, unless we include occupation dummies. Model-predicted income explains 26 percent of the variation in average income at the survey area level, and the coefficient implies a Fréchet shape parameter of $\hat{\epsilon} = 7.1$, similar to estimates in the urban economics literature (6.83 in Ahlfeldt et al., 2015). In columns 2-3, employment density and distance have slightly lower and slightly higher predictive power, respectively.[13] The coefficient on model-predicted income is almost unchanged when controlling for these variables (column 4), showing that the model contains information not available in these other measures. In column 5, we include model-predicted *residential* income. While the two model measures are highly correlated, the positive correlation with survey *workplace* income is loaded onto model *workplace* income. The coefficient on residential income is negative and significant, yet less precisely estimated.

Results are broadly similar with *residual* income as the dependent variable (panel B). All coefficients are roughly half the size of those in Panel A, which suggests that part of the variation in average income captured by any of the three measures is due to sorting. However, model workplace income remains significant in all specifications, and the explanatory power is similar to Panel A. Hence, our revealed-preference approach has similar predictive power after controlling for workplace sorting along observable worker characteristics.

Results are robust to several alternate specifications of the gravity equation estimation

[12]This specification is exact in a model where income heterogeneity due to demographic factors enters multiplicatively in equation (1) (Appendix A.2).

[13]Note that averaging within relatively coarse geographic areas favors the distance to CBD measure. Indeed, when averaging, while the range of distance to CBD remains roughly unchanged, the variance of average model-predicted income goes down, which tends to decrease R^2.

(Appendix Table C.5). Appendix Table C.6 uses individual survey data and shows that our main result is robust to controlling for origin survey area fixed effects, geographic area of destination location, and travel time.

In terms of predictive power, our measures capture about one-quarter of the spatial variation of surveyed income. This suggests that factors outside the model are also important. In particular, distance to CBD has a significant, negative coefficient even after controlling for model income (Table 2, column 4). This may happen if we underestimate travel costs towards the CBD in our gravity equation, for example if traffic jams are valued beyond the time lost, or if locations near the CBD exhibit workplace disamenities, so that higher wages are required to compensate.

By comparison, Tsivanidis (2018) uses a similar model estimated with population counts and finds $R^2 \approx 0.3$ between model-predicted wages and independent survey wage data in 19 urban zones in Bogotá, Colombia. Severen (2019) performs a similar exercise in Los Angeles on a model estimated with commuting flows at the census tract level and finds no correlation with real wage data, citing unobserved location-pair-specific factors as a potential explanation.

4.2 Model-Predicted Residential Income and Nighttime Lights

In our second validation, we use high-resolution nighttime light satellite data (Visible Infrared Imaging Radiometer Suite; VIIRS) to assess the model's performance in Dhaka and Colombo.[14] Nighttime lights are an established income proxy in data-scarce environments (Henderson et al., 2010; Chen and Nordhaus, 2011), and the high resolution of VIIRS data (approximately 500 meters) makes it ideal for studying urban areas.

We use the model to predict *residential* (take-home) income at the cell tower level. We test the assumption that nightlights measure residential income by also including model *workplace* income in our regressions. We also include two alternate variables: residential population density, and distance to the CBD.

Model residential income is strongly related to nightlights at the cell tower level (Table 3). The R^2 is high at 0.71 and 0.86 in Dhaka and Colombo. Residential density and distance to CBD are also highly correlated with nightlights, and they explain smaller shares of the variance.

The model performs well at fine spatial resolution. The coefficient on model-predicted income remains large when including sub-district fixed effects (55 units in Dhaka and 42 units in Colombo), and when controlling for residential density, distance to CBD, and

[14] We use a monthly cloud-free composite (January 2014) based on nights with zero moonlight and areas without clouds, curated by the Earth Observation Group (EOG).

model-predicted *workplace* income (column 4 in Table 3). In particular, the model-predicted *residential* income is robustly correlated, while *workplace* income is not statistically significant. Together with results in Table 2, this suggests that the model picks up real patterns of how income "moves" around the city.

Our results are robust to using the gravity equation estimated with model origin terms, using home-work commuting flows, constructing residential income excluding nearby destination towers, and using log mean residential income (instead of mean log income) (Appendix Table C.7). In Dhaka not area-adjusting destination fixed effects reverses the sign of the correlation with nightlights, because of very large cells far away from the city center. Our results are broadly robust to only using data from the urban core of Dhaka (the area covered by the transport survey used in Section 4.1), except that the slope is shallower (Appendix Table C.8).

We also repeat the residential income validation exercise using survey income. We do not find any correlation between model income and survey income (Appendix Table C.9, Panel A). This may be due to lower underlying differences in average income at the *residential* level (compared to the workplace level), and hence a more noisy measure when using survey data. Consistent with this idea, density and distance to CBD also explain very little variation in average residential survey income.

5 Application: The Economic Costs of *Hartal*

We now illustrate how high-frequency commuting data and our detailed model-predicted income measure can be used in an application.

Hartals are a form of political strike that involves a partial shutdown of urban transportation and businesses. They are common in South Asia, and especially in Bangladesh (UNDP 2005). On hartal days, typically announced a few days in advance, groups of people (some paid) enforce the transportation shutdown, especially on major roads and in certain locations.

We use our data and model estimates to quantify the short-term impact of hartal on forgone income. We analyze how predicted income and travel patterns differ on hartal days compared to other days. We focus on the overall effect of hartals on commuting behavior, inclusive of potential changes in traffic congestion and commuting routes. Note, our empirical strategy cannot quantify direct impacts of hartals on worker productivity, nor long-term adaptation costs.

We use daily individual commuting data from cell phone records. The sample covers commuters with distinct long-term home and work locations (towers) identified with the

procedure in Section 2.1, accounting for 27% of all users in the data.[15] We only observe travel behavior if a user makes calls on a given day, and call behavior itself may differ on hartal days. We include commuter fixed effects to ensure that our results are not driven by selection across different types of commuters. Moreover, restricting to frequent callers yields very similar results (Appendix Table C.10).

Given that users may travel to different destinations on different days, we use two definitions of predicted income. First, we assign the (model-predicted) wage of the worker's destination j on that particular day, as long as she makes a proper trip (a trip with distinct origin and destination). We assign zero income to stationary (non-proper) trips. This is a way to capture that workers may earn income from different destinations on different days. Second, we use the wage from her long-term destination if the worker makes a proper trip to that location. We assign zero income to stationary trips and trips to any other destination. In both cases, predicted income is $\exp\left(\hat{\psi}_j^R/\epsilon\right)$ where log wages $\hat{\psi}_j^R$ are estimated using only non-hartal weekdays, and $\epsilon = 6.4$.[16] Since we are interested in *changes* in income during hartal days, the scale of $\hat{\psi}_j^R$, which is not identified, does not matter for this exercise.

Our main specification is:

$$y_{ct} = \beta^H Hartal_t + \beta^F Friday_t + \beta^S Saturday_t + \beta^{Ho} Holiday_t + \mu_c + \gamma_{Month(t)} + \varepsilon_{ct}$$

where c denotes a commuter, t denotes a calendar date, and the outcome y_{ct} is predicted income or a dummy for making a proper trip. The unit of observation is a commuter times calendar date, for dates when we observe a trip (either proper and stationary). The main coefficient of interest is β^H, which measures the difference in outcome on hartal days relative to workdays (non-holiday weekdays). We use β^F, the effect on Fridays, the main free day in Bangladesh, as a benchmark.

We use the data set of hartal dates in Dhaka from Ahsan and Iqbal (2015). They identify 33 hartal days over the 4 months in our sample. The study period preceded parliamentary elections and was marked by general instability, and hence hartals were more frequent than in previous years. Hence, our results may not directly generalize to periods with lower hartal intensity.

Cell phone data picks up stark differences in behavior on hartal, weekends and holidays. Appendix Figure B.6 plots the change in predicted income relative to workdays, by calendar date. Predicted income is systematically lower during hartals compared to weekdays, yet

[15] We are interested in canceled trips due to hartals, which are difficult to observe for users with identical home and work towers. Results with all users are qualitatively similar and smaller in magnitude.

[16] This is our point estimate of the Fréchet parameter using a structural estimation method (Appendix A.3). The regression coefficient from Table 2 of log survey income (measuring ω_j) on the destination fixed effect ($\epsilon\omega_j$) implies a very similar number ($\hat{\epsilon} = 7.1 = 0.14^{-1}$).

not as much as on Fridays and on some important holidays, such as the end of Ramadan. Predicted income is higher than usual on the two Fridays that immediately follow long hartal spells (December 6th and 20th), consistent with temporal substitution. The results are also consistent with longer hartals having lower impacts.

Table 4 shows the main results. To facilitate interpretation, all coefficients indicate proportional changes relative to the outcome mean on workdays. Odd columns assign predicted income to any proper trip, while in even columns outcomes are non-zero only when the commuter travels to her long-term workplace.

There are four key insights. First, commuters in Dhaka earn on average 4.4 to 4.8% less on hartal days compared to workdays. These effects are significantly smaller compared to Fridays, when predicted income is 20 to 45% lower on average (panel A). Second, hartal days affect "all trips" and "work trips" roughly equally, while on Fridays, work trips are disproportionately affected. This suggests a limited "destination selection" effect of hartals; on average, commuters do not switch to lower-income destinations. Third, the reduction in predicted income is driven primarily by the extensive margin, namely fewer trips (panel B). Fourth, commuters working in high-income destinations reduce trips relatively more. In columns 3-4, we fully interact the model with an indicator for commuters whose long-term workplace location is below median in the predicted wage distribution. The interaction results show that the proportional reduction in trips is concentrated among high-income commuters, both on hartal and Fridays. Columns 5-6 document that this heterogeneity is not due to commute distance (as long-distance commuters are also more affected).

These results show that commuters broadly succeed to maintain their workday travel routines on hartal days, which limits the short-term impact of hartal on economic activity. These results are consistent with previous studies on hartals in more specific settings (Ashraf et al., 2015; Ahsan and Iqbal, 2015).

6 Conclusion

This paper provides a theory-based toolkit for using cell phone data to understand the spatial distribution of economic activity in cities. This framework is especially suited to measuring and interpreting the short-term impact of urban shocks such as floods, or of transportation incidents or improvements, on commuting and economic activity. Together with officials statistics, they can be used to investigate spatial discrepancies between formal and informal economic activity.

Big data, such as cell phone or smartphone mobility records, credit card transactions, or user-generated reviews, are rapidly gaining popularity due to their ability to *predict* behavior,

individual characteristics and economic conditions (Blumenstock et al., 2015; Jean et al., 2016; Glaeser et al., 2017; Björkegren and Grissen, 2018).

However, big data also contain a wealth of information regarding individual *choices*. This allows economists to apply revealed preference techniques to infer attributes of choice options, such as workplace wages in our paper or spatial aspects of consumption behavior (Athey et al., 2018; Davis et al., 2018; Agarwal et al., 2018). We believe that this type of applications is a promising path for using "big data" in economics.

References

Agarwal, S., F. Monte, and B. Jensen (2018): "The Geography of Consumption," *NBER Working Paper No. 23616*.

Ahlfeldt, G. M., S. J. Redding, D. M. Sturm, and N. Wolf (2015): "The Economics of Density: Evidence from the Berlin Wall," *Econometrica*, 83, 2127–2189.

Ahsan, R. and K. Iqbal (2015): "Political Strikes and its Impact on Trade: Evidence from Bangladeshi Transaction-level Export Data," *IGC Working Paper*.

Alonso, W. (1960): "A Theory of the Urban Land Market," *Papers and Proceedings Regional Science Association*, 6, 149–157.

Ashraf, A., R. Macchiavello, A. Rabbani, and C. Woodruff (2015): "The Effect of Political and Labour Unrest on Productivity: Evidence from Bangladeshi Garments," *IGC Working Paper*.

Athey, S., D. Blei, R. Donnelly, F. Ruiz, and T. Schmidt (2018): "Estimating Heterogeneous Consumer Preferences for Restaurants and Travel Time Using Mobile Location Data," *AEA Papers and Proceedings*, 108, 64–67.

Björkegren, D. and D. Grissen (2018): "The Potential of Digital Credit to Bank the Poor," *AEA Papers and Proceedings*, 108, 68–71.

Blumenstock, J., G. Cadamuro, and R. On (2015): "Predicting Poverty and Wealth from Mobile Phone Metadata," *Science*, 350.

Calabrese, F., G. Di Lorenzo, L. Liu, and C. Ratti (2011): "Estimating Origin-Destination Flows Using Mobile Phone Location Data," *IEEE Pervasive Computing*, 10, 36–44.

Chen, X. and W. D. Nordhaus (2011): "Using luminosity data as a proxy for economic statistics." *Proceedings of the National Academy of Sciences of the United States of America*, 108, 8589–8594.

Davis, D., J. Dingel, J. Monras, and E. Morales (2018): "How Segregated is Urban Consumption?" *Accepted, Journal of Political Economy*.

Duncan, C. (2005): *Beyond Hartals: Towards Democratic Dialogue in Bangladesh*, United Nationas Development Programme.

Fally, T. (2015): "Structural gravity and fixed effects," *Journal of International Economics*, 97, 76–85.

Financial Times (2014): "Nigeria almost doubles GDP in recalculation," https://www.ft.com/content/70b594fe-bd94-11e3-a5ba-00144feabdc0.

Frank, P. and T. Murtha (2010): "Trips Underway by Time of Day by Travel Mode and Trip Purpose for Metropolitan Chicago," Tech. rep., CMAP Congestion Management Process, Chicago Metropolitan Agency for Planning.

Glaeser, E. L., H. Kim, and M. Luca (2017): "Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity," *Harvard Business School Working Paper, No. 18-022*.

Head, K. and T. Mayer (2014): "Gravity equations: Workhorse, toolkit, and cookbook," in *Handbook of international economics*, Elsevier, vol. 4, 131–195.

Heblich, S., S. Redding, and D. Sturm (2018): "The Making of the Modern Metropolis: Evidence from London," *Working Paper*.

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

Iqbal, M. S., C. F. Choudhury, P. Wang, and M. C. González (2014): "Development of Origin-destination Matrices Using Mobile Phone Call Data," *Transportation Research Part C: Emerging Technologies*, 40, 63–74.

Japan International Cooperation Agency (2010): "Preparatory Survey Report on Dhaka Urban Transport Network Development Study (DHUTS) in Bangladesh : Final Report." Tech. rep., Japan International Cooperation Agency, http://open_jicareport.jica.go.jp/pdf/11996774_03.pdf.

Japan Times (2019): "Government to review statistics after revelation that faulty Japan jobs data spanned 14 years," https://www.japantimes.co.jp/news/2019/01/11/national/government-review-statistics-revelation-faulty-japan-jobs-data-spanned-14-years/.

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon (2016): "Combining satellite imagery and machine learning to predict poverty," *Science*, 353, 790–794.

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, 197–210.

Muth, R. (1968): *Cities and Housing*, Chicago: University of Chicago Press.

Redding, S. and D. Weinstein (2019): "Aggregation and the Gravity Equation," *NBER Working Paper 25464*.

Severen, C. (2019): "Commuting, Labor, and Housing Market Effects of Mass Transportation: Welfare and Identification," *Working Paper*.

Silva, J. S. and S. Tenreyro (2006): "The log of gravity," *The Review of Economics and statistics*, 88, 641–658.

Tsivanidis, N. (2018): "The Aggregate And Distributional Effects Of Urban Transit Infrastructure: Evidence From Bogota's TransMilenio," *Working Paper*.

Wang, P., T. Hunter, A. M. Bayen, K. Schechtner, and M. C. González (2012): "Understanding Road Usage Patterns in Urban Areas," *Scientific Reports*, 2, 1001.

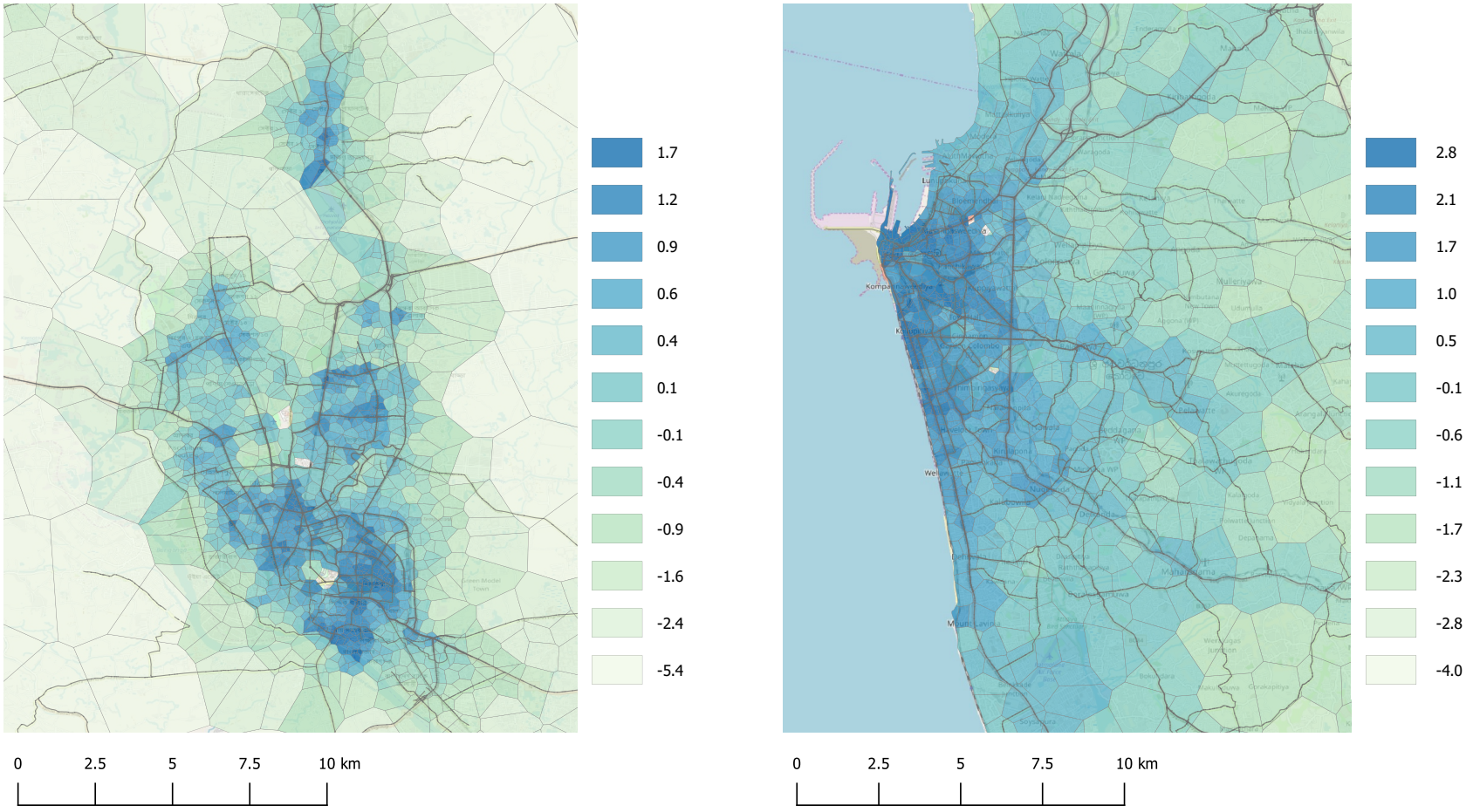
7 Figures and Tables

Table 1: Gravity Equation and Destination Fixed Effects

	log Commuting Flow			
	(1)	(2)	(3)	(4)
log Travel Time	-1.65^{***}	-1.93^{***}	-1.76^{***}	-2.30^{***}
	(0.01)	(0.01)	(0.03)	(0.02)
City	Dhaka	Dhaka	Colombo	Colombo
Model Constraints	No	Yes	No	Yes
Number of Destination FE	$1.9 \cdot 10^3$	$1.9 \cdot 10^3$	$1.2 \cdot 10^3$	$1.2 \cdot 10^3$
Number of Trips	$19.3 \cdot 10^6$	$19.3 \cdot 10^6$	$129 \cdot 10^6$	$129 \cdot 10^6$
Observations	$1.5 \cdot 10^6$	$1.5 \cdot 10^6$	$1.1 \cdot 10^6$	$1.1 \cdot 10^6$
Adjusted R^2	0.61	0.65	0.75	0.69

Notes. This table reports estimates of the gravity equation (3). The outcome variable is log total commuting flows $\log(V_{ij})$ between a pair of cell phone towers, computed from cell phone data and aggregated over weekdays. (This is equivalent to using log commuting probabilities $\log(\pi_{ij}) = \log(V_{ij}/\sum_s V_{is})$ as in equation (3), as the origin fixed effects capture the denominator.) In Bangladesh, we exclude hartal days. For each trip, the origin is the first location (tower) between 5 am and 10 am, and the destination is the last location between 10 am and 3 pm. Travel time between towers from the Google Maps API. The sample is all tower pairs with travel time between 180 seconds and the 99th percentile. Columns (1) and (3) report the coefficient $-\hat{\beta}$ from an OLS regression with origin and destination fixed effects. Columns (2) and (4) report the estimates from an iterative procedure, where at step $k+1$ the model-predicted origin term $\mu_i^k = \log\left(\sum_s \exp\left(\hat{\psi}_s^k - \hat{\beta}^k d_{is}\right)\right) + \log(\sum_s V_{is})$ is subtracted from log flow before running an OLS regression without origin fixed effects (Appendix A.1). Two-way clustered standard errors at the origin and destination level are reported in parentheses. $^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$

Figure 1: Estimated log Wages in Dhaka and Colombo



(A) Dhaka (B) Colombo

Notes. These figures plot area-adjusted destination fixed effects (the model measure proportional to log wages) at the level of cell phone tower Voronoi cells in Dhaka and Colombo. Log wages are kernel smoothed with an adaptive kernel bandwidth (proportional to the radius of the equivalent-area circle of the Voronoi cell, see Appendix A.6).

Table 2: Average Workplace Income: Model Predictions and Survey Data in Dhaka

	(1)	(2)	(3)	(4)	(5)
Panel A. Outcome: log Survey Income (workplace)					
log Model Income (workplace)	0.14***			0.12**	0.27***
	(0.03)			(0.05)	(0.06)
log Employment Density		0.15***		−0.05	−0.01
		(0.04)		(0.07)	(0.07)
log Dist. to CBD			−0.18***	−0.14***	−0.18***
			(0.03)	(0.03)	(0.03)
log Model Income (residential)					−0.32**
					(0.13)
Observations	88	88	88	88	88
Adjusted R^2	0.26	0.16	0.33	0.42	0.46
Panel B. Outcome: Residual log Survey Income (workplace)					
log Model Income (workplace)	0.07***			0.08**	0.15***
	(0.02)			(0.03)	(0.03)
log Employment Density		0.08***		−0.03	−0.01
		(0.02)		(0.03)	(0.04)
log Dist. to CBD			−0.08***	−0.05***	−0.07***
			(0.02)	(0.02)	(0.02)
log Model Income (residential)					−0.16*
					(0.09)
Observations	88	88	88	88	88
Adjusted R^2	0.23	0.13	0.17	0.28	0.3

Notes. This table compares survey and model predictions of average workplace income. The unit of analysis is a survey area from the DHUTS survey. The survey sample is 11,006 commuters who live and work inside the Dhaka City Corporation, who report positive income, excluding students, homemakers, the unemployed, and government workers. The outcome in panel A is the average income of survey respondents who work in a survey area, using log income truncated at the 99th percentile. In panel (B), it is the residual of log income on gender, age, years of education, occupation and job sector dummies. Model-predicted workplace income in survey area b is $\sum_{j \in b} y_j V_j^W / V_b^W$ where j is a cell phone tower, $y_j = \hat{\psi}_j^R$ is the area adjusted destination fixed effect at j, $V_j^W = \sum_i V_{ij}$ and $V_b^W = \sum_{j \in b} V_j^W$ denote workplace population in tower j and survey area b, respectively. Regressions are weighted by survey area employment population. The Central Business District (CBD) is Shapla Chatter in Motijheel. Appendix Figure B.4 shows the corresponding scatter plots. Robustness in Appendix Tables C.5 and C.6. The analogous exercise at the residential level in Appendix Table C.9. Conley standard errors with 5 km distance cutoff shown in parentheses. $^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table 3: Average Residential Income: Model Prediction and Nighttime Lights

	log VIIRS Nighttime Lights			
	(1)	(2)	(3)	(4)
Panel A. Dhaka, Bangladesh				
log Model Income (residential)	0.57***			0.53***
	(0.04)			(0.07)
log Residential Density		0.26***		0.01
		(0.06)		(0.02)
log Dist. to CBD			−0.52***	−0.11
			(0.10)	(0.09)
log Model Income (workplace)				−0.04
				(0.05)
Sub-district FE (count)				X (55)
Observations	1,868	1,868	1,868	1,868
Adjusted R^2	0.71	0.31	0.34	0.84
Panel B. Colombo, Sri Lanka				
log Model Income (residential)	0.79***			0.49***
	(0.02)			(0.05)
log Residential Density		0.63***		−0.07*
		(0.03)		(0.03)
log Dist. to CBD			−1.08***	−0.30***
			(0.12)	(0.10)
log Model Income (workplace)				0.01
				(0.04)
Sub-district FE (count)				X (42)
Observations	1,199	1,199	1,199	1,197
Adjusted R^2	0.87	0.66	0.76	0.93

Notes. This table compares high resolution nighttime lights with model predicted average residential income. The unit of analysis is a cell phone tower in the greater metropolitan area of each city. The outcome variable is log mean light intensity in the VIIRS nighttime light data inside each Voronoi cell (weighting each VIIRS cell by its overlap with the Voronoi cell). Average model residential (take-home) income at tower i is $\sum_j y_j V_{ij}/V_i^H$ where j indexes workplace towers, $y_j = \hat{\psi}_j^R$ is the area adjusted destination fixed effect at j, and V_i^H is total residential population at i. Regressions are weighted by tower residential population. Column 4 controls for 55 sub-district (thana) fixed effects for Dhaka (panel A), and 42 sub-districts (Divisional Secretariat) fixed effects for Colombo (panel B). Appendix Figure B.5 shows corresponding scatter plots. Robustness to definitions of model income in Appendix Table C.7. Conley standard errors with 5 km distance cutoff shown in parentheses.$^*p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$

Table 4: Impact of Hartal on Predicted Income, Travel Behavior, and Workplace Attendance

	(1)	(2)	(3)	(4)	(5)	(6)
	All Coefficients: % **Change** From Workday Mean					
	All Trips	Work Trips	All Trips	Work Trips	All Trips	Work Trips
Panel A. Predicted Income						
Hartal	-0.048^{***}	-0.044^{***}				
	(0.009)	(0.015)				
Friday (free day)	-0.208^{***}	-0.452^{***}				
	(0.007)	(0.014)				
Observations	$22.5 \cdot 10^6$	$22.5 \cdot 10^6$				
Panel B. Make a Trip						
Hartal	-0.038^{***}	-0.037^{***}	-0.045^{***}	-0.068^{***}	-0.050^{***}	-0.079^{***}
	(0.007)	(0.014)	(0.008)	(0.017)	(0.008)	(0.018)
Friday (free day)	-0.159^{***}	-0.430^{***}	-0.190^{***}	-0.543^{***}	-0.209^{***}	-0.602^{***}
	(0.006)	(0.013)	(0.007)	(0.018)	(0.008)	(0.019)
Hartal x Low Income			0.015^{***}	0.067^{***}	0.013^{***}	0.063^{***}
			(0.003)	(0.010)	(0.003)	(0.010)
Friday x Low Income			0.066^{***}	0.239^{***}	0.058^{***}	0.215^{***}
			(0.004)	(0.013)	(0.004)	(0.012)
Hartal x Short Commute					0.014^{***}	0.031^{***}
					(0.002)	(0.004)
Friday x Short Commute					0.053^{***}	0.165^{***}
					(0.003)	(0.005)
Observations	$22.5 \cdot 10^6$	$22.5 \cdot 10^6$	$22.5 \cdot 10^6$	$22.5 \cdot 10^6$	$22.5 \cdot 10^6$	$22.5 \cdot 10^6$
Workday Mean	0.79	0.37	0.79	0.37	0.79	0.37

Notes. This table shows differences in predicted income and travel probability on hartal days and Fridays relative to workdays. All coefficients show proportional changes relative to workdays. The sample is all days with commuting data (including stationary trips) for commuters with distinct long term home and workplace towers (27% of all users). For commuter c on calendar date t, denote their trip origin by i_{ct}, destination by j_{ct}, and c's long-term workplace by j_c^W. In panel A, the outcome is predicted income. In column (1), commuters earn the destination wage $\exp\left(\hat{\psi}^R_{j_{ct}}/\epsilon\right)$ for any proper trip and zero otherwise. In column (2), commuters earn positive income only when $i_{ct} \neq j_{ct} = j_c^W$. In both cases, the gravity equation is estimated on non-hartal weekdays, and we use $\epsilon = 6.4$. In panel B, the outcome is a dummy for proper trip ($j_{ct} \neq i_{ct}$) in odd columns, and a dummy for proper workplace trip ($j_{ct} = j_c^W \neq i_{ct}$) in even columns. All regressions include commuter and month fixed effects, and dummies for Saturday and holidays. In columns (3)-(6), we fully interact the model with dummies for low-wage commuters (c's long-term workplace wage $\hat{\psi}^R_{j_c^W}$ is below-median) and short-commute commuters (c's travel time between long-term home and work is below-median). Reported coefficients are proportional changes relative to non-hartal, non-holiday weekday mean. Standard errors clustered at the calendar date level in parentheses. $^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$

A Appendix

A.1 Estimation of Constrained Gravity Equation

The structural gravity equation (2) implies a constraint between the destination fixed effects ψ_j, origin fixed effects μ_i, and the distance coefficient β in the empirical gravity equation (3). In our main specification, we do not impose these constraints, but we show that this does not affect the estimation results. Here, we describe the constrained estimation procedure in more detail.

First, we estimate a version of gravity equation (3) without origin fixed effects by OLS:

$$\log(\pi_{ij}) = \psi_j - \beta \log(D_{ij}) + \varepsilon_{ij}, \tag{5}$$

Next, using the estimated destination fixed effects $\hat{\psi}_s^1$ and the distance coefficient $\hat{\beta}^1$ from the OLS regression, we compute the model-predicted origin terms $\tilde{\mu}_i^2 = \log\left(\sum_s \exp\left(\hat{\psi}_s^1 - \hat{\beta}^1 d_{is}\right)\right)$. Next, we estimate (5) using $\log(\pi_{ij}) - \tilde{\mu}_i^2$ as outcome variable. This leads to the step 2 estimates $\hat{\psi}_s^2$ and $\hat{\beta}^2$.

In general, after step k, we construct model terms $\tilde{\mu}_i^{k+1} = \log\left(\sum_s \exp\left(\hat{\psi}_s^k - \hat{\beta}^k d_{is}\right)\right)$ and run (5) using $\log(\pi_{ij}) - \tilde{\mu}_i^{k+1}$ as outcome variable. We iterate this procedure until the vectors $\left(\hat{\psi}_s^k\right)_s$ and $\left(\hat{\psi}_s^{k+1}\right)_s$ converge in L^2 norm, with a tolerance of 10^{-6}.

The procedure is identical to SILS (structurally iterated least squares) proposed in trade gravity literature (Head and Mayer, 2014), except that our model constraints are only on the origin fixed effects, but not on the destination fixed effects. Fally (2015) discusses the potential bias of the estimated gravity fixed effects without imposing these model constraints.

A.2 Model Extension: Worker Heterogeneity in Effective Labor Supply

In Section 3, we assumed that workers are ex-ante identical. However, in panel B in Table 2, we measure the model's predictive power *after* netting out individual demographic characteristics from survey income. Here, we show how this validation regression arises directly in a specific model with worker heterogeneity.

Assume that worker ω supplies ξ_ω effective units of labor. $\omega's$ income from working in j is $\xi_\omega W_j$ instead of simply W_j. Otherwise, workers have the same disutility of commuting, and face the same profile of wages. This implies that workers living at the same location i face the same workplace location choice, regardless of ξ_ω. Hence, in aggregate, the gravity equation (2) continues to hold unchanged.

However, the average ξ_ω of commuters working in j affects average income at that location. Hence, the correct validation regression should control for average ξ_ω at location j from individual income. To the extent that ξ_ω depends on observable characteristics (gender, age, education level, occupation, job sector), this is exactly what the specification in panel B, Table 2 achieves.

A.3 Structural Estimation: How Much do Individual Shocks and Travel Time Affect Income

In the main analysis, we assume that an agent earns income directly proportional to her wage. Formally, the Fréchet shocks $Z_{ij\omega}$ and travel time D_{ij} affect utility but not income. Here, we relax this assumption and allow $Z_{ij\omega}$ and D_{ij} to partly affect income; for example, they may affect productivity or labor supply. We derive a transparent method that allows survey income data to speak as to the role of shocks and travel time for income.

Model. Assume that income is given by $Y_{ij\omega}^{\alpha_z,\alpha_d} = W_j Z_{ij\omega}^{\alpha_z} D_{ij}^{-\tau\alpha_d}$, where $\alpha_z, \alpha_d \in [0,1]$ respectively control the extent to which the shocks $Z_{ij\omega}$ and travel time D_{ij} affect income. For example, when $\alpha_z = 1$ and $\alpha_d = 0$, shocks affect utility and income equally, while travel time only affects utility. We derive formulas for expected income in the following four extreme extreme cases:

$$\begin{aligned}
&\mathrm{E}y_{ij\omega}^{0,0} = w_j \\
&\mathrm{E}y_{ij\omega}^{0,1} = w_j - \tau d_{ij} \\
&\mathrm{E}y_{ij\omega}^{1,1} = \frac{1}{\epsilon}\log\left(\sum_s \exp\left(\epsilon w_j - \epsilon\tau d_{ij}\right)\right) - \frac{K}{\epsilon} \text{ for some absolute constant } K \\
&\mathrm{E}y_{ij\omega}^{1,0} = \mathrm{E}y_{ij\omega}^{1,1} + \tau d_{ij}
\end{aligned} \tag{6}$$

When neither shocks nor travel time affect income, income is simply the destination wage. In the second case, travel time fully affects labor earnings. When the shocks $Z_{ij\omega}$ affect income, as in the third and fourth cases, log income for a worker commuting between i and j depends on the distribution of the shock *conditional* on destination j being chosen. By virtue of the Fréchet distribution, the conditional distribution $y_{ij\omega}|j \in \arg\max_s U_{is\omega}$ is also Fréchet with the same shape parameter ϵ and scale $T_i = \sum_s T_{is} = \sum_s \left(W_s D_{is}^{-\tau}\right)^{\epsilon}$. In particular, this distribution only depends on the origin i and thus expected log income is the same for all destinations j.

In the general case, log income is a convex combination of the following four extreme cases:

$$y_{ij\omega}^{\alpha_z,\alpha_d} = \alpha_z\alpha_d \cdot y_{ij\omega}^{1,1} + \alpha_z\left(1-\alpha_d\right) y_{ij\omega}^{1,0} + \left(1-\alpha_z\right)\alpha_d \cdot y_{ij\omega}^{0,1} + \left(1-\alpha_z\right)\left(1-\alpha_d\right) y_{ij\omega}^{0,0}. \tag{7}$$

Using (6) and dropping the constant K, this simplifies to

$$\mathrm{E}y_{ij\omega}^{\alpha_z,\alpha_d} = \frac{\alpha_z}{\epsilon}\left[\log\left(\sum_s \exp\left(\epsilon w_j - \epsilon\tau d_{ij}\right)\right) + \epsilon\tau d_{ij}\right] + \frac{1-\alpha_z}{\epsilon}\left[\epsilon w_j\right] + \frac{\alpha_d}{\epsilon}\left[-\epsilon\tau d_{ij}\right] \tag{8}$$

The intuition of this expression is as follows. For the third term, if travel time affects income, we expect that people who commute further away have lower income. The difference between the first two terms is more subtle. If Fréchet shocks affect income, then the first term is the best explanatory variable for income.[17] If shocks do not affect income, the wage at the destination should be the best predictor of income.

Estimating Parameters $\alpha_z, \alpha_d, \epsilon$**.** We are now in a position to estimate the parameters α_z, α_d and ϵ. Specifically, we estimate by OLS the equation:

$$y_{ij\omega}^S = \rho_1 \hat{X}_{ij}^1 + \rho_2 \hat{X}_j^2 + \rho_3 \hat{X}_{ij}^3 + \varepsilon_{ij\omega}^S, \tag{9}$$

where $y_{ij\omega}^S$ is survey-based income of commuter ω who lives at i and works at j, and $\hat{X}_{ij}^1 = \log\left(\sum_s \exp\left(\hat{\psi}_s - \hat{\beta} d_{ij}\right)\right) + \hat{\beta} d_{ij}$, $\hat{X}_j^2 = \hat{\psi}_j$ and $\hat{X}_{ij}^3 = -\hat{\beta} d_{ij}$ are estimators of the three terms in square brackets in (8), computed using the gravity equation estimates. (Recall that $\hat{\psi}_j$ is a consistent estimator for ϵw_j, and $\hat{\beta}$ is a consistent estimator for $\epsilon\tau$.) Asymptotically, we have

$$\hat{\alpha}_z = \frac{\hat{\rho}_1}{\hat{\rho}_1 + \hat{\rho}_2},\ \hat{\alpha}_d = \frac{\hat{\rho}_3}{\hat{\rho}_1 + \hat{\rho}_2},\ \text{and } \hat{\epsilon} = \frac{1}{\hat{\rho}_1 + \hat{\rho}_2}. \tag{10}$$

Table C.11 reports the estimates of α_z, α_d, and ϵ based on estimating equation (9) with OLS, and using transformation (10). We report two types of standard errors: based on the Delta method (in round parentheses) and based on bootstrapping at the origin survey area level (in square parentheses).[18] In columns 1-2, we estimate the full equation (9), and we find that $\hat{\alpha}_d$ is close to zero with a small and insignificant negative value, and the other

[17] The first term is analogous to the market access term in gravity trade literature, except that it includes the compensation income from commuting cost in utility.

[18] For the Delta method, following the equations in (10), define $(\epsilon, \alpha_z, \alpha_d) = h(\rho_1, \rho_2, \rho_3) = \left(\frac{1}{\rho_1+\rho_2}, \frac{\rho_1}{\rho_1+\rho_2}, \frac{\rho_3}{\rho_1+\rho_2}\right)$. The Delta method states that the asymptotic covariance matrix of $(\hat{\epsilon}, \hat{\alpha}_z, \hat{\alpha}_d)$ is given by $J(h)^T \Sigma J(h)$ where $J(h)$ is the Jacobian of h and Σ is the asymptotic covariance matrix of $(\hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3)$. The Jacobian is obtained by differentiating h with respect to the ρ's:

$$J(h) = (\rho_1 + \rho_2)^{-2} \begin{bmatrix} -1 & -1 & 0 \\ \rho_2 & -\rho_1 & 0 \\ -\rho_3 & -\rho_3 & \rho_1 + \rho_2 \end{bmatrix}$$

When the distance coefficient α_d is constrained to zero ρ_3 is also zero, we have $(\epsilon, \alpha_z) = h(\rho_1, \rho_2) = \left(\frac{1}{\rho_1+\rho_2}, \frac{\rho_1}{\rho_1+\rho_2}\right)$ and the Jacobian is

$$J(h) = (\rho_1 + \rho_2)^{-2} \begin{bmatrix} -1 & -1 \\ \rho_2 & -\rho_1 \end{bmatrix}$$

parameters are imprecisely estimated when using bootstrapped standard errors. Given that the model restricts $\rho_3 \geq 0$ (from $\alpha_d \in [0, 1]$), in columns 3-4 we restrict the coefficient on travel time to be equal to zero ($\rho_3 = 0$) and estimate the other two parameters. This does not affect the point estimates for $\hat{\alpha}_z$ and $\hat{\epsilon}$ while improving precision.

These results show that idiosyncratic shocks partly affect income, while travel time is most consistent with a pure utility cost. Robustness exercises using "home" and "work" commuting flows and iterated gravity equation give qualitatively similar results (not reported).

A.4 Discussion: Estimating Wages without Commuting Flows

In the absence of detailed commuting flows data, Ahlfeldt et al. (2015) use an exactly identified procedure to infer wages from total residential and total workplace population counts at each location, as well as knowledge of the distance coefficient β (which in turn is estimated from a gravity equation at a coarse level). While the two procedures are equivalent if equation (3) does not have noise, estimating wages using bilateral commuting flows is more robust against measurement error and idiosyncratic departures from the model.[19] If random shocks affect bilateral commuting flows proportionally, as in equation (3), then in the procedure using total counts, commuting pairs with high commuting flow (e.g. nearby pairs) have a large, noisy influence on estimated wages. Indeed, we find that the gravity equation explains only between 0.6 and 0.75 of the variation in bilateral log commuting flows. Severen (2019) documents that time-invariant link-specific effects explain a large share of the variance in commuting flows in Los Angeles. Related, bilateral commuting data allows us to test the stability of our results, i.e., the estimated wages should be similar when using a sub-sample of location pairs for estimation. Indeed, we show that results are similar when estimating (3) with or without nearby pairs of towers (for which commuting flows are more likely to be mismeasured). Finally, with the gravity equation we can in principle estimate the effect of distance non-parametrically and simultaneously with destination fixed effects. Appendix Figure B.3 shows that a linear fit as a function of log travel time is, in fact, appropriate.

A.5 Model: Approximate Invariance to Aggregation Level

The model has a general (approximate) invariance property with respect to the level of geographic aggregation, both at the origin and at the destination level.

[19] The constrained gravity equation (2) yields asymptotically identical results as the procedure in Ahlfeldt et al. (2015) given the knowledge of $\beta = \epsilon\tau$ if we estimate the model using the employment distribution as moments ($\sum_i H_i \pi_{ij}$), where H_i is the number of residents in location i, instead of the bilateral commuting probability $\log(\pi_{ij})$ as moments, as in our procedure.

At the origin level, the model is approximately invariant with respect to the origin aggregation level, because the basic discrete choice problem is individual specific.

At the destination level, the aggregation level affects the interpretation of wages W_j in a straight-forward way. Assume that location j is in fact composed of several sub-locations $k_1, k_2, ..., k_{N_j}$, and we estimate the model at the higher level (j) and ignore the sub-locations. The wage we obtain, $W_j = \left(\sum_{\ell=1}^{N_j} W_{k_\ell}^{\epsilon}\right)^{1/\epsilon}$, represents a C.E.S. aggregate with elasticity of the true underlying wages at all sub-locations within j. (This is easy to prove using the standard properties of the Fréchet distribution.) In particular, this implies a simple adjustment for the destination fixed effect $\psi_j = \epsilon w_j$ estimated using the gravity model. Assume that the "real" underlying wage is constant and denoted by W_j^R within each location j, then the C.E.S. relationship becomes $W_j = N_j^{1/\epsilon} W_j^R$, or in logs the underlying wage is given by $w_j^R = w_j^{1/\epsilon} - \log(N_j)$. In terms of estimated quantities, this becomes $\hat{\psi}_j^R = \hat{\psi}_j - \log(N_j)$. The underlying destination fixed effect $\hat{\psi}_j^R$ is obtained from the fixed effect $\hat{\psi}_j$, estimated ignoring sub-locations, minus an adjustment factor equal to the log of the number of true underlying locations where shocks are realized, N_j. This relationship is exact if the distances between each sub-location in location j and all other locations do not depend on the sub-location. Redding and Weinstein (2019) derive an exact relationship by using all the distance profiles in the context of gravity equations of trade models.

A.6 Data: Smoothing procedure for Google Map Travel Time

Due to the large number of bilateral tower pairs (on the order of $\sim 10^6$), we obtain travel time estimates from the Google Maps API for 90,000 randomly selected pairs in each country, and interpolate the travel time for remaining tower pairs. The data was collected in June 2016 for Colombo and in July 2017 for Dhaka. Specifically, for each pair we query the Google Maps Distance Matrix API for the typical driving time on a weekday with departure time at 8 am.[20]

To interpolate the travel time for tower pairs without actual data, we use the following procedure. For any such tower pair i, j, we compute the kernel smoothed speed (travel time divided by straight line distance) of trips starting at origins a near i and ending at destinations b near j, for pairs a, b for which we have Google Maps data. Formally, the smoothed speed $\hat{s}_{ij}$ between i and j is given by

$$\hat{s}_{ij} = \frac{\sum_{a \neq i, b \neq j} \frac{1}{h_a h_b} K\left(\frac{d_{i,a}}{h_a}\right) K\left(\frac{d_{j,b}}{h_b}\right) s_{a,b} GM_{a,b}}{\sum_{a \neq i, b \neq j} \frac{1}{h_a h_b} K\left(\frac{d_{i,a}}{h_a}\right) K\left(\frac{d_{j,b}}{h_b}\right) GM_{a,b}},$$

[20]Specifically, we queried Google Maps for travel times on Friday, August 26, 2016, in Colombo, and Wednesday, September 13, 2017, in Dhaka. The queries were sent about one month before these dates.

where $GM_{a,b}$ is a dummy indicator for having Google Maps data between a and b, and $d_{x,y}$ is the distance in decimal degrees between towers x and y. We use the two-dimensional Epanechnikov kernel $K(d) = \frac{3}{\pi}|1 - d^2|_+$, together with an "adaptive" kernel bandwidth $h_x \equiv h\frac{\sqrt{A_x}}{\bar{A}}$ where A_x is the area (in square kilometers) of Voronoi cell of cell phone tower x, $\bar{A}$ is the average of $\sqrt{A_x}$ over all towers x, and $h = 0.03$ and $h = 0.1$ are the (manually-chosen) bandwidth for average-radius towers in Dhaka and Colombo, respectively. A tower with average square root area will have a bandwidth of about 3.5 kilometers in Dhaka, and 11 kilometers in Colombo.

We evaluate the predictive power of this procedure using a leave-one-out test using only the tower pairs with actual Google Maps data. The R^2 of $\hat{s}_{a,b,}$ on $s_{a,b}$ is 0.98 and 0.96 and Colombo and Dhaka, respectively. Note, the prediction $\hat{s}_{a,b}$ is computed without using $s_{a,b}$.

A.7 Availability of Conventional Data Sources

Fine-grained spatially disaggregated data on wages at the firm location is rare and difficult to access in developing countries. For example, the Bangladesh economic census does not include labor costs data, and we were not able to acces Sri Lanka economic census microdata.

As a case study, here we document the availability of firm census data in Sub-Saharan Africa, a region undergoing rapid urban growth and urban transformation. We collected data on the 27 largest countries that account for over 95% of the population in the region. Of these, 16 ever had an economic census, 11 covered informal firms. However, at most 4 included wage data, which accounts for between 5.6 and 8.6% of the urban population of all countries in the sample. (The 2014 Ghana and 2015 Zimbabwe censuses included wage data, while for the ongoing censuses in Mali and Togo we do not know if wage data was collected.)

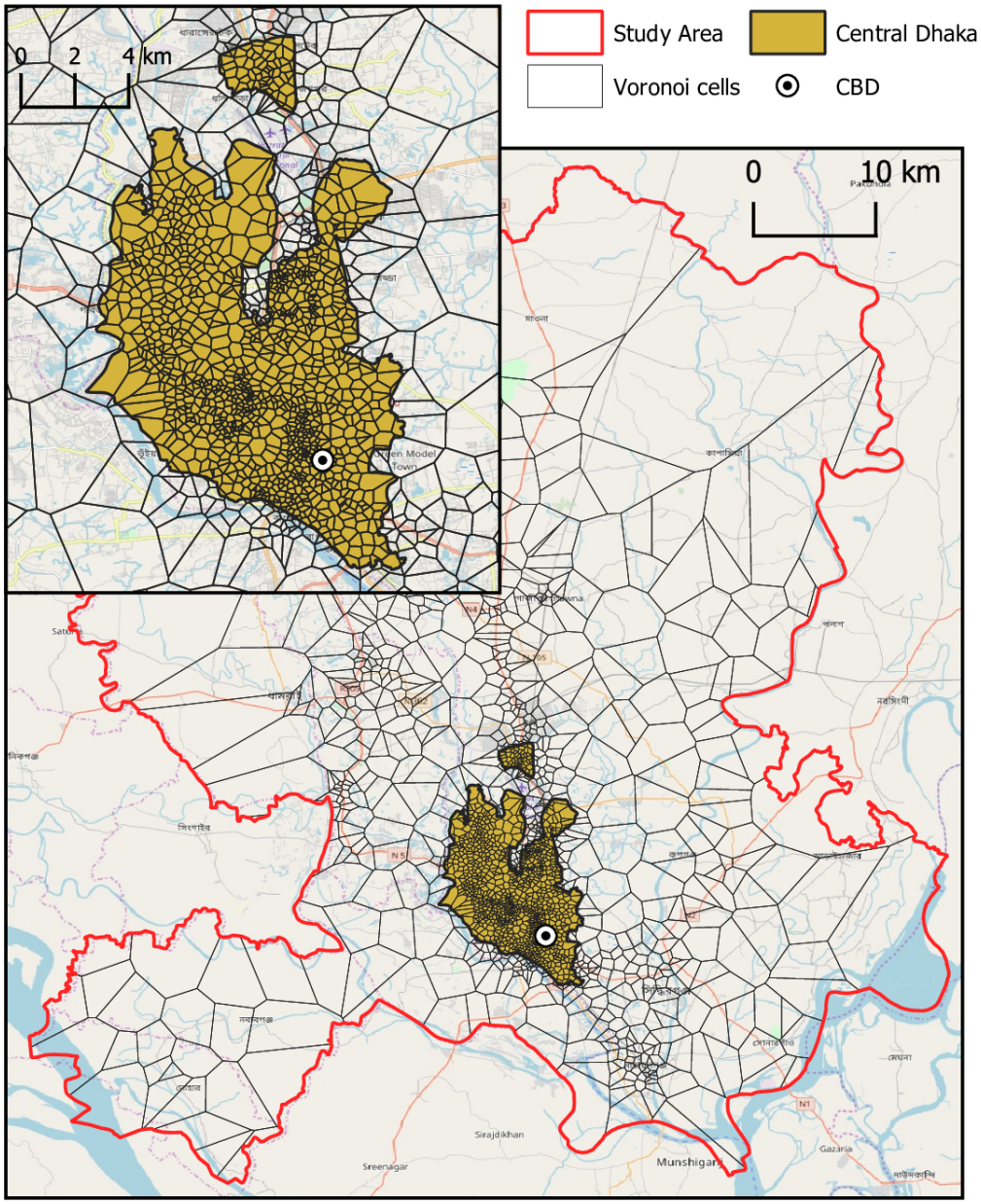
Table A.1: Sub-Saharan African Countries with Economic Cenuses

Country	Urban Population	Year of Last Census	Covers Informal	Wage Data
Nigeria	98,525,244	-		
South Africa	38,112,552	-		
D.R.C.	37,382,220	-		
Ethiopia	22,367,255	2004	Yes	No
Angola	20,157,104	2002	No	
Tanzania	19,972,890	2014	Yes	No
Ghana	16,529,104	2014	Yes	Yes
Cameroon	13,918,524	2016	No	
Kenya	13,756,737	2017	No	
Ivory Coast	12,652,168	-		
Uganda	10,536,394	2010	Yes	No
Mozambique	10,990,322	2014	Yes	No
Madagascar	9,769,765	-		
Mali	8,101,667	2019	?	?
Senegal	7,690,895	2016	Yes	No
Zambia	7,659,992	2011	?	No
Burkina Faso	5,806,985	2009	Yes	No
Zimbabwe	5,446,070	2015	Yes	Yes
Benin	5,432,724	2008	Yes	
Guinea	4,711,991	-		
Niger	3,659,066	-		
Chad	3,546,586	-		
Togo	3,332,216	2018	?	?
Malawi	3,238,839	-		
South Sudan	2,532,134	-		
Rwanda	2,150,199	2017	Yes	No
Burundi	1,458,139	-		

Notes: For each country, we checked the national statistics agency website as well as the Google Search results for the terms "economic census," "firm census," "establishment census," "enterprise census," and "business registry," in English, French or Portuguese. We could not find official census reports for Ethiopia and Zambia, while the Mali and Togo censuses are still ongoing. Detailed results available upon request. Data on urban population from
`https://en.wikipedia.org/wiki/Urbanization_by_country` and
`https://en.wikipedia.org/wiki/List_of_sovereign_states_and_dependent_territories_in_Africa`.

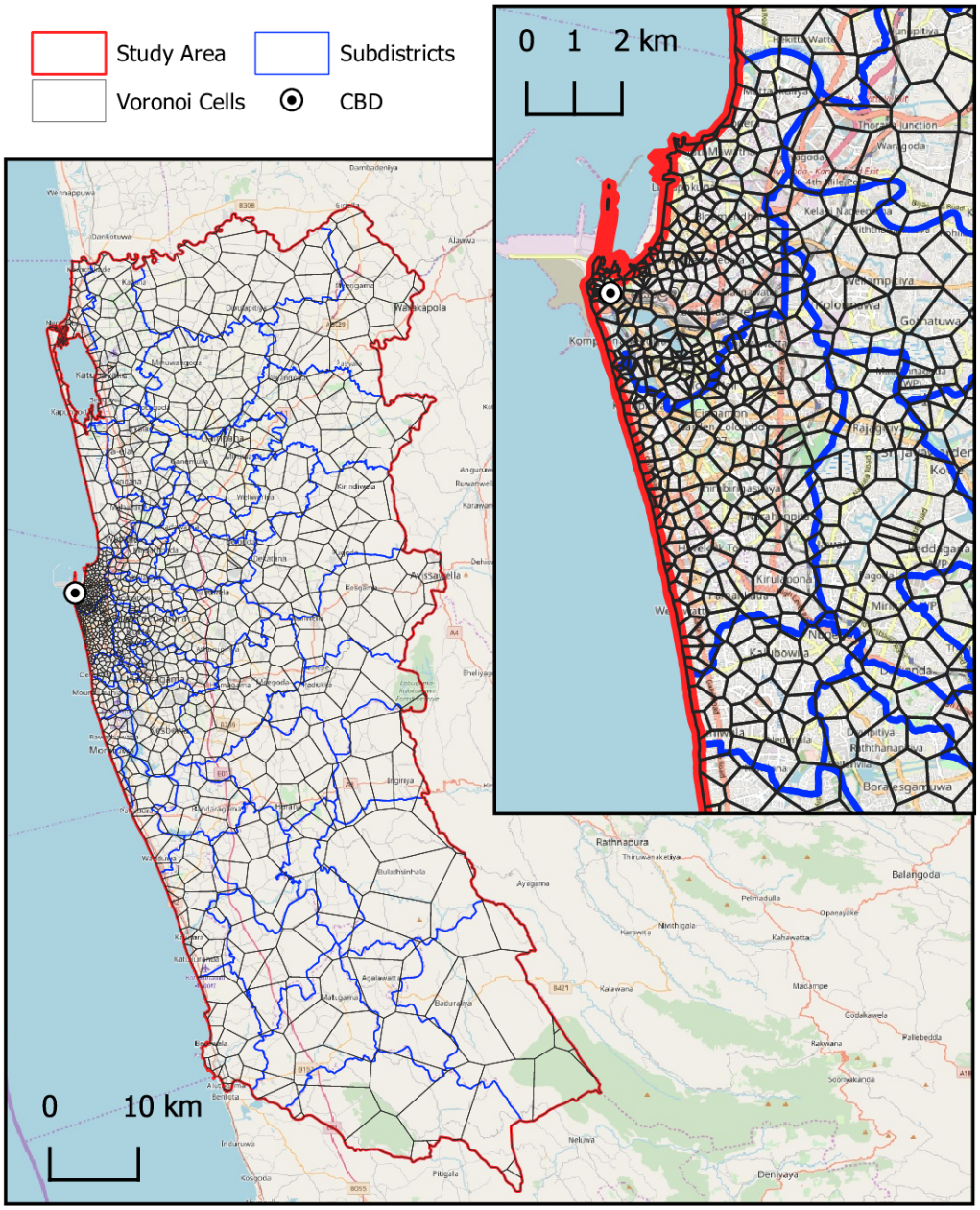
B Additional Figures

Figure B.1: Administrative Units and Cell Phone Voroni Cells in Dhaka



Notes. This figure shows the map of cell phone tower Voronoi cells in Dhaka, Bangladesh. The yellow shaded area is the Dhaka City Corporation (DCC), the urban core of Dhaka, the main sample in the DHUTS transportation survey. The overall study area covers three districts in Bangladesh: Dhaka, Gazipur, and Narayanganj. The Voronoi cell of a tower is the locus of all points closer to that tower than to any other tower.

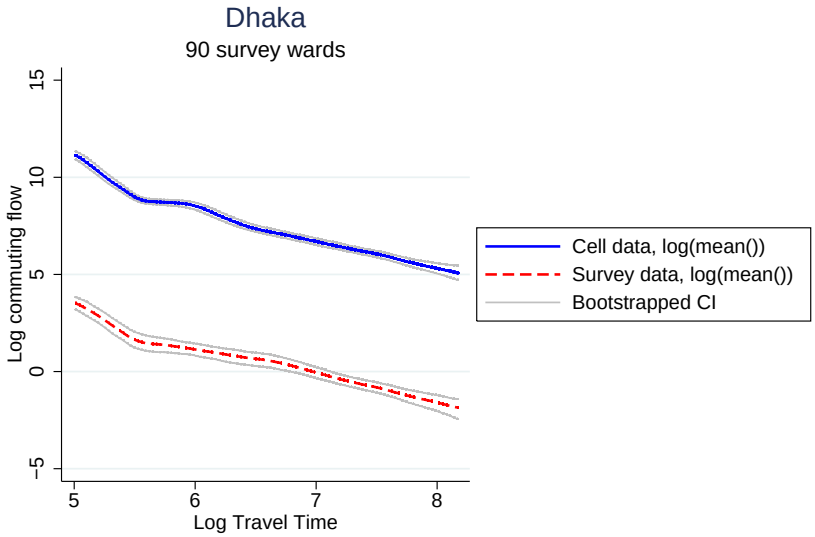
Figure B.2: Administrative Units and Cell Phone Voroni Cells in Colombo



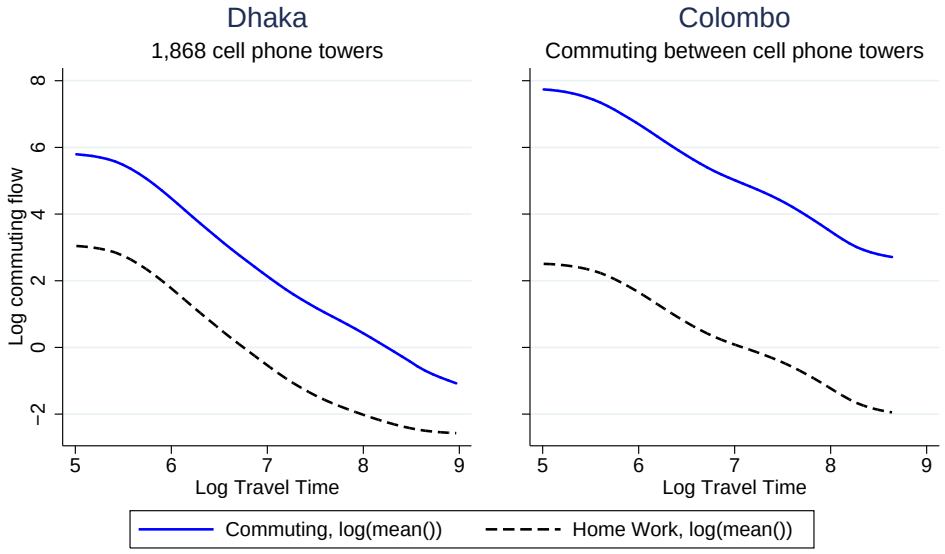
Notes. This figure shows the map of cell phone tower Voronoi cells in Colombo, Sri Lanka. The study area covers the entire Western Province in Sri Lanka. The Voronoi cell of a tower is the locus of all points closer to that tower than to any other tower.

Figure B.3: Commuting Flows from Survey Data and Cell Phone Data

Panel (A) Survey vs Cell Phone Data



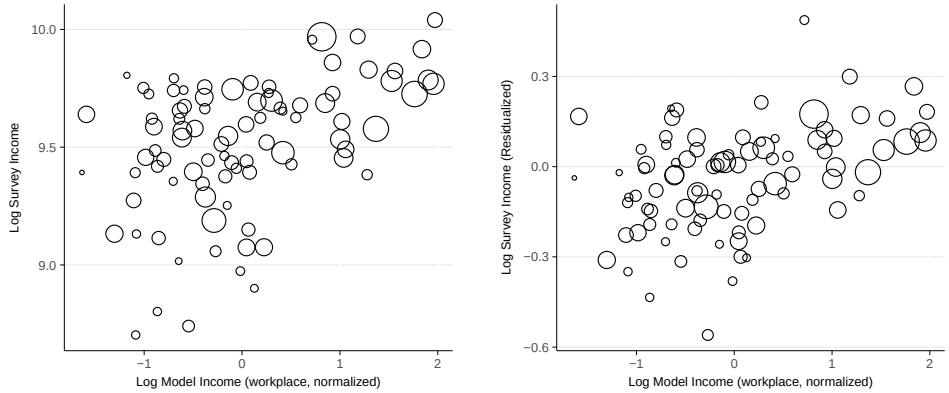
Panel (B) Commuting Flows vs Home-Work Flows



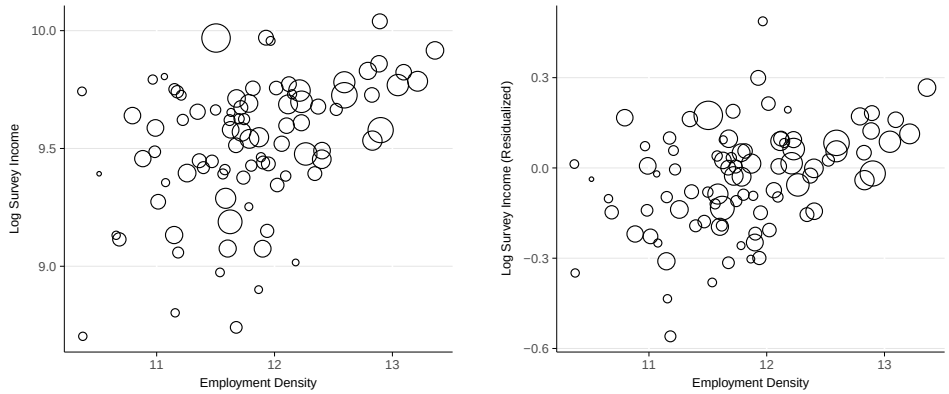
Notes. This figure compares the decay of commuting flows with travel time in survey and cell phone data. The unit of analysis is 7,836 survey area pairs in Panel A, and $1.6 \cdot 10^6$ and $1.4 \cdot 10^6$ tower pairs in Dhaka and Colombo in Panel B, respectively. Panel A compares commuting flows from the DHUTS survey (red, dash) and from cell phone data (blue, solid) in Dhaka. Panel B compares daily commuting trips (blue, solid) and home-work commuting trips (black, dash). See Section 2.1 for the definition of home-work commuting trips. In each graph, commuting flows are first averaged within each of 100 equal bins of log travel time below the 99th percentile, and the plot shows the local linear regression of log mean commuting flow on log travel time. This procedure avoids the bias due to zero commuting flows, which is important for survey and home-work commuting data. The DHUTS sample (described in Table C.2) has 12,510 commuters. The cell phone data sample has $18 \cdot 10^6$ trips in Panel A, and $38 \cdot 10^6$ daily trip and $5.2 \cdot 10^6$ for home-work trips in Dhaka, and $237 \cdot 10^6$ daily trips and $2.6 \cdot 10^6$ home-work trips in Colombo, in Panel B. In Panel A, pointwise bootstrapped 95% confidence intervals clustered at the origin survey area shown in gray.

Figure B.4: Average Workplace Income: Model Predictions and Survey Data

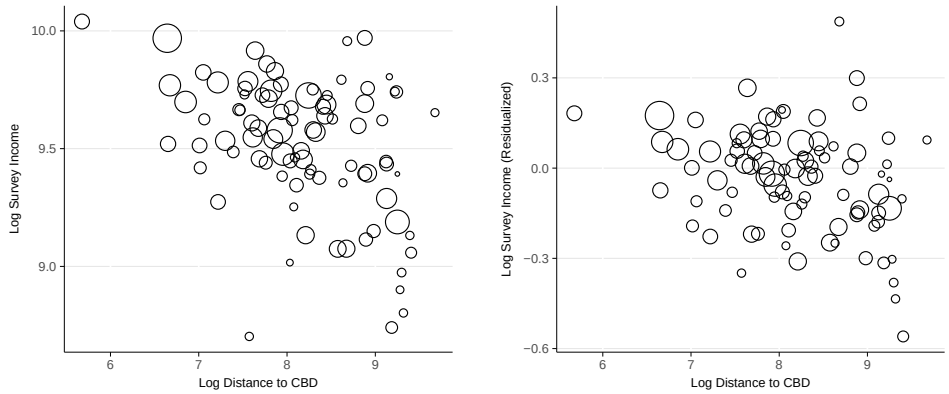
(A) Model and Survey Workplace Income



(B) Employment Density and Survey Workplace Income



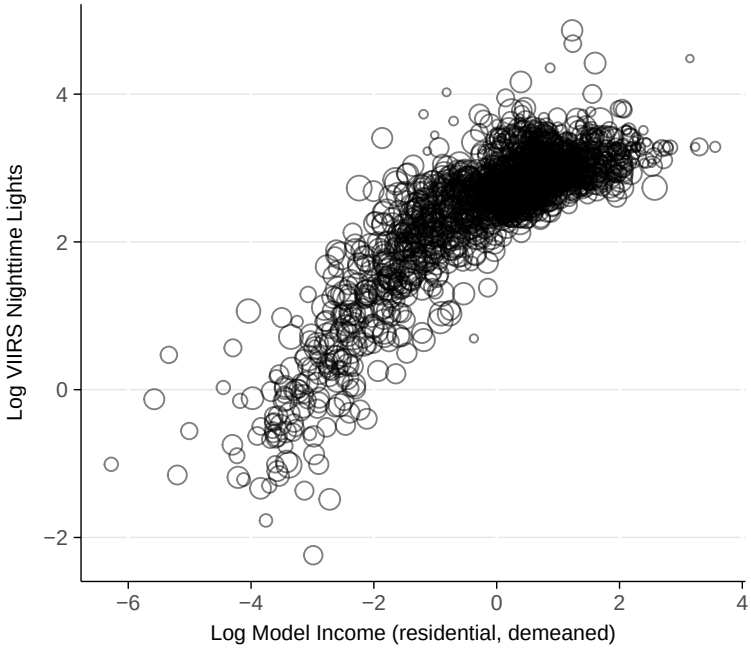
(C) Distance to City Center and Survey Workplace Income



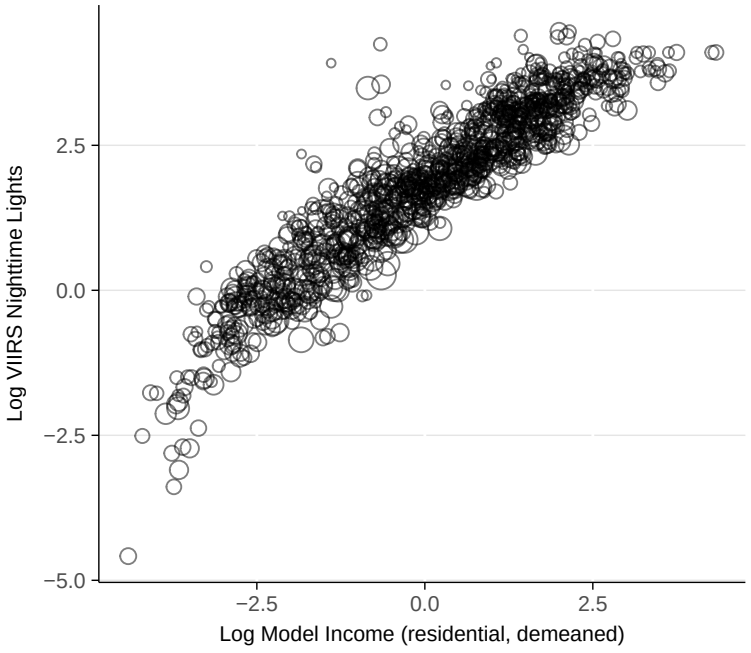
Notes. The scatter plots in the three columns correspond to columns 1, 2, and 3 in Table 2. The left-side graphs correspond to panel A (log survey income), and the right-side graphs to panel B (the residual of log survey income after regressing on demographics). Each circle represents a workplace survey area, with radius proportional to employment population obtained from the cell phone data.

Figure B.5: Average Residential Income: Model Prediction and Nighttime Lights

(A) Dhaka

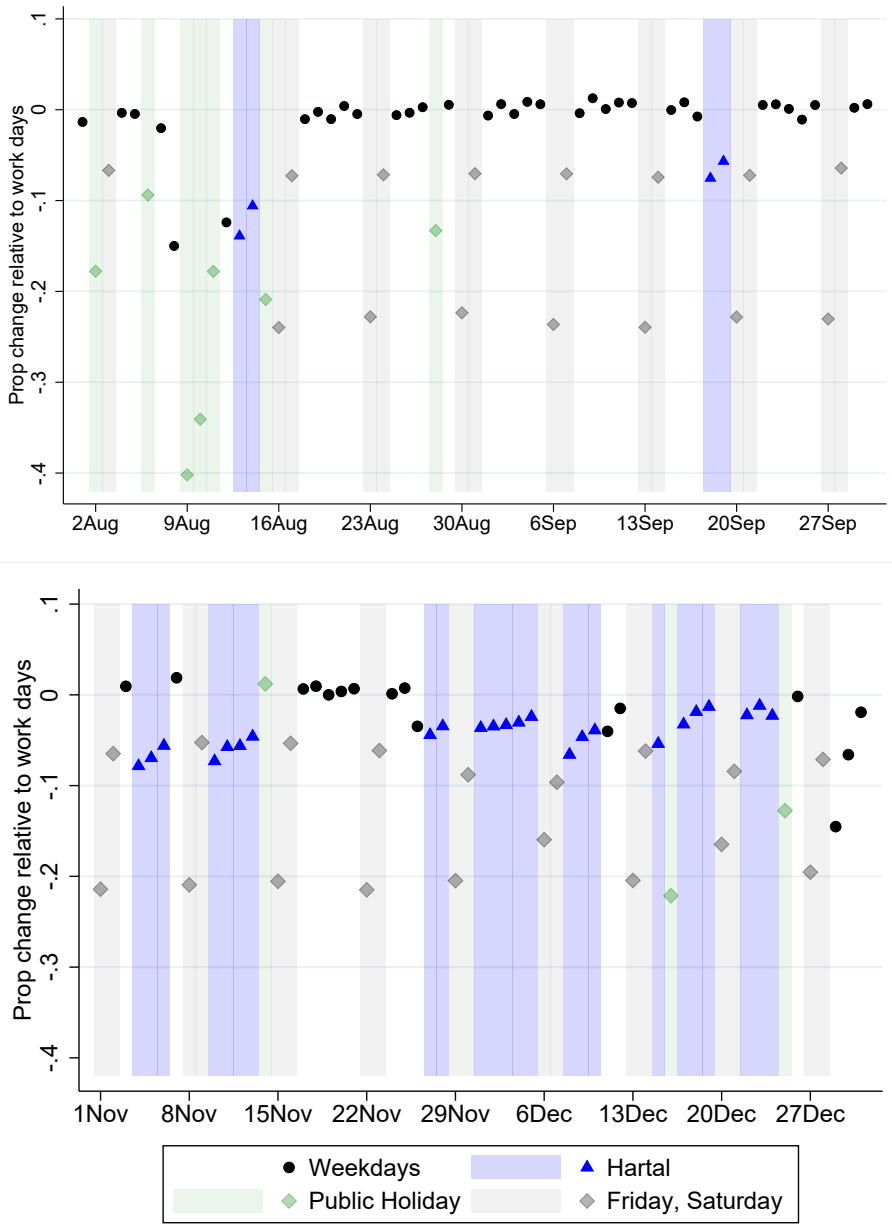


(B) Colombo



Notes. The two scatter plots correspond to the first column, panels A and B, in Table 3. Each circle represents residential cell phone tower, with radius proportional to residential population obtained from the cell phone data.

Figure B.6: Model-Predicted Income by Calendar Date (Hartals, Holidays and Weekends)



Notes. This figure shows average predicted income by calendar date. The Y axis plots the proportional change relative to the mean on workdays. The sample and outcome are as in Panel A, Column 1 in Table 4. The figure plots calendar date fixed effects from a regression of normalized trip predicted income (equal to zero when the person does not travel) on commuter and calendar date fixed effects. Hartal dates are from Ahsan and Iqbal (2015) and public holidays from `https://www.timeanddate.com/holidays/bangladesh/`. Friday is the main free day in Bangladesh, and Saturday is the other weekend day. August 2 is Jumatul Bidah, August 6 is Shab-e-qadr, August 9-12 is the Eid ul-Fitr (end of Ramadan), August 15 is the National Mourning Day, August 28 is Janmashtami, November 14 is Ashura, December 16 is Victory Day, and December 25th is Christmas Day. The last week in December preceded the General Election of January 5, 2014.

C Additional Tables

Table C.1: Cell Phone Data Coverage at User-Day Level

		Dhaka, Bangladesh	**Colombo, Sri Lanka**
(1)	Users in sample	$5.3 \cdot 10^6$	$3.0 \cdot 10^6$
(2)	Days in sample	122	395
(3)	All user-days possible = (1)×(2)	$6.5 \cdot 10^8$	$1.2 \cdot 10^9$
(4)	User-days with data	$2.9 \cdot 10^8$	
(5)	User-days with data (5-10am)	$1.5 \cdot 10^8$	
(6)	User-days with data (10am-3pm)	$2.4 \cdot 10^8$	
(7)	User-days with data (5-10am and 10am-3pm)	$1.0 \cdot 10^8$	$3.4 \cdot 10^8$
(8)	Coverage rate =(7)/(3)	16.1%	28.8%

Notes: This table describes data coverage in the two countries. The first row indicates the number of unique users (who appear at least once in the data set). The second row shows the total number of calendar dates with data. The third row is the product of the previous two, which is the theoretical upper bound of user-day combinations that could appear in the data. (Note that in practice some users only start using a cell phone partway through the period, so this is an overestimate.) Rows 4-6 describe the actual number of user-days in the Bangladesh data under different restrictions. The seventh row shows the number of user-days for which we have at least one location between 5 am and 10 am, and at least one location between 10 am and 3 pm – this corresponds to the data necessary to define commuting behavior for that user and that day.

Table C.2: Comparison of Commuting Flows from Survey Data and Cell Phone Data

	Flow survey data (DHUTS)			
	(1)	(2)	(3)	(4)
Log flow cell phone data	0.69***	0.80***	0.32***	0.75***
	(0.024)	(0.028)	(0.068)	(0.059)
Log duration			-1.13***	-0.14
			(0.18)	(0.14)
Origin and destination fixed effects		Yes		Yes
Observations	7915	7915	7915	7915

Notes: This table shows the relationship between commuting flows from two different data sets in Dhaka: the DHUTS transportation survey (outcome) and from cell phone data (explanatory variable). An observation is a pair of survey areas from the DHUTS survey. The coefficients show the estimates from the Poisson pseudo-maximum-likelihood (PPML) estimation of DHUTS commuting flow on log flows from cell phone. We use PPML to deal with the presence of zeros in DHUTS commuting flows (Silva and Tenreyro, 2006). If cell phone commuting flow data is a perfect measure of commuting flows, one would expect coefficients equal to one; the results indicate that cell phone commuting flows contain some measurement error. Standard errors are clustered at the origin survey area level. $^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table C.3: Comparison of Residential Population from Cell Phone Data and Population Census

	log Residential Density (cell phone)		log Residential Population (cell phone)	
	(1)	(2)	(3)	(4)
log Residential Density (census)	1.16***	1.16***		
	(0.03)	(0.14)		
log Residential Population (census)			0.57***	0.40***
			(0.07)	(0.04)
City	Dhaka	Colombo	Dhaka	Colombo
Observations	1,866	1,201	1,866	1,201
Adjusted R^2	0.61	0.49	0.25	0.24

Notes: This table shows the relationship between residential population and population density from cell phone data and from the population census. The unit of analysis is a Voronoi cell around each cell phone tower in the greater metropolitan area of each city (Dhaka, Gazipur, and Narayanganj districts in Bangladesh, and Western Province in Sri Lanka). In cell phone data, residential population is defined as out-commuting flow, namely the total number of commuting trips from a given origin excluding stationary trips (including them yields virtually identical results). Census residential population in a Voronoi cell is computed as the average census population in census geographic units (Mauza for Bangladesh, Grama Niladhari for Sri Lanka), weighted by their overlap with the Voronoi cell. The high adjusted R-squared in columns (1) and (2) indicates a strong association between the geographic density from the two data sources. The slope above one indicates that the cell phone data slightly over-represents residential population in denser areas. The comparatively lower adjusted R-squared in columns (3) and (4) is likely because cell phone operators tend to assign cell phone towers to equalize the subscriber coverage per tower. Conley standard errors with 5 km distance cutoff shown in parentheses. $^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table C.4: Gravity Equation Robustness: Destination Fixed Effects

	Destination Fixed Effects (Benchmark)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dest FE (Gravity w/ Model Origin FE)	0.94***				0.97***			
	(0.002)				(0.004)			
Dest FE (Home-Work Flows)		1.92***				2.48***		
		(0.02)				(0.04)		
Dest FE (Volume + 1)			1.41***				1.08***	
			(0.01)				(0.002)	
Dest FE (Full Sample)				0.98***				1.00***
				(0.001)				(0.0004)
City	Dhaka	Dhaka	Dhaka	Dhaka	Colombo	Colombo	Colombo	Colombo
Observations	1,868	1,860	1,868	1,868	1,201	1,201	1,201	1,201
Adjusted R^2	0.99	0.81	0.94	1.00	0.98	0.75	1.00	1.00

Notes. This table compares destination fixed effects computed under different assumptions. The outcome in the first (last) four columns is the destination fixed effects from the first (third) column in Table 1. The first row uses fixed effects (FE) from the gravity equation estimated using the iterative procedure (columns 2 and 4 in Table 1). The second row uses FE from the gravity equation estimated using home-work flows instead of daily commuting trips (see Section 2.1 for the definition). The third row uses FE from the gravity equation estimated using $\log(V_{ij}+1)$ instead of $\log(V_{ij})$ as outcome variable. The last row uses FE from the gravity equation estimated on all tower pairs below the 99th percentile of the travel time, with travel time censored from below at 180 seconds. Most coefficients are close to 1 and the R^2 is above 0.9, except for home-work flows. This is due to slightly lower coverage for home-work flows with positive values. Some commuting pairs have zero home-work flows, and are hence omitted from the gravity equation. This sample selection tends to overestimate destination fixed effects for locations with low wages. This leads to a flatter profile of destination fixed effects. Standard errors in parentheses. $^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table C.5: Robustness: Average Workplace Income and Survey Income Comparison

(A) Log Survey Income

	log Survey Income (workplace)									
	(1) Gravity w/ Model Origin FE		(2) Home-Work Flows		(3) Excluding Neighboring Towers		(4) Without Area Adjustment		(5) Include All Origins	
log Model Income (workplace)	0.14***	0.26***	0.15***	0.44***	0.14***	0.27***	0.31***	0.21***	0.13***	0.25***
	(0.03)	(0.06)	(0.04)	(0.11)	(0.03)	(0.06)	(0.05)	(0.05)	(0.03)	(0.06)
Geographic Controls		X		X		X		X		X
Adjusted R2	0.26	0.46	0.19	0.44	0.26	0.46	0.3	0.45	0.23	0.49
Observations	88	88	88	88	88	88	88	88	89	89

(B) Log Survey Income Residual on Demographics

	log Survey Income (workplace, residual)									
	(1) Gravity w/ Model Origin FE		(2) Home-Work Flows		(3) Excluding Neighboring Towers		(4) Without Area Adjustment		(5) Include All Origins	
log Model Income (workplace)	0.07***	0.15***	0.08***	0.25***	0.07***	0.15***	0.16***	0.12***	0.07***	0.13***
	(0.02)	(0.03)	(0.02)	(0.07)	(0.02)	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)
Geographic Controls		X		X		X		X		X
Adjusted R2	0.23	0.3	0.17	0.28	0.23	0.3	0.25	0.29	0.22	0.31
Observations	88	88	88	88	88	88	88	88	89	89

Notes. Robustness for columns 1 and 5 from Table 2 comparing model predicted workplace income and survey income. Odd and even columns correspond to the specifications in columns 1 and 5 in Table 2. The first two columns use destination fixed effects from the gravity estimated using the iterative procedure (column 2 in Table 1). The next two columns use home-work commuting flows instead of daily commuting flows (see Section 2.1 for the definition). The next two columns define workplace income at the survey-area level excluding commuters whose origin towers are within 180 seconds of the destination cell tower, when we aggregate up from cell tower level. The next two columns use destination fixed effects not adjusted for Voronoi cell tower. The last two columns include commuters from DHUTS survey whose origin locations are outside the DCC area (see Section 2.1).

Table C.6: Individual Income: Model Predictions and Survey Data

	log Survey Income		
	(1)	(2)	(3)
Model log Income (workplace)	0.14***	0.05***	0.04***
	(0.02)	(0.02)	(0.01)
log Travel Time		0.12***	0.09***
		(0.02)	(0.01)
log Dest. Dist. to CBD		−0.09***	−0.03
		(0.02)	(0.02)
log Dest. Commuting Zone Area		−0.07**	−0.02
		(0.03)	(0.02)
Male			0.48***
			(0.03)
Age			0.02***
			(0.001)
Level of education			0.21***
			(0.01)
Origin FE		X	X
Occupation and Sector FE			X
Observations	11,006	11,006	11,006
Adjusted R^2	0.02	0.09	0.50

Notes: This table regresses log income from the DHUTS survey on model-predicted income and controls. The unit of observation is a survey respondent in the sample described in Table 2. Model-predicted income for a pair of origin and destination survey areas is the weighted average of tower-pair model income, with weights given by tower-to-tower commuting flows. Formally, for survey areas a and b, $y_{ab} \equiv \sum_{i \in a, j \in b} V_{ij}/V_{ab} \cdot y_j$, where $i \in a$ and $j \in b$ index towers, $y_j = \hat{\psi}_j^R$ is the area-adjusted destination fixed effect at j, and $V_{ab} \equiv \sum_{i \in a, j \in b} V_{ij}$ is the total flow between a and b. We assign to each survey respondent the predicted income between his or her home and work survey areas. Columns 2 and 3 include origin survey area fixed effects, and column 3 includes occupation and job sector fixed effects. Two-way clustered standard errors clustered by origin and destination survey area reported in parentheses. $^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$

Table C.7: Robustness: Average Residential Income and Nightlights Comparison

(A) Dhaka

	log VIIRS Nighttime Lights									
	(1) Gravity w/ Model Origin FE		(2) Home-Work Flows		(3) Excluding Neighboring Towers		(4) log of Mean Income $\epsilon = 6.84$		(5) Without Area Adjustment	
log Model Income (residential)	0.59^{***}	0.54^{***}	0.54^{***}	0.31^{***}	0.65^{***}	0.30^{***}	3.99^{***}	3.17^{***}	-1.03^{***}	-0.32^{***}
	(0.04)	(0.07)	(0.04)	(0.08)	(0.04)	(0.04)	(0.31)	(0.53)	(0.11)	(0.09)
Geographic Controls		X		X		X		X		X
Sub-district FE		X		X		X		X		X
Observations	1,868	1,868	1,854	1,854	1,867	1,867	1,868	1,868	1,868	1,868
Adjusted R^2	0.70	0.84	0.71	0.84	0.65	0.83	0.70	0.83	0.19	0.82

(B) Colombo

	log VIIRS Nighttime Lights									
	(1) Gravity w/ Model Origin FE		(2) Home-Work Flows		(3) Excluding Neighboring Towers		(4) log of Mean Income $\epsilon = 6.84$		(5) Without Area Adjustment	
log Model Income (residential)	0.83^{***}	0.50^{***}	0.88^{***}	0.31^{***}	0.90^{***}	0.39^{***}	5.50^{***}	2.93^{***}	0.60	-0.32^{***}
	(0.03)	(0.05)	(0.03)	(0.07)	(0.06)	(0.04)	(0.17)	(0.29)	(0.41)	(0.04)
Geographic Controls		X		X		X		X		X
Sub-district FE		X		X		X		X		X
Observations	1,199	1,197	1,195	1,193	1,199	1,197	1,199	1,197	1,199	1,197
Adjusted R^2	0.85	0.93	0.86	0.93	0.82	0.94	0.86	0.93	0.03	0.93

Notes. Robustness for columns 1 and 4 in Table 3 comparing model predicted residential income and nighttime lights. Odd and even columns correspond to the specifications in columns 1 and 4 in Table 3. The first two columns use destination fixed effects from the gravity estimated using the iterative procedure (column 2 in Table 1). The next two columns use home-work commuting flows instead of daily commuting flows. The next two columns define workplace income at the survey-area level excluding commuters whose origin towers are within 180 seconds of the destination cell tower. The next two columns compute model-predicted log average income (rather than average log-income), namely $\log\left(\frac{1}{N}\sum_j \exp\left(\hat{\psi}_j^R/\epsilon\right)\right)$, using $\epsilon = 6.84$ from Ahlfeldt et al. (2015) (results are not sensitive to ϵ). The next two columns use destination fixed effects not adjusted for Voronoi cell tower.

Table C.8: Robustness: Average Residential Income and Nightlights (central Dhaka only)

	log VIIRS Nighttime Lights					
	(1)	(2)	(3)	(4)	(5)	(6)
log Model Income (residential)	0.24***			0.15**	0.12***	0.09
	(0.03)			(0.07)	(0.04)	(0.06)
log Residential Density		−0.02*		−0.06***	−0.02***	−0.02**
		(0.01)		(0.02)	(0.004)	(0.01)
log Dist. to CBD			−0.16***	−0.08**	−0.16***	−0.16***
			(0.04)	(0.04)	(0.04)	(0.04)
log Model Income (workplace)				0.05		0.03
				(0.05)		(0.03)
Sub-district FE (count)					X (39)	X (39)
Observations	1,202	1,202	1,202	1,202	1,202	1,202
Adjusted R^2	0.23	0.01	0.15	0.29	0.66	0.66

Notes. Version of Table 3 restricting to cell phone towers within the Dhaka City Corporation (DCC) area, which is the area covered by survey data in Table 2. (For the spatial extent of DCC, see panel B in Appendix Figure B.1). The lower slope is consistent with the curvature visible in Appendix Figure B.5, panel A. We lose significance in the last column, which includes sub-district fixed effects, geographic controls, and the model workplace income measure.

Table C.9: Average Residential Income and Survey Income

(A) log Survey Income

	log Survey Income (residential)				
	(1)	(2)	(3)	(4)	(5)
log Model Income (residential)	0.06			−0.10	−0.12
	(0.06)			(0.09)	(0.25)
log Residential Density		0.11*		0.14**	0.14**
		(0.06)		(0.06)	(0.07)
log Dist. to CBD			−0.09*	−0.10*	−0.10
			(0.05)	(0.05)	(0.06)
log Model Income (workplace)					0.01
					(0.14)
Adjusted R2	0	0.02	0.05	0.06	0.04
Observations	73	73	73	73	73

(B) Nighttime Lights at Survey Area Level (central Dhaka only)

	log VIIRS Nighttime Lights				
	(1)	(2)	(3)	(4)	(5)
log Model Income (residential)	0.25***			0.22	0.03
	(0.09)			(0.14)	(0.24)
log Residential Density		0.15*		−0.09	−0.10
		(0.08)		(0.10)	(0.09)
log Dist. to CBD			−0.18***	−0.13**	−0.15**
			(0.03)	(0.06)	(0.06)
log Model Income (workplace)					0.12
					(0.09)
Adjusted R2	0.2	0.06	0.24	0.29	0.3
Observations	73	73	73	73	73

Notes. Versions of Table 3 at the level of the 73 survey areas with at least one observation with reported income and residence in that survey area in the DCC. Panel A uses average residential income from survey data, panel B uses nighttime lights. The low explanatory power in all columns in panel A is consistent with low signal-to-noise ratio for average residential income measured using survey income (i.e., in Dhaka, at this level of aggregation, residential income is more equally distributed compared to workplace income). The fact that results in Panel B are more statistically significant suggests that nighttime lights may be a more precise measure of average residential income. Weaker results in panel B relative to Table 3 indicate that our model weakly performs better at finer spatial units. (This is not because of the exclusion of samples outside urban core of Dhaka; see Appendix Table C.8.)

Table C.10: Impact of Hartal: Robustness with Very Frequent Callers

	(1)	(2)	(3)	(4)	(5)	(6)
	All Coefficients: % **Change** From Workday Mean					
	All Trips	Work Trips	All Trips	Work Trips	All Trips	Work Trips
Panel A. Predicted Income						
Hartal	-0.048^{***}	-0.048^{***}				
	(0.010)	(0.016)				
Friday (free day)	-0.236^{***}	-0.486^{***}				
	(0.008)	(0.015)				
Observations	$7.3 \cdot 10^6$	$7.3 \cdot 10^6$				
Panel B. Make a Trip						
Hartal	-0.038^{***}	-0.042^{***}	-0.043^{***}	-0.070^{***}	-0.046^{***}	-0.077^{***}
	(0.008)	(0.015)	(0.008)	(0.019)	(0.009)	(0.020)
Friday (free day)	-0.176^{***}	-0.462^{***}	-0.207^{***}	-0.576^{***}	-0.226^{***}	-0.632^{***}
	(0.007)	(0.014)	(0.008)	(0.018)	(0.009)	(0.019)
Hartal x Low Income			0.012^{***}	0.069^{***}	0.010^{***}	0.066^{***}
			(0.003)	(0.012)	(0.003)	(0.011)
Friday x Low Income			0.074^{***}	0.269^{***}	0.068^{***}	0.251^{***}
			(0.004)	(0.012)	(0.004)	(0.012)
Hartal x Short Commute					0.012^{***}	0.025^{***}
					(0.002)	(0.006)
Friday x Short Commute					0.061^{***}	0.188^{***}
					(0.003)	(0.005)
Observations	$7.3 \cdot 10^6$	$7.3 \cdot 10^6$	$7.3 \cdot 10^6$	$7.3 \cdot 10^6$	$7.3 \cdot 10^6$	$7.3 \cdot 10^6$
Workday Mean	0.82	0.40	0.82	0.40	0.82	0.40

Notes. Version of Table 4 restricted to the sample of commuters with commuting data on at least 75% of all days (90 out of 122 days), who are 9% of all commuters.

Table C.11: How Preference Shocks and Travel Time Affect Income: Estimated Structural Parameters

	(1)	(2)	(3)	(4)
	Full model		Constrained model ($\alpha_d = 0$)	
Shock productive α_z	0.47	0.32	0.57	0.57
	(0.06)	[3.37]	(0.07)	[0.09]
Shock distance α_d	-0.19	-0.47	0	
	(0.51)	[5.99]		
Shape parameter ϵ	7.33	9.27	6.32	6.40
	(4.22)	[38.07]	(0.78)	[0.73]
Observations	11,006	11,006	11,006	11,006
SE Method	Delta	Bootstrap	Delta	Bootstrap
Bootstrap clusters		70		70

Notes. This table reports estimates of the structural parameters that control the degree to which idiosyncratic shocks affect income (α_z), travel time affects income (α_d), and the Fréchet shape parameter ϵ, using the procedure described in Appendix A.3. We estimated equation (9) by regressing individual log survey income from the DHUTS survey on the three model-predicted terms. In columns 3 and 4, we restrict the third coefficient that corresponds to travel time to be zero ($\rho_3 = 0$). The estimates for α_z, α_d and ϵ in this table are transformations of the estimated OLS coefficients as detailed in equation (10). Columns 1 and 3 report standard errors computed using the Delta method. Columns 2 and 4 report results from 100 bootstrap runs where we cluster at the origin survey area level (70 survey areas with at least one out-commuter in DHUTS survey): the median estimate in the first row and standard errors in square parentheses.