NBER WORKING PAPER SERIES

MEASURING COMMUTING AND ECONOMIC ACTIVITY INSIDE CITIES WITH CELL PHONE RECORDS

Gabriel E. Kreindler
Yuhei Miyauchi

Working Paper 28516
http://www.nber.org/papers/w28516

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2021

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 are 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 Lauren Li, Akira Matsushita and Zhongyi Tang, who provided excellent research assistance. We sincerely thank David Atkin, Alexander Bartik, Abhijit Banerjee, Sam Bazzi, Arnaud Costinot, Dave Donaldson, Esther Duflo, Gilles Duranton, Jean-Benoit Eymeoud, Ed Glaeser, Seema Jayachandran, Sriganesh Lokanathan, Danaja Maldeniya, Melanie Morten, Ben Olken, Steve Redding, members of the LIRNEasia BD4D team, and seminar participants at MIT, LIRNEasia, NEUDC 2016, the Harvard Urban Development Mini-Conference, ADB Urban Development, and Economics Conference, UEA 2019, NBER Cities and Global Economy 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Measuring Commuting and Economic Activity inside Cities with Cell Phone Records
Gabriel E. Kreindler and Yuhei Miyauchi
NBER Working Paper No. 28516
February 2021
JEL No. C55,E24,R14

ABSTRACT

We show how to use commuting flows to infer the spatial distribution of income within a city. A simple workplace choice model predicts a gravity equation for commuting flows whose destination fixed effects correspond to wages. We implement this method with cell phone transaction data from Dhaka and Colombo. Model-predicted income predicts separate income data, at the workplace and residential level, and by skill group. Unlike machine learning approaches, our method does not require training data, yet achieves comparable predictive power. We show that hartals (transportation strikes) in Dhaka reduce commuting more for high model-predicted wage and high-skill commuters.

Gabriel E. Kreindler
Department of Economics
Harvard University
1805 Cambridge Street
Cambridge, MA 02138
and NBER
gkreindler@fas.harvard.edu

Yuhei Miyauchi
Department of Economics, Room 412A
270 Bay State Road
Boston University
Boston, Mass 02215
United States
miyauchi@bu.edu

Measures of urban economic activity at fine temporal and spatial scales are important yet rare. 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. At the same time, comprehensive new data sources on urban behavior, especially individual mobility and commuting, are becoming available worldwide. Indeed, academic research publications using call detail records (CDR) or cell phone data covered at least 19 out of 62 countries in Africa and Asia as of 2020. Meanwhile, less than 10% of the urban population in sub-saharan African countries is covered by a census of firms with wage data.[1]

In this paper, we provide a theory-based method to predict the spatial distribution of urban economic activity from commuting choices. The revealed-preference logic of our approach is simple. A core function of cities is to connect workers and jobs (Duranton and Puga 2015). 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 wage at a location based on its "attractiveness" as a commuting destination.

We formalize this intuition by building on recent urban economics models of commuting choices. In these models, 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 models developed to evaluate urban policies and transport infrastructure (Redding and Turner 2015, Redding and Rossi-Hansberg 2017). We also show how to apply a similar method when wages and the commuting elasticity differ flexibly by skill group.

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 (Björkegren 2018), and it contains phone user location for every transaction. We construct individual home and work locations by observing a user's

[1] Authors' calculations (Appendix A).

location at different times of the day over time. Commuting flows constructed this way have very fine geographic resolution and correlate strongly with commuting flows from a transportation survey from Dhaka. We use this data to estimate the model gravity equation.

Next, we assess how well this simple measure captures real differences in income. First, we show that model *workplace* income is a robust predictor of *workplace* commuter income data from a large transportation survey in Dhaka. Second, in both cities, model-predicted *residential* income is a robust predictor of a census *residential* income proxy.

To set the right benchmark for the predictive performance, we also implement a supervised learning approach (elastic net regularization) using hundreds of features from cell-phone data and geographic measures (Zou and Hastie 2005, Blumenstock et al. 2015). The model prediction, despite being computed without any training data, explains between 70% and 90% as much income variation compared to the supervised learning approach. Hence, the destination fixed effects act as a "summary statistic" for most information in cell phone data.

We also estimate and validate the model extension with multiple worker skills.

The ideal application of our income-prediction method and of the high-frequency commuting data is to trace out heterogeneous impacts of urban events and policies, such as transportation shocks, lockdowns or floods. We study *hartals*, a type of strike intended to disrupt transportation and economic activity in Bangladesh. We find larger reductions in commuting on hartal days for commuters with higher model-predicted income, high skill, and long commute.

We build on a growing literature using quantitative urban models. While papers in this literature often use gravity equations to estimate structural parameters as part of a broader exercise (Ahlfeldt et al. 2015, Monte et al. 2018, Owens et al. 2020, Tsivanidis 2019, Severen 2019, Heblich et al. 2018, Dingel and Tintelnot 2020), our focus is to use gravity equations to construct proxies of the spatial distribution of wages. Another related literature uses machine learning to empirically predict wealth and consumption at individual or regional level (Blumenstock et al. 2015, Jean et al. 2016, Glaeser et al. 2017). A key feature of our approach is that it does not require training data, relying instead on a simple and general theory of commuting behavior.

We believe that one promising path for "big data" in economics is using 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).

1 Cell-Phone Data and Commuting Flows

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.[2]

We construct commuting trips by assigning "home" and "work" locations for each user. Home (work) locations are identified as the most frequent towers with a transaction between 9pm to 5am of the next day (10am to 3pm) during weekdays excluding hartal days. For robustness, we also construct *daily* commuting trips.[3] We then aggregate over users to obtain an origin-destination (OD) matrix of commuting flows between every pair of cell towers.

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. 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. We use predicted time without traffic congestion. Using predicted time with traffic congestion in Colombo, where such data was available, yields virtually identical model-predicted wages (Table G.4).

[2] 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.

[3] To construct *daily* commuting trips, 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. If transaction data is missing in either time interval, commuting behavior is not observed for that user-day (Table G.1).

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. Home and work locations are at the level of 90 "survey areas." Our main analysis sample covers 12,510 commuters who live and work within the DCC, with positive income from work, excluding students, homemakers, and the unemployed.

Population Censuses. We use census data from 2011 in Bangladesh and 2012 in Sri Lanka, the closest years to our cell phone data. Since the census does not report income in either country, we obtain the first principal component of houshold assets (house building materials, toilet facilities, water and electricity connection) at the finest geographic unit available.[4] The residential income proxy at the cell tower level is the average across overlapping census units, weighted by overlap area with the tower.

Representativeness of Commuters in Cell Phone Data. In Dhaka, commuting flows derived from cell phone data are strongly related to those from the DHUTS commuting survey, including when controlling for log travel time, origin and destination survey area fixed effects (Table G.2). This is consistent with previous research validating cell-phone-based commuting flows (Calabrese et al. 2011, Wang et al. 2012, Iqbal et al. 2014). The decay of commuting flows with travel time is virtually identical between the two data sources (Figure G.2, Panel A).

Residential population density from cell phone data is correlated with census population density at the level of 1,866 and 1,201 cell phone towers in the two cities, with $R^2 = 0.61$ in Dhaka and $R^2 = 0.49$ in Colombo (Table G.3). The slope is 1.16 for both cities, hence cell phone data slightly over-represents population in denser areas. This type of bias does not automatically affect our results as our approach uses workplace choice shares conditional on a residential location. In section 2.4 we further allow choices to differ by worker skill.

[4] In the study areas, there are 2,381 Grama Niladhari (GN) in Sri Lanka, and 3,704 mauza in Dhaka.

2 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).

2.1 Workplace Choice Model

Space is partitioned into a finite set of locations, which may serve as both residential and work locations. We discuss how to map model locations to the data in section 2.3.

There is a unit mass of workers, and each worker ω sequentially decides her residential location (or origin) i, and then her work location (or destination) j. Conditional on residing in location i, the utility of worker ω 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 normazlied to one and shape parameter ϵ.[5] Each worker supplies one unit of labor and earns income W_j.

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[6]

$$\pi_{ij|i} = \frac{(W_j / D_{ij}^{\tau})^{\epsilon}}{\sum_s (W_s / D_{is}^{\tau})^{\epsilon}} \tag{2}$$

[5] We follow Ahlfeldt et al. (2014) and assume that $Z_{ij\omega}$ are preference shocks. Tsivanidis (2019) alternatively assumes that $Z_{ij\omega}$ are productivity shocks and derives expected income as "commuter market access". Appendix C shows that our results are robust to alternate assumptions, and estimates a model where $Z_{ij\omega}$ and D_{ij} partly affect productivity and utility.

[6] Assuming joint home and work location choice leads to the same conditional choice probabilities $\pi_{ij|i}$ as residential terms (amenities, rent) cancel out (Ahlfeldt et al. 2015). However, if workers choose their workplace first and then the home location (e.g. migrants), destination fixed effects would not exclusively capture destination wages.

Taking logs, and denoting log quantities by lowercase letters:

$$\log(\pi_{ij|i}) = \epsilon w_j - \epsilon\tau d_{ij} - \log\left(\sum_s \exp\left(\epsilon w_s - \epsilon\tau d_{is}\right)\right) \tag{3}$$

2.2 Estimating the Gravity Equation

We estimate equation (3) using the empirical Poisson pseudo-maximum likelihood (PPML) method with two-way fixed effects:

$$\log(E[\pi_{ij|i}]) = \psi_j - \beta d_{ij} + \mu_i \tag{4}$$

where μ_i and ψ_j are origin and destination fixed effects. We use PPML, rather than OLS, to deal with zero commuting flows (Silva and Tenreyro 2006, Dingel and Tintelnot 2020).[7]

This equation allows us to measure the attractiveness of each destination j after accounting for the fact that conditional commuting probabilities also depend on log commute duration. Importantly, in our model, ψ_j is proportional to the (relative) log wage at j with a factor of ϵ, the Fréchet dispersion parameter. Our main goal is to recover the ψ_j's from observed commuting choices. For this purpose, it is not necessary to model explicitly how wages are determined in equilibrium. The 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.[8]

To obtain relative wage *levels*, we further need to know ϵ, the Fréchet parameter, which governs the variance of idiosyncratic preference shocks. In section 3, we will estimate ϵ in Dhaka using data on average wages by location in the city. When such data is not available, ϵ may also be estimated from the overall variance of wages in the entire city, as in Ahlfeldt et al. (2015).

[7] Log travel time as a measure of commuting costs offers a good fit (Figure G.2).

[8] Our model does not include workplace amenities. If these differ considerably across space, the gravity destination fixed effects will capture the combined effect of wages and amenities. Our empirical results in section 3 address empirically the extent to which our measure is correlated with wages.

2.3 Mapping Model Locations to Geographic Areas

A key advantage of the model is that locations can be mapped directly to two-dimensional urban data. However, exactly how we map model locations to locations in the data may not be innocuous for inferring wages. The difficulty of defining appropriate geographic units is related to the "modifiable areal unit problem" in geography (Fotheringham and Wong 1991), and to the problem of defining urban areas in economics (Rozenfeld et al. 2011, Baragwanath et al. 2019).

We show how to solve this issue in a theory-consistent way. Assume that each Voronoi cell j consists of N_j sub-locations and workers draw independent Fréchet shocks at the levels of these sub-locations. Hence, the model implies a gravity equation at the level of sub-locations. However, a gravity equation continues to hold at the level of the larger Voronoi cells under two simplifying assumptions: that all sub-locations within j offer the same log wage w_j^R, and that all sub-location in j are located in the same place in space (and hence they are equally distant to any given origin i). Using standard Fréchet properties, the underlying log wage w_j^R is expressed as:[9]

$$\psi_j^R = \epsilon w_j^R = \psi_j - \log\left(N_j\right), \tag{5}$$

where ψ_j is the destination fixed effect from the gravity equation between Voronoi cells. In other words, we need to adjust the destination fixed effects by subtracting a term ($\log(N_j)$) that is increasing in the number of sub-locations in each cell.

Implementing this adjustment requires taking a stand on N_j, i.e. the number of independent Fréchet shocks drawn per Voronoi cell j. Theory does not offer strong guidance on how to select N_j. Here, we assume that N_j is proportional to the geographic area of Voronoi cell j, that is, Fréchet shocks are drawn for each vertex on a square lattice. We also show robustness to assuming $N_j = 1$ and to aggregating our raw data at the level of square grid cells (Tables G.6, G.9, and Figure G.5).

[9] When wages differ across sub-locations within j, ψ_j^R corresponds to a CES aggregate of these wages.

2.4 Estimating Wages Separately by Skill Group

Skill inequality is salient in cities like Dhaka and Colombo, and our baseline model does not capture this. We now extend our method to estimate wages separately by skill group.

We consider two types of models. In our first model, workers are either low-skill (L) or high-skill (H) and face separate log wage profiles w_j^L and w_j^H and elasticities of commuting cost with respect to travel time τ^L and τ^H. All agents have a common Fréchet shape parameter ϵ. As before, our goal is to estimate composite wage terms $\psi_j^s = \epsilon w_j^s$ for all j up to a constant and distance terms $\beta^s = \epsilon\tau^s$ for each skill $s \in \{L, H\}$. Equation (2) expresses commute probabilities $\pi_{ij|i}^s$ separately by skill s. The average commuting probability (that we observe in our cell phone data) is $\pi_{ij|i} = (1 - \lambda_i^H)\pi_{ij|i}^L + \lambda_i^H \pi_{ij|i}^H$ where λ_i^H is the share of high skill at residential location i (measured from census data). Because $\pi_{ij|i}$ is not log-linear in the parameters of interest, we estimate the model using maximum likelihood (Appendix B).

In our second model, each residential location i has a "representative agent" with preferences given by a weighted mean, with weights that depend on λ_i^H, the share of high skill at i. Agents from i earn log wages $w_j^H\lambda_i^H + w_j^L(1 - \lambda_i^H)$ at j and have commuting elasticity $\tau^H\lambda_i^H + \tau^L(1 - \lambda_i^H)$. Plugging into equation (4), our estimating equation becomes:

$$\log(\mathrm{E}[\pi_{ij|i}]) = \psi_j^L(1 - \lambda_i^H) + \psi_j^H\lambda_i^H - \beta^L(1 - \lambda_i^H)d_{ij} - \beta^H\lambda_i^H d_{ij} + \mu_i \tag{6}$$

This equation is intuitive: each destination has two levels of "attractiveness," $\psi_j^L = \epsilon w_j^L$ and $\psi_j^H = \epsilon w_j^H$. Loosely speaking, ψ_j^H are identified from commuting probability patterns from origins with large shares of high-skilled in the census. We estimate (6) using PPML with origin and destination fixed effects, log duration interacted with skill shares, and destination-specific linear slopes in λ_i^H.

The two models are identical when residences are completely segregated by skills ($\lambda_i^H = 0$ or 1). In the intermediate case, we show by simulation that the two data generating process are closely aligned when using the same profiles of w_j^L and w_j^H (Appendix B). Because of the simpler estimation procedure, we use the second model as our benchmark. Appendix B has results from the first model.

2.5 Estimation Results: Gravity and Wages

We estimate gravity equation (4) 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 commuting flows between 1,859 locations in Dhaka (columns 1-3) and between 1,201 locations in Colombo (columns 4-6). The gravity equation is estimated with commuting flows constructed from assigned home and work locations for 1.5 and 1 million commuters in the two cities (columns 1,3,4, and 6), and using the commuting flows identified at the daily level (columns 2 and 5).

Commuting probability decreases strongly with travel time. Interestingly, although the average commuting trip is 25% longer on average in Sri Lanka, the coefficients are similar (-2.44 in Column 1 and -2.19 in Column 4). This is a substantive finding, as the two cities differ in terms of economic development, population, and urban structure (mono- vs poly-centric). At the same time, these elasticities are substantially lower in magnitudes than in the United States (between 4 to 8, Monte et al. (2018) and Owens et al. (2020)), potentially reflecting differences in the disutility of commuting time.

Turning to our model with skill heterogeneity, we estimate equation (6) using PPML, data on aggregate bilateral commuting flows, and residential-level skill shares from the population census. High-skilled is defined as literate in Dhaka, and having secondary education in Colombo. The travel time elasticity is smaller in absolute value for the high skill group, implying that low skill workers commute closer to home. Tsivanidis (2019) finds a similar result using commuting flows data by skill. Our method only requires aggregate commuting flows and the residential skill shares.

Figure 1 displays smoothed estimated wages in Dhaka and Colombo using choropleth maps. Our estimated, area adjusted measure ψ_j^R is proportional to log wages, with factor ϵ

[10] In Dhaka, we 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 and same tower pairs are virtually identical (Table G.4).

(the Fréchet shape parameter). 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: estimating on disjoint samples (Table D.1), using daily commuting flows instead of home and work assignment, when we use travel times with congestion in Colombo, and when we include neighboring and same tower pairs in the samples (Table G.4). Using OLS instead of PPML leads to a flatter profile of destination fixed effects due to many zero commuting flows (57% of all possible tower pairs in Bangladesh and 15% in Sri Lanka). We obtain virtually identical results when allowing travel time to enter non-parametrically (as deciles of the travel time distribution).

Low- and high-skill destination fixed effects are positively correlated with the destination fixed effects from the benchmark model, with respective correlation coefficients 0.71 and 0.80 in Dhaka and 0.43 and 0.67 in Colombo. However, the model also points to independent variation by skill: low- and high-skill destination fixed effects are weakly correlated in Dhaka and *negatively* correlated in Colombo, with correlation coefficients 0.17 and -0.37, respectively.

3 Validation: Model-Predicted and Survey Workplace Income in Dhaka

Our first validation exercise compares income from the model and survey income from the DHUTS survey. We compute average income at the workplace level in each survey area in the DCC, the finest geographic location available in the DHUTS survey.[11]

Our procedure predicts that the area adjusted destination fixed effects $\hat{\psi}_j^R$ corresponds to the log labor income multiplied by ϵ, the Fréchet shape parameter of worker unobserved preferences. Panel (A) of Table 2 shows the scatter plot between $\hat{\psi}_j^R$ against survey income, at the level of 88 survey areas. We expect a slope of ϵ^{-1} and obtain a regression slope of 0.12 (with standard deviation 0.03), implying $\hat{\epsilon} = 8.3$. This estimate is slightly higher than previous estimates in Berlin (6.83; Ahlfeldt et al. 2015) and London in the 19th century (5.25;

[11] Given that government jobs are typically paid less yet include large non-monetary benefits (such as job tenure) and are centrally located, our baseline estimation sample excludes government workers. Including them does not substantially change our results (Table G.7).

Heblich et al. 2018), suggesting that idiosyncratic shocks are less important in our context.

Panel (A) shows that model-predicted wages are significantly correlated with survey wages. In Panel (B), we repeat this exercise separately by skill, using the extended model described in Section 2.4. We find that model income for one skill group predicts survey income for the same skill, with a much weaker relationship *across* skills.

Turning to predictive power, the bivariate relationship in Panel (A) has $R^2 = 0.25$ and Root-Mean Squared Error (RMSE) $= 0.22$. A perfect fit would imply $R^2 = 1$ and RMSE$= 0$ if wages are measured without noise in the DHUTS data. Our results might indicate that our model has poor predictive power, or that wages are measured with noise. We believe the latter concern is warranted. For example, 25% of survey areas have fewer than 50 observations, and the survey records total monthly income, which may include non-labor income.

To set the right benchmark given the potential for measurement error in this validation data, we compare the model prediction with a supervised-learning approach (elastic net regularization) using 498 features from cell-phone data and geographic data. We view this exercise as the state of the art method to do prediction when training data is available (Blumenstock et al. 2015). To implement this, we randomly select half of all survey areas as "training data," and predict survey income in the other half as "test data," using either OLS or elastic net regularization. See Appendix E for details.

Panel (C) of Table 2 reports the average test and training R^2 and RMSE statistics over 100 random splits. Model-predicted income alone predicts 22% of the variation in the test data (column 1). The area of the tower voronoi cell, an intuitive predictor of economic activity from cell phone data, has test $R^2 = 0.09$ (column 2).[12] Including all features from cell-phone data raises test R^2 to 0.24, a slight improvement over just using the model-predicted income (Column 3). The same pattern holds for RMSE. This result indicates that the model-predicted income (one statistic computed from cell phone data) summarizes nearly all information about predicting workplace income in this context, despite the parsimonious model and functional form specification for travel cost.

Another important factor for R^2 is geographic extent. In Section 5, we show that including

[12] Cell phone operators tend to locate more towers in locations with high activity (Figure G.1).

peri-urban areas drastically increases R^2.[13]

Severen (2019) and Tsivanidis (2019) perform validation exercises for commuting-choice models similar to ours, estimated using commuting data from survey or administrative data. Using fine divisions (census tracts), Severen (2019) finds that model wages estimated using commuting flows barely predict tract-level wages in Los Angeles. Tsivanidis (2019) finds that model-predicted wages across 19 urban areas in Bogotá predict survey wages, with an $R^2 = 0.36$. These results suggest that the geographic aggregation level may also matter. However, in our setting, aggregating up does not change predictive power significantly (Figure G.5).

Table G.5 shows that the model income is statistically significantly correlated with survey income after controlling for employment density, distance to the central business district, and log model *residential* income. Table G.6 shows that our results are robust to several alternate gravity equation specifications. Table G.7 uses an individual-level specification and shows that our main result is robust to controlling for workplace sorting along observable worker characteristics and other controls.

4 Validation: Model-Predicted Income and Residential Income Proxies

We next use a residential income proxy constructed from population census data to validate the model prediction at the residential location level. Model-predicted residential income at tower i is defined as

$$\sum_j \hat{\psi}_j^R V_{ij} / V_i \tag{7}$$

where j indexes workplace towers, $\hat{\psi}_j^R$ is the area adjusted destination fixed effect at j, V_i is total residential population at i, and V_{ij} is the commuting volume from i to j. We focus on model fit and not the magnitude of the slope in this exercise, because the income proxy is not measured in the same units as actual income.

Table 3 shows the results in Dhaka and Colombo. Model residential income is a strong predictor of the income proxy at the cell tower level (panel A). The $R^2 = 0.55$ (Dhaka) and

[13] In a machine learning application, Blumenstock et al. (2015) finds $R^2 = 0.41$ when restricting to 37 urban DHS clusters in Rwanda, compared to 0.62 when including both rural and urban clusters. Jean et al. (2016) do not report results separately by urban areas. Within entire countries, R^2 ranges between 0.37 and 0.55 for DHS-cluster level predicted consumption.

0.77 (Colombo) is high, partly because of the coverage of suburban areas.

We next benchmark the predictive power to a supervised learning method (panel (B) Table 3) in Dhaka.[14] The procedure is similar to Panel (B) in Table 2. Test R^2 is 0.55 when using model-predicted income alone (column 1). The test R^2 when using the cell phone tower Voronoi cell area alone is 0.71 (column 2), and the supervised-learning method using all features increases it to 0.73. Model-predicted residential income alone achieves about 75% of the predictive power of using all the cell-phone data metrics.

5 Practical Guidance: Spatial Coverage and Resolution

What factors affect the predictive performance of our method? We explore this issue using model-predicted residential income and census income proxy data to guide future applications of this method.

Our approach performs better when the analysis includes both urban core and peri-urban areas. In both cities, the explanatory power of model residential income increases as we include areas further away from the city center. Adjusted R^2 goes from 0.2 at the 10 km cutoff to 0.5 and 0.7 at 30 km cutoff in Dhaka and Colombo, respectively (Figure G.3).[15]

Model performance is not sensitive to the level of spatial aggregation. We combine cell phone towers into square grid cells, and estimate the gravity equation at this level. In both cities, regressing the census income proxy on model-predicted residential income yields stable adjusted R^2 between 0.6 and 0.9 for grid cells of between 2 and 10km wide (Figure G.5). (Gravity distance slopes are also stable.)

Overall, the method is best suited to applications that cover a large urban area, and is likely not sensitive to the spatial aggregation unit.

[14] Unfortunately, we do not have access to features of cell phone data to implement the supervised learning approach in Colombo.

[15] Related, average residential income from the (DHUTS) survey data – only available in Dhaka's urban core – is difficult to predict, both using model residential income, and using other measures such as distance to CBD or residential density.

6 Application: The Heterogeneous Impacts of Hartal on Commuting

Cities in developing countries experience frequent shocks that disrupt commuting, such as floods, transportation strikes, protests or violence. Measuring how different types of commuters are affected by such shocks is a key step in understanding volatility in urban economic activity. We now show how to use high-frequency cell phone data combined with our method to study the heterogeneous impacts of urban shocks.

We study this question in the context of hartal, 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 by unions or political groups, groups of people enforce the transportation shutdown, especially on major roads and in certain locations. There were 33 hartal days over the 4 months in our sample (Ahsan and Iqbal 2015).[16]

We estimate the impact of hartal on commuting to work and heterogeneity using the following specification:

$$C_{\omega ijt} = \beta \cdot hartal_t + \beta_W \cdot hartal_t \cdot \overline{\psi}_j + \beta_D \cdot hartal_t \cdot \overline{d}_{ij} + \mu_\omega + \eta_{m(t)} + \varepsilon_{\omega ijt} \qquad (8)$$

where $C_{\omega ijt}$ is a dummy for whether commuter ω with home location i and work location j travelled from home to work on day t, and $hartal_t$ is a dummy for hartal dates.

We first focus on the interactions between hartal and $\overline{\psi}_j$, the standardized area adjusted wage at j, and $\overline{d}_{ij}$, standardized log commute duration between i and j. We estimate $\overline{\psi}_j$ using commuting data on non-hartal weekdays following the procedure in Section 2. We include month fixed effects $\eta_{m(t)}$ and commuter fixed effects μ_ω, to account for potential differences in calling behavior on hartal days that may affect the measure of commuting.[17]

Table 4 shows the results. Commuting to work falls by 7.7% on Hartal days (column 1). The magnitude of the effect is consistent with previous studies on hartals in more specific settings (Ashraf et al. 2015, Ahsan and Iqbal 2015). Appendix F reports robustness exercises.

[16] The study period preceded parliamentary elections and was marked by general instability and more frequent hartals than in previous years.

[17] Results are similar for a sample of frequent callers (Table F.1).

Commuters working in high-productivity areas (high $\overline{\psi}_j$) are more sensitive to hartal disruptions (column 2).[18] Specifically, commuters with model-predicted wage one standard deviation above the mean reduce their commuting to work by 10.6%, an effect that is 26% larger than the average effect. This could be due to more disruption of production activity in more productive areas, or due to higher commuting cost (for example, due to the danger of physical violence) for higher-income commuters. Distinguishing between these mechanisms is an interesting question beyond the scope of this paper.[19] Hartal affects long-distance commuters more: a standard deviation longer commute is associated with a 2.7pp more negative hartal effect. However, the coefficient on the destination wage barely changes, showing that the productivity effect is mostly independent of commute duration.

We now use our method to study how hartal affects workers of different skills. We use model-derived wages for low- and high-skilled ($\overline{\psi}_j^L$ and $\overline{\psi}_j^H$) estimated in section 2.5 and standardized. Assume that the average hartal treatment effect for commuter ω of skill $s \in \{L, H\}$ is $\beta^s + \beta_D^s \overline{d}_{ij} + \beta_W^s \overline{\psi}_j^s$. This yields the following specification

$$\begin{aligned} C_{\omega ijt} =& \beta^L \cdot hartal_t \cdot (1 - \widehat{\lambda}_{ij}^H) + \beta^H \cdot hartal_t \cdot \widehat{\lambda}_{ij}^H \\ & \beta_W^L \cdot hartal_t \cdot \overline{\psi}_j^L \cdot (1 - \widehat{\lambda}_{ij}^H) + \beta_W^H \cdot hartal_t \cdot \overline{\psi}_j^H \cdot \widehat{\lambda}_{ij}^H + \\ & \beta_D^L \cdot hartal_t \cdot \overline{d}_{ij} \cdot (1 - \widehat{\lambda}_{ij}^H) + \beta_D^H \cdot hartal_t \cdot \overline{d}_{ij} \cdot \widehat{\lambda}_{ij}^H + \\ & \mu_\omega + \eta_{m(t)} + \epsilon_{\omega ijt} \end{aligned}$$

where we interact the hartal terms in (8) with the predicted share of high-skilled among commuters from i to j, $\widehat{\lambda}_{ij}^H = \widehat{V}_{ij}^H / (\widehat{V}_{ij}^L + \widehat{V}_{ij}^H)$, where $\widehat{V}_{ij}^s = V_i^s \widehat{\pi}_{ij|i}^s$ is the model-predicted commuting flow of skill s between i and j, and V_i^s is population of skill s from the census.

For both skills, commuting to work is negatively affected by hartal, and the effect is stronger for commuters with higher (skill-specific) wages (column 4). The high-skilled are more affected on average, yet the low-skill hartal effect is more sensitive to (low-skilled) des-

[18] Mean-reversion is a potential concern because destination fixed effects are estimated on non-hartal days. To investigate this, we ran a placebo hartal exercise, estimating the gravity equation for a dummy switched on for a random set of weekdays. Coefficients in Table 4 become an order of magnitude smaller.

[19] Remote work is unlikely to explain this pattern. Dingel and Neiman (2020) estimate that in 2020, only 12% of jobs in Bangladesh can be done entirely at home.

tination wage. Hence, the low-wage, low-skilled are least affected.

The distance interactions are both negative and statistically indistinguishable. Skill differences documented above are likely not due to commuting technologies, such as access to private vehicles.

Overall, we find consistent results of stronger hartal disruptions for high productivity and high-skill workers in Dhaka. These results illustrate how our methods to infer the spatial distribution of income can be used to measure heterogeneous effects of high-frequency urban events.

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 (2014): "The Economics of Density: Evidence from the Berlin Wall," *NBER Working Paper Series*.

——— (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.

Baragwanath, K., R. Goldblatt, G. Hanson, and A. K. Khandelwal (2019): "Detecting urban markets with satellite imagery: An application to India," *Journal of Urban Economics*, 103173.

Björkegren, D. (2018): "The Adoption of Network Goods: Evidence from the Spread of Mobile Phones in Rwanda," *The Review of Economic Studies*, 86, 1033–1060.

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.

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

Dingel, J. and F. Tintelnot (2020): "Spatial Economics for Granular Settings," *NBER working Paper*, 27287.

Dingel, J. I. and B. Neiman (2020): "How Many Jobs Can be Done at Home?" Working Paper 26948, National Bureau of Economic Research.

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

Duranton, G. and D. Puga (2015): "Chapter 8 - Urban Land Use," in *Handbook of Regional and Urban Economics*, ed. by G. Duranton, J. V. Henderson, and W. C. Strange, Elsevier, vol. 5 of *Handbook of Regional and Urban Economics*, 467 – 560.

Fotheringham, A. S. and D. W. Wong (1991): "The modifiable areal unit problem in multivariate statistical analysis," *Environment and planning A*, 23, 1025–1044.

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*.

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

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.

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.

Monte, F., S. Redding, and E. Rossi-Hansberg (2018): "Commuting, Migration and Local Employment Elasticities," *American Economic Review*, 108, 3855–3890.

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

Owens, R., E. Rossi-Hansberg, and P. D. Sarte (2020): "Rethinking Detroit," *American Economic Journal: Economic Policy*, 12, 258–305.

Redding, S. J. and E. Rossi-Hansberg (2017): "Quantitative Spatial Economics," *Annual Review of Economics*, 9, 21–58.

Redding, S. J. and M. A. Turner (2015): "Transportation Costs and the Spatial Organization of Economic Activity," in *Handbook of Regional and Urban Economics*, 5, 1339–1398.

Rozenfeld, H. D., D. Rybski, X. Gabaix, and H. A. Makse (2011): "The Area and Population of Cities: New Insights from a Different Perspective on Cities," *American Economic Review*, 101, 2205–25.

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.

Steele, J. E., P. R. Sundsøy, C. Pezzulo, V. A. Alegana, T. J. Bird, J. Blumenstock, J. Bjelland, K. Engø-Monsen, Y. A. De Montjoye, A. M. Iqbal, K. N. Hadiuzzaman, X. Lu, E. Wetter, A. J. Tatem, and L. Bengtsson (2017): "Mapping poverty using mobile phone and satellite data," *Journal of the Royal Society Interface*, 14.

Tsivanidis, N. (2019): "Evaluating the Impact 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.

Zou, H. and T. Hastie (2005): "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, 67, 301–320.

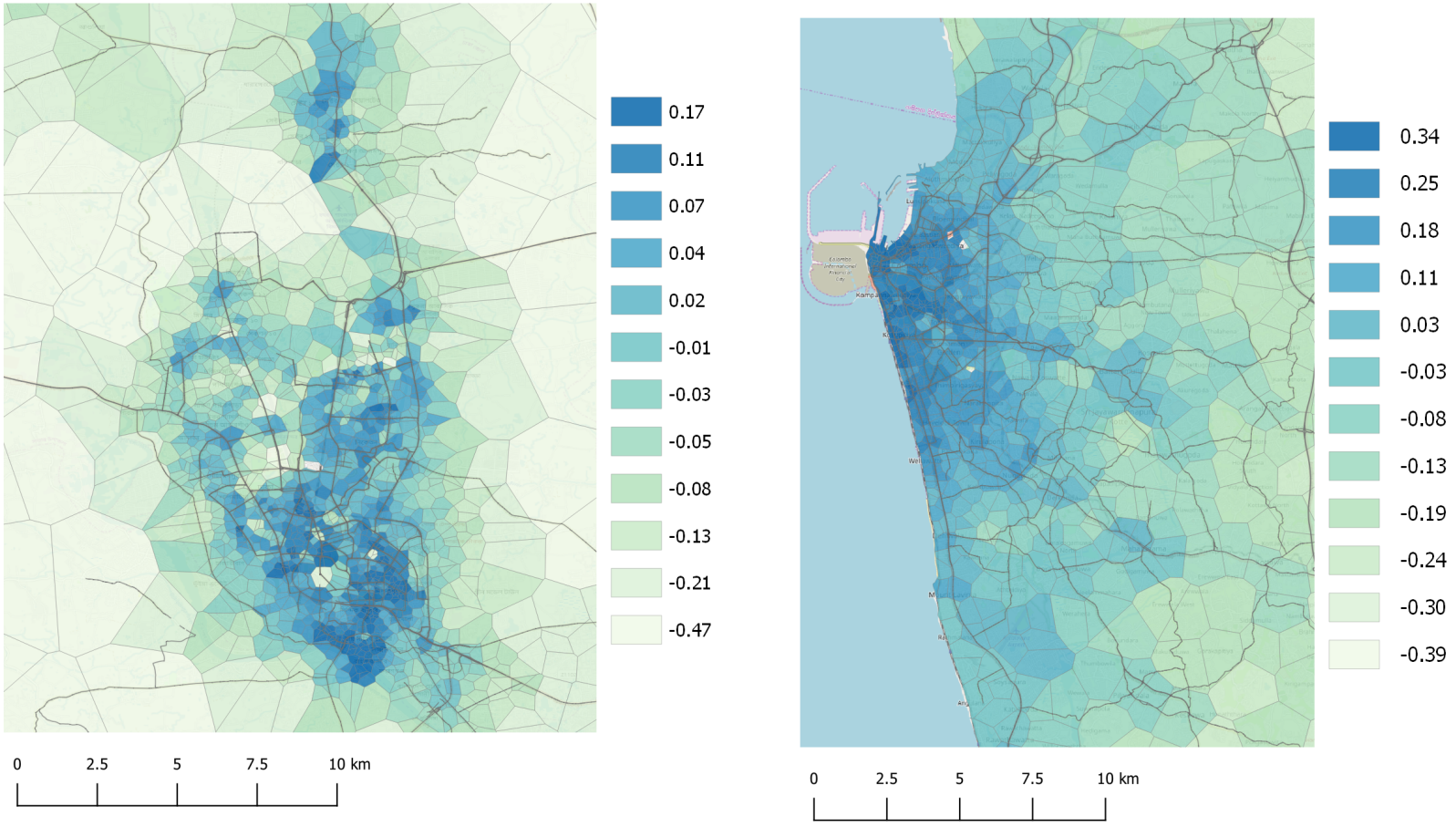
Figures and Tables

Table 1: Gravity Equation Estimation Results

	Commuting Probability					
	(1)	(2)	(3)	(4)	(5)	(6)
log Travel Time	-2.44	-2.55		-2.19	-2.22	
	(0.0011)	(0.0003)		(0.0015)	(0.0001)	
log Travel Time × Low Skill			-3.68			-5.00
			(0.11)			(0.36)
log Travel Time × High Skill			-1.91			-1.57
			(0.04)			(0.08)
City	Dhaka	Dhaka	Dhaka	Colombo	Colombo	Colombo
Commuting Measure	Home-Work	Daily	Home-Work	Home-Work	Daily	Home-Work
Number of Destination FE	1,859	1,868	1,859	1,201	1,201	1,201
Number of Trips	1.5e+6	1.9e+7	1.5e+6	9.4e+5	1.3e+8	9.4e+5
Observations	3.4e+6	3.4e+6	3.4e+6	1.3e+6	1.3e+6	1.3e+6
Pseudo R^2	0.67	0.82		0.66	0.88	

Notes. This table reports estimates of the gravity equations (4) and (6) by Poisson pseudo-maximum likelihood (PPML) method with two-way fixed effects. The outcome variable is commuting probability $\pi_{ij|i}$ from cell phone tower i to j conditional on residing in i, computed from cell phone data and aggregated over weekdays. In Bangladesh, we exclude hartal days. Commuting flows are constructed from assigned home and work locations (columns 1,3,4,6) and using the commuting flows identified at the daily level (columns 2 and 5) using cell phone data as described in Section 1. 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. High-skilled is defined as literate in Dhaka (67% of the population), and having secondary education or more in Colombo (80% of the population). 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



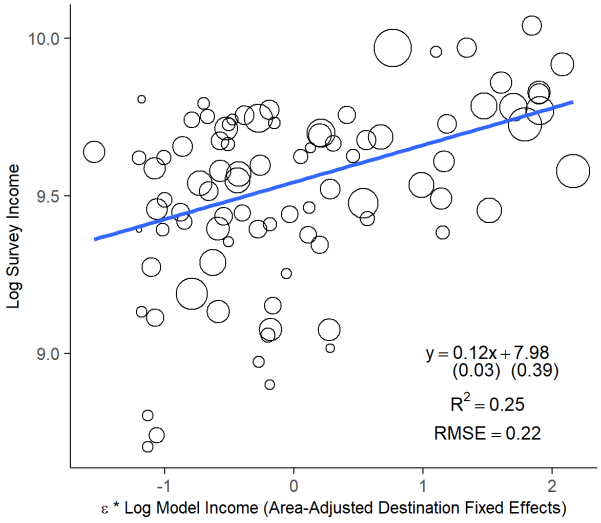
(A) Dhaka

(B) Colombo

Notes. These figures plot our model measure of log wages, the area-adjusted destination fixed effects $\hat{\psi}_j^R$ divided by the Fréchet shape parameter ϵ, at the level of cell phone tower Voronoi cells in Dhaka and Colombo. We use $\epsilon = 8.3$, as estimated in section 3. Log wages are kernel smoothed with an adaptive kernel bandwidth (proportional to the radius of the equivalent-area circle of the Voronoi cell.

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

(A) Raw Correlation



(B) Raw Correlation By Skill

	log Survey Income (workplace)					
	Low Skill			High Skill		
	(1)	(2)	(3)	(4)	(5)	(6)
ϵ × log Model Income (workplace)						
Low-Skilled	0.06***		0.06***	0.01		-0.01
	(0.02)		(0.02)	(0.01)		(0.01)
High-Skilled		0.04	-0.00		0.07***	0.08***
		(0.03)	(0.03)		(0.01)	(0.02)
Observations	87	87	87	87	87	87
Adjusted R2	0.09	0.01	0.08	0.00	0.23	0.23
RMSE	0.30	0.32	0.31	0.17	0.15	0.15

(C) Comparison with supervised learning using features derived from cell-phone data

Features	(1) log Model Income (workplace)	(2) log Tower Area	(3) All CDR Features	(4) (3) + log Model Income (workplace)
Training R^2	0.26	0.16	0.44	0.44
Training RMSE	0.20	0.22	0.17	0.17
Test R^2	0.22	0.09	0.24	0.24
Test RMSE	0.22	0.24	0.22	0.22
Observations	88	88	88	88

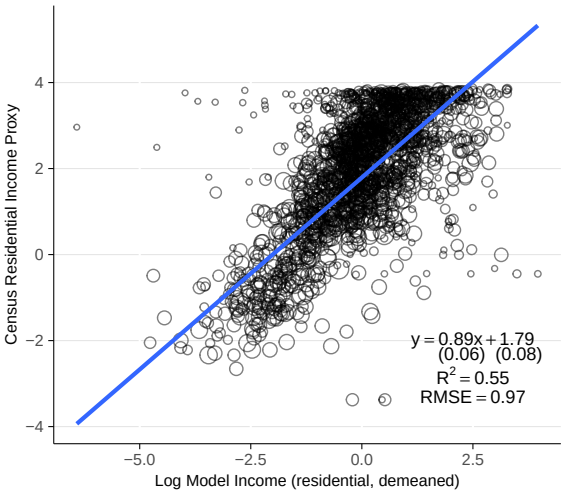
Notes. This table compares survey and model predictions of average workplace income. The slopes in panels (A) and (B) identify ϵ^{-1}, the inverse Fréchet shape parameter. The unit of analysis is a survey area from the DHUTS survey. The survey sample is 11,006 commuters who live and work inside the Dhaka City Corporation, who report positive income, excluding students, homemakers, the unemployed, and government workers. The outcome variable is the average income of survey respondents who work in a survey area, using log income truncated at the 99th percentile. In panel (B) we compute average survey income separately by skill; high skill is defined as literate (67% of the population). 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 (V_{ij} is the commuting volume from i to j). In panel (B) we use skill-specific destination fixed effects estimated using equation (6) and skill-specific predicted volume, defined as $V_{ij}\frac{\hat{V}_{ij}^L}{\hat{V}_{ij}^L+\hat{V}_{ij}^H}$ where $\hat{V}_{ij}^s$ is predicted commuting volume of skill s. Regressions in panels (A) and (B) are weighted by survey area employment population from the DHUTS survey (skill-specific in panel B).

In Panel (C), Test (Training) R^2 and RMSE indicate the average R^2 in the test (training) data over 100 random splits. See Appendix E for the description of the supervised learning method (elastic-net regularization) and cell phone data feature construction.

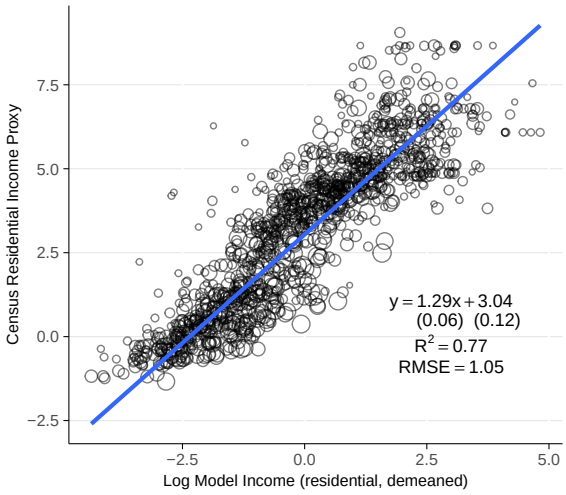
Table 3: Average Residential Income: Model Prediction and Residential Income Proxy

(A) Raw Correlation

(i) Dhaka



(ii) Columbo



(B) Comparison with supervised learning using features derived from cell-phone data (Dhaka)

Features	(1) log Model Income (residential)	(2) log Tower Area	(3) All CDR Features	(4) (3) + log Model Income (residential)
Training R^2	0.56	0.71	0.77	0.80
Training RMSE	0.93	0.75	0.67	0.63
Test R^2	0.55	0.71	0.73	0.73
Test RMSE	0.94	0.75	0.72	0.72
Observations	1844	1844	1844	1844

Notes. This table compares a census proxy and model predictions of average residential income. The unit of analysis is a cell phone tower. Income proxy is the first principal component of census residential assets (weighting each census block by its area overlap with the Voronoi cell). Average model residential (take-home) income at tower i is $\sum_j y_j V_{ij} / V_i$ where j indexes workplace towers, $y_j = \hat{\psi}_j^R$ is the area adjusted destination fixed effect at j, V_i is total residential population at i, and V_{ij} is the commuting volume from i to j. Regressions in both panels are weighted by tower residential population (from cell phone data). Panel (B) repeats the analysis in Table 2 panel (B). See Appendix E for details. Table G.8 shows that model residential income is statistically significantly correlated with survey income after controlling for residential density, distance to the central business district, and model workplace income.

Table 4: The Heterogeneous Impacts of Hartal on Commuting

	Work Commute (% change vs weekday)			
	(1)	(2)	(3)	(4)
Hartal	-0.077***	-0.078***	-0.078***	
	(0.004)	(0.004)	(0.004)	
Interactions: Hartal ×				
(β^L) % Low Skill				-0.050***
				(0.010)
(β^H) % High Skill				-0.075***
				(0.004)
Dest. FE (z)		-0.028***	-0.025***	
		(0.005)	(0.005)	
(β^L_W) % Low Skill × Dest. FE Low Skill (z)				-0.051***
				(0.013)
(β^H_W) % High Skill × Dest. FE High Skill (z)				-0.014***
				(0.005)
Log Duration (z)			-0.027***	
			(0.002)	
(β^L_D) % Low Skill × Log Duration (z)				-0.032***
				(0.007)
(β^H_D) % High Skill × Log Duration (z)				-0.026***
				(0.002)
Commuter FE	X	X	X	X
P-value $\beta^L = \beta^H$				0.02
P-value $\beta^L_W = \beta^H_W$				0.01
P-value $\beta^L_D = \beta^H_D$				0.46
Observations	1.5e+07	1.5e+07	1.5e+07	1.5e+07

Notes. This table reports the impact of hartal days on the probability to commute to work. The sample includes all commuters with distinct home- and work- locations (35% of all users), who travel at least once during hartal and non-hartal days. The sample is all days with data on commuting (including days when the user is observed to not travel), excluding holidays and weekends. All specifications include commuter and month fixed effects. The outcome is normalized to 1 on weekdays, so coefficients represent percentage changes on hartal days relative to weekdays. All variables interacted with hartal are standardized. The destination fixed effects, as well as skill-specific destination fixed effects are estimated in Table 1. The share of high skilled among i to j commuters is $\widehat{\lambda}^H_{ij} = \widehat{V}^H_{ij} / (\widehat{V}^L_{ij} + \widehat{V}^H_{ij})$ where $\widehat{V}^H_{ij} = V^H_i \widehat{\pi}^H_{ij|i}$, and V^H_i is the high-skilled residential population at i (measured from census data), and $\widehat{\pi}^H_{ij|i}$ is the model-predicted commuting probability for high-skilled. Table F.1 shows similar results when restricting to a sample of frequent callers.

Online Appendix for "Measuring Commuting and Economic Activity inside Cities with Cell Phone Records" (Gabriel Kreindler and Yuhei Miyauchi)

Table of Contents

A Conventional and Modern Data Availability in Developing Countries

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.

To illustrate, we document data availability for the 27 largest countries in Sub-Saharan Africa (accounting for $> 95\%$ of the population in the region). We chose this region as it is undergoing rapid urban growth and urban transformation.[1]

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.)

By contrast, big data that can be used to construct urban commuting flows is increasingly available in developing countries.

To illustrate, we identified 19 countries out of 62 countries in South Asia and Africa, where call detail record (CDR) data have been used in academic papers.[2] In other contexts, public transport transactions, geolocated tweets or other smartphone app location data, may be used to construct urban commuting flows. Smartphone location data is becoming increasingly popular and available to researchers, and even more so since the onset of the Covid-19 pandemic.

B Model Extension and Estimation: Worker Skill Heterogeneity

In section 2.4 we introduced two model extensions with skill heterogeneity. In this section we provide additional detail on model derivation, estimation, and simulation results. We also

[1] 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.

[2] In August 2020, we searched on Google Scholar using the following keywords "call detail records" and the country name.

present results from the first model.

First, we derive the expression for aggregate commuting flows in the first model. Assume that workers are either low-skill L or high-skill H. The two skills face different log wage profiles $(w_j^L)_j$ and $(w_j^H)_j$ and different commuting elasticities τ^L and τ^H, and have the same Fréchet shape parameter ϵ. Equation (3) now holds separately by skill, and the aggregate commuting flow is Poisson distributed with mean given by:

$$\mathrm{E}\, V_{ij} = V_i^L \frac{\exp(\epsilon w_j^L - \epsilon \tau^L d_{ij})}{\sum_s \exp(\epsilon w_s^L - \epsilon \tau^L d_{is})} + V_i^H \frac{\exp(\epsilon w_j^H - \epsilon \tau^H d_{ij})}{\sum_s \exp(\epsilon w_s^H - \epsilon \tau^H d_{is})} \tag{B.1}$$

In our data we observe *aggregate* commuting flows V_{ij}, not separately by skill. However, equipped with census data on V_i^L and V_i^H, the low- and high-skill residential populations at i, we can estimate $\psi_j^s = \epsilon w_j^s$ and $\beta^s = \epsilon \tau^s$ for all j and $s \in \{L, H\}$ in (B.1). We use maximum likelihood and implement a standard gradient ascent algorithm that has good convergence properties, yet is not guaranteed to find the global maximum.

In the second model, we assume that a representative commuter has preferences given by a weighted mean with weights given by the skill shares at their residential location. Agents from i earn log wages $w_j^H \lambda_i^H + w_j^L (1 - \lambda_i^H)$ and have commuting elasticity $\tau^H \lambda_i^H + \tau^L (1 - \lambda_i^H)$. Plugging into equation (4), the estimating equation becomes:

$$\log(\mathrm{E}[\pi_{ij|i}]) = \psi_j^L (1 - \lambda_i^H) + \psi_j^H \lambda_i^H - \beta^L (1 - \lambda_i^H) d_{ij} - \beta^H \lambda_i^H d_{ij} + \mu_i \tag{B.2}$$

Before applying these methods on real data, we explore their performance on data that is simulated based on (B.1), and using the geographic structure in Dhaka. Both methods perform well to recover underlying parameters. Table B.1 shows that the distance slopes are broadly accurate, and that the log-linear specification (B.2) disentangles the two vectors of destination fixed effects to a great extent (although not perfectly). Indeed, the off-diagonal terms in columns 2 and 3 are smaller than the diagonal terms. However, the log-linear specification performs worse for low-skilled workers.

We next estimate the gravity equation with two skills. In Dhaka, we define high-skill as literate, and use the fraction of population that is literate from the population census. (Overall, 67% of the population is literate. Interpolated at the tower level, this fraction ranges from 31% to 100% with a mean of 76% and standard deviation 11%.) In Colombo, we define high-skill as having secondary education or more. (Overall, 80% of the population has a secondary education. At the tower level, this fraction ranges from 57% to 95% with a mean of 82% and standard deviation 6%.)

Table B.2 reports results from the gravity equation with two skills. Columns 1 and 3 replicate columns 3 and 6 in Table 1. In both countries, the high skilled have a shallower slope on travel time. This could be due to a lower disutility of distance (e.g. if high-skilled can afford faster or more convenient travel modes).

Table B.3 replicates the validation exercise from Table 2 by skill, at the level of 87 survey area in the DCC that appear in the DHUTS survey. The one-skill model income predicts both low-skilled and high-skilled survey income, with higher R^2 for the latter. Using the log-linear gravity equation, low-skilled model income predicts low-income survey income. Importantly, columns 4 and 5 show that the "off-diagonal" terms are zero, meaning that this method is successful in discriminating between low- and high-skill wage patterns. Using the maximum likelihood estimate of the gravity equation, only the high-skilled model income is positively predictive of survey income (columns 7-10).

Table B.1: Numerical Simulation Check: Estimating Gravity with Two Skill Groups

	(1)	(2)	(3)	(4)	(5)	(6)
Estimation Method:	Pooled	Log-linear		MLE		
Outcome	$\hat{\psi}_j$	$\hat{\psi}_j^L$	$\hat{\psi}_j^H$	$\hat{\psi}_j^L$	$\hat{\psi}_j^H$	
Panel A. Destination Fixed Effects						
True Low Skill FE ψ_j^L	0.26***	0.73***	0.03*	0.95***	0.01	
	(0.01)	(0.04)	(0.02)	(0.01)	(0.00)	
True High Skill FE ψ_j^H	0.71***	0.25***	0.96***	0.00	1.00***	
	(0.01)	(0.04)	(0.02)	(0.01)	(0.00)	
Observations	1,840	1,840	1,840	1,859	1,859	
Adjusted R2	0.95	0.54	0.88	0.94	0.99	
Panel B. Distance Slopes						
Estimation Method:	Pooled	Log-linear		MLE		True parameter
log Travel Time	-2.43					
log Travel Time × Low Skill		-3.61		-4.00		-4.00
log Travel Time × High Skill		-1.98		-2.00		-2.00

Notes. This table uses simulated data to compare estimated parameter values with true values. Data is simulated for the 1,859 towers in Dhaka and the actual skill-specific population. Destination fixed effects for the two skills are the sum of a common normal component (sd=0.8) and a skill-specific component (sd=0.4). Commuting flows are drawn from a Poisson distribution with mean given by (B.1). The distance slopes for low- and high-skilled are $\beta^L = \epsilon\tau^L = -4$ and $\beta^H = \epsilon\tau^H = -2$. The first column shows results from the pooled regression (3) estimated with PPML. The next two columns use the log-linear specification (B.2) estimated with PPML. The last two columns use maximum likelihood estimates of (B.1), using zero as initial values for both sets of destination fixed effects. Panel A regresses estimated low- and high- skill destination fixed effects on the true values. Panel B reports the estimate (and true) distance coefficients.

Table B.2: Gravity Equation with Skills: Estimation Results

	(1)	(2)	(3)	(4)
Low Skill × log Travel Time	-3.68***	-3.88	-5.00***	-3.20
	(0.11)		(0.36)	
High Skill × log Travel Time	-1.91***	-2.10	-1.57***	-2.10
	(0.04)		(0.08)	
City	Dhaka	Dhaka	Colombo	Colombo
Estimation Method	Log-Linear	MLE	Log-Linear	MLE
Corr($\hat{\psi}_j^{L,MLE}, \hat{\psi}_j^{L,LL}$)		0.80		0.90
Corr($\hat{\psi}_j^{H,MLE}, \hat{\psi}_j^{H,LL}$)		0.86		0.77
Number of Destination FE	1,859	1,859	1,201	1,201
Number of Trips	1.5e+6	1.5e+6	9.4e+5	9.4e+5
Observations	3.4e+6	3.4e+6	1.3e+6	1.3e+6

Notes. This table estimates the gravity equation for models with two skill groups. Columns 1 and 3 estimate the log-linear equation (B.2) using PPML, while columns 2 and 4 estimate equation (B.1) using a custom gradient ascent algorithm and maximum likelihood, where destination fixed effect are initialized at the values from columns 1 and 3, respectively. Columns 2 and 4 report in each country the correlation between the destination fixed effects obtained from the two methods, for each skill group.

Table B.3: Average Workplace Income by Skill: Model Prediction and Survey Data in Dhaka

	Outcome: log Survey Income (workplace)									
	(1) Low	(2) High	(3) Low	(4) High	(5) Low	(6) High	(7) Low	(8) High	(9) Low	(10) High
Explanatory variables: $\epsilon\times$ log Model Income (workplace)										
Pooled	0.09**	0.08***								
	(0.04)	(0.02)								
Log-linear Low			0.06***		0.06***	-0.01				
			(0.02)		(0.02)	(0.01)				
Log-linear High				0.07***	-0.00	0.08***				
				(0.01)	(0.03)	(0.02)				
MLE Low							0.02		-0.08	-0.05**
							(0.03)		(0.05)	(0.02)
MLE High								0.07***	0.15**	0.11***
								(0.01)	(0.06)	(0.02)
Observations	87	87	87	87	87	87	87	87	87	87
Adjusted R2	0.07	0.21	0.09	0.23	0.08	0.23	-0.00	0.21	0.08	0.25

Notes. This table compares survey and model predictions of average workplace income, by skill. See notes for Table 2. In the DHUTS survey, low-skilled is defined as at most primary school (38% of all commuters).

C Model Extension and Estimation: Preference or Productivity Shocks

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 affect income instead. We show that the model income continues to be correlated with survey income. Furthermore, we develop a method to estimate *the degree* to which $Z_{ij\omega}$ and D_{ij} affect income instead of preferences.

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:

$$\begin{aligned}
\mathbb{E}y_{ij\omega}^{0,0} &= w_j \\
\mathbb{E}y_{ij\omega}^{1,1} &= \frac{1}{\epsilon}\log\left(\sum_s \exp\left(\epsilon w_s - \epsilon\tau d_{is}\right)\right) - K \\
\mathbb{E}y_{ij\omega}^{0,1} &= w_j - \tau d_{ij} \\
\mathbb{E}y_{ij\omega}^{1,0} &= \mathbb{E}y_{ij\omega}^{1,1} + \tau d_{ij}
\end{aligned} \tag{C.1}$$

where K is a constant term that does not depend on locations.

When neither shocks nor travel time affect income (as assumed in our main specification and in Ahlfeldt et al. 2014), income is simply the destination wage. In the second case, travel time and income affect income but not preference directly (as assumed in Tsivanidis 2019). In this case, the expression for expected income has the form of "commuting market access" (Tsivanidis 2019).

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

$$y_{ij\omega}^{\alpha_z,\alpha_d} = \alpha_z\alpha_d \cdot y_{ij\omega}^{1,1} + \alpha_z(1-\alpha_d)\, y_{ij\omega}^{1,0} + (1-\alpha_z)\,\alpha_d \cdot y_{ij\omega}^{0,1} + (1-\alpha_z)(1-\alpha_d)\, y_{ij\omega}^{0,0}. \quad \text{(C.2)}$$

Using (C.1) and dropping the constant K, this simplifies to

$$\mathbb{E}y_{ij\omega}^{\alpha_z,\alpha_d} = \frac{\alpha_z}{\epsilon}\left[\log\left(\sum_s \exp\left(\epsilon w_s - \epsilon\tau d_{is}\right)\right) + \epsilon\tau d_{ij}\right] + \frac{1-\alpha_z}{\epsilon}\left[\epsilon w_j\right] + \frac{\alpha_d}{\epsilon}\left[-\epsilon\tau d_{ij}\right] \quad \text{(C.3)}$$

Validation for the four extreme cases. Table C.1 shows the results of the OLS regression of average workplace survey income on model workplace income under the four extreme cases in equation (C.3). In all of these regressions, we expect the slope of ϵ^{-1}.

In all cases, we find that model income is significantly correlated with survey income. In terms of the model fit (R^2 and RMSE), we find the best fit when $Z_{ij\omega}$ is on income and D_{ij} is on preference (Column 3). Our baseline assumption (both $Z_{ij\omega}$ and D_{ij} are on preference; Column 1) performs the second, followed by the case with both $Z_{ij\omega}$ and D_{ij} are on income; Column 2). We also find a larger regression slope in Column (2). This indicates that the estimates of ϵ may differ depending on the model assumptions.

Table C.1: Robustness of Workplace Income Validation with Different Assumptions on Idiosyncratic Shocks and Travel Cost

	log Survey Income (workplace)			
	(1)	(2)	(3)	(4)
$\epsilon\times$ log Model Income (workplace)	0.12^{***}	0.22^{***}	0.12^{***}	0.12^{***}
	(0.03)	(0.06)	(0.03)	(0.04)
$Z_{ij\omega}$	Preference	Income	Income	Preference
D_{ij}	Preference	Income	Preference	Income
Adjusted R2	0.25	0.2	0.31	0.06
Root Mean Squared Error	0.22	0.23	0.22	0.25
Observations	88	88	88	88

Notes. The results of OLS regressions between survey income and model income under four different assumptions on idiosyncratic shocks and travel cost expressed in equation (C.1).

Estimating Parameters $\alpha_z, \alpha_d, \epsilon$ in a general case. The above framework also allows us to estimate $\alpha_z, \alpha_d, \epsilon$ using survey income data. These structural parameters illustrate the sources of spatial frictions in intra-city labor market, hence they are of independent interest aside from the prediction of income.

We estimate the parameters α_z, α_d and ϵ by OLS the equation:

$$y_{ij\omega}^{S} = \rho_1 \hat{X}_{ij}^{1} + \rho_2 \hat{X}_{j}^{2} + \rho_3 \hat{X}_{ij}^{3} + \varepsilon_{ij\omega}^{S}, \quad \text{(C.4)}$$

where $y_{ij\omega}^{S}$ is survey-based income of commuter ω who lives at i and works at j. Asymptotically, we have

$$\hat{\alpha}_z = \frac{\hat{\rho}_1}{\hat{\rho}_1 + \hat{\rho}_2},\ \hat{\alpha}_d = \frac{\hat{\rho}_3}{\hat{\rho}_1 + \hat{\rho}_2},\ \text{and}\ \hat{\epsilon} = \frac{1}{\hat{\rho}_1 + \hat{\rho}_2}. \quad \text{(C.5)}$$

Table C.2 reports the estimates of α_z, α_d, and ϵ based on estimating equation (C.4) with OLS, and using transformation (C.5). 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). In columns 1-2, we estimate the full equation (C.4), and we find that $\hat{\alpha}_d$ is close to zero with a small and insignificant negative value, and the other 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 increases the point estimate for $\hat{\alpha}_z$ and slightly lowers that for $\hat{\epsilon}$ while improving precision.

These results show that idiosyncratic shocks partly affect income, while travel time is most consistent with a pure utility cost.

Table C.2: How Pref. Shocks and Travel Time Affect Income: Estimated Structural Parameters

	(1)	(2)	(3)	(4)	(5)
	Full model			Constrained model ($\alpha_d = 0$)	
Shock productive α_z	0.21	-0.10	0.27	0.56	0.55
	(0.05)	[4.68]	[0.26]	(0.10)	[0.10]
Shock distance α_d	-0.57	-1.09	0.03	0	0
	(0.50)	[7.89]	[0.07]		
Shape parameter ϵ	12.84	16.97	11.85	9.09	9.11
	(7.59)	[60.25]	[3.80]	(1.16)	[1.36]
Observations	10,947	10,947	10,947	10,947	10,947
Bootstrap clusters		71	71		71

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 C. We estimated equation (C.4) by regressing individual log survey income from the DHUTS survey on the three model-predicted terms. In columns 4 and 5, 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 (C.5). Columns 1 and 4 report standard errors computed using the Delta method. Columns 2, 3, 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 coefficient is the median estimate and standard errors in square parentheses. Column 3 censors $\hat{\rho}_1 \geq 0$ and $\hat{\rho}_2 \geq 0$.

D Robustness: Gravity Equation Over-identification

In this section, we estimate the gravity equation on two disjoint samples to understand the stability of the estimated parameters. In each city, we estimate the gravity equation on the sample of nearby towers, and on the sample of distant towers. We use as cutoff the travel time such that aggregate commuting flows are roughly equal below and above the cutoff (13 minutes in Dhaka and 18 minutes in Colombo).

Table D.1 shows the results. Panel A shows that the distance coefficient is stable when estimating on these disjoint samples. (In Colombo, it is slightly steeper when estimated for long commutes.) Moreover, the resulting destination fixed effects estimated on the two disjoint samples are highly correlated (0.88 and 0.86 in the two cities).

Panel B repeats the validation exercise in Bangladesh. Column 1 repeats the analysis Panel A in Table 2, while the next two columns use model income computed using destination

fixed effects from the gravity equation estimated on the "close" and on the "far" samples, respectively. In both cases, the model income measure is predictive of log survey income, with a similar slope. The explanatory power is higher when using the "far" sample.

Table D.1: Overidentification: Estimating on "Close" and "Far" Tower Samples

	(1)	(2)	(3)	(4)	(5)	(6)
		Dhaka			Colombo	
Sample:	Pooled	Close	Far	Pooled	Close	Far
Panel A. Gravity Equation						
log Travel Time	-2.19***	-2.10***	-2.43***	-2.44***	-2.42***	-2.49***
	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)
$Corr(\hat{\psi}_j^{Close}, \hat{\psi}_j^{Far})$			0.88			0.86
Number of Destination FE	1201	1193	1199	1859	1741	1852
Number of Trips	9.4e+5	4.7e+5	4.7e+5	1.5e+6	7.3e+5	7.4e+5
Observations	1.3e+6	1.8e+5	1.2e+6	3.4e+6	1.6e+5	3.1e+6
Pseudo R2	0.66	0.80	0.49	0.67	0.80	0.44
Panel B. Validation (Outcome: log Survey Income, Workplace)						
$\epsilon\times$ log Model Income (workplace)	0.12***	0.09***	0.11***			
	(0.03)	(0.02)	(0.02)			
Observations	88	88	88			
Adjusted R2	0.25	0.15	0.27			

Notes. This table reports results when estimating the gravity equation using only nearby (or only distant) tower pairs. Panel A estimates gravity equation using home-work commuting flows (columns 1 and 4 reproduce results from Table 1). The sample is all tower pairs at least 180 seconds away. In columns 2 and 3, we restrict to towers below and above 13 minutes, respectively. In columns 5 and 6, we restrict to towers below and above 18 minutes, respectively. In columns 3 and 6 we report the correlation between the two vectors of destination fixed effects using the two disjoing samples. Panel B regresses log survey income at the workplace level on the log of our model income measure (at the workplace level). Column 1 reproduces Table 2, while the next two columns use destination fixed effects estimated using the two disjoint samples. Robust standard errors in parentheses.

E Supervised-Learning Method Details

In Sections 3 and 4, we compare the predictive power of a single model-predicted income measure, and of a supervised learning approach that uses multiple features derived from cell phone data. This appendix describes the details of the supervised-learning approach.

The main steps of our procedure as as follows. We begin by computing a large set of cell phone tower-level metrics from cell phone data. Following Blumenstock et al. (2015), we then use elastic net regularization (Zou and Hastie 2005) to fit a linear model without over-fitting the data. We then assess the predictive power on a hold-out testing sample. The rest of this section explains the details of feature construction, model fitting and hyper-parameter calibration, and of the comparison with the model-predicted income measure.

Extracting a Large Set of Quantitative Metrics from Cell-Phone Data

To construct our set of features from cell phone data, whenever the data allows we closely follow Steele et al. (2017), who use cell phone data to map poverty in Bangladesh. We then add additional hour-and-location level metrics.[3] To capture nonlinear patterns, for each variable described below, we include both the variable and its logarithm. Altogether, we have 498 tower-level features from this procedure.

[3] "Transactions" refers to outgoing call in Bangladesh, as only this type of call is recorded in the data.

User-level characteristics averaged at home and work locations. The first set of features measures averages at users' home and work towers. We construct the following statistics for each user for the entire sample period.

1. Number of transactions
2. Number of places: unique number of towers that the user ever visits
3. Radius of Gyration: the sum of squared distances from each visited tower (each transaction) to the centroid of all visited towers
4. Entropy of places: $-\sum_{i \in N_i} P_i \log P_i$, where P_i is the fraction of transactions at tower i, and N_i is the set of all towers visited by i

For each tower, we then take the average of these metrics, once for all users for whom this tower is their *home* location, and once for all users for whom this tower is their *work* location. Altogether, we obtain 8 metrics (4 metrics $\times$ 2 (home and work)).

Hourly statistics at the tower level. The second set of features is constructed for each hour of the day and each tower. We first compute the following statistics for each tower, date and hour:

1. Number of transactions
2. Number of unique users who made transactions
3. Average travel time distance to home locations of users who made at least one transaction at the tower on the specified date and hour
4. Average travel time distance to work locations of users who made at least one transaction at the tower on the specified date and hour
5. Average duration of calls

We then aggregate these statistics at the tower level, separately for weekends and weekdays (excluding *Hartal* days). Together, we have 240 (5 metrics $\times$ 24 hours $\times$ 2 (weekdays/weekends)) features.

Tower areas. The last statistic is the geographic area of the voronoi cell that contains the tower. We choose this statistic as a particularly compelling predictor of economic activity because cell phone operators tend to strategically locate towers at a high spatial frequency in areas where they expect high (cell phone) activity.

Our final set of cell phone features includes all the variables above, and for each one, its logarithm. In total, we have 498 features (2 $\times$ (8+240+1)).

Elastic Net Regularization for Relevant Feature Selection

Given the large number of features (or variables) relative to the number of observations, our next step is to use a supervised learning model that has good out-of-sample predictive power and does not overfit the training data set. Following Blumenstock et al. (2015), we use elastic net regularization, which is a regularized linear regression method that minimizes the sum of squared deviations from a linear model, minus a penalty term. The penalty term is the sum of an absolute value or L^1 penalty (as in LASSO regression) and a quadratic or L^2 penalty (as in ridge regression):

$$\lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|) \tag{E.1}$$

where β_j is the coefficient on feature j, and λ and α are hyperparameters.

We implement the elastic net regularization in the following steps. First, we randomly

select 50% of our survey areas as our "training data," and predict the survey income of the remaining survey areas as "test data." Second, we implement the elastic net regularization to select relevant features and fit the model. Third, we assess the predictive performance of the model in the test data. Our primary measure is test R^2, defined by the sum of squared prediction error divided by the total sum of squares. Lastly, we repeat this exercise 100 times, and report the average test R^2 (as well as the training R^2).

Our baseline results use $\alpha = 0.5$. We show in robustness exercises below that this parameter choice does not significantly affect our results. For λ, a typical strategy used in the literature is cross-validation. Due to the very small sample (88 observations), this does not perform well in our case. Instead, we select λ to maximize the R-squared in the test data over 100 random splits of the data into training and test. Given that we are using the *test* data for choosing λ, the predictive power we obtain is likely an upper bound of the true predictive power. Below, we show that choosing λ based on cross-validation within the training data set performs worse (for survey workplace income prediction).

Additional Robustness Results with DHUTS Survey Workplace Income

Hyperparameter λ using cross-validation. Here we replicate Table 2 panel (B) where the elastic net hyperparameter λ is computed via cross-validation. For each iteration of splitting the training and test data set, we further split the training data set into N folds. Within these N set of samples, we repeat training the data with $N-1$ subsets and predict the in remaining subset. We repeat this procedure N times, and compute the sum of squared prediction residuals. We choose λ that minimizes the prediction error, and we use the chosen λ to once again train the model with the entire training data set, and evaluate the predictive performance using the test data set.

Table E.1 reports the results. Column (1) is the OLS prediction with the model-predicted income, and Columns (2)-(7) are the results of the elastic net using all cell phone data features. Column (2) simply reproduces Panel (B) of Table 2 where λ is chosen to maximize the test R^2. Columns (3)-(7) show the results when we choose λ based on different number of folds for cross-validation.

Table E.1: Predicting Workplace Income: Choosing Hyperparameter with Cross-Validation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	OLS	Elastic Net					
	(log Model Income)	(All CDR Features)					
		Maximize Test R^2	CV	CV	CV	CV	CV
Training R^2	0.26	0.44	0.44	0.48	0.50	0.51	0.53
Test R^2	0.22	0.24	0.19	0.18	0.13	0.16	0.12
Number of Folds for CV			3	5	10	20	44
Observations	88	88	88	88	88	88	88

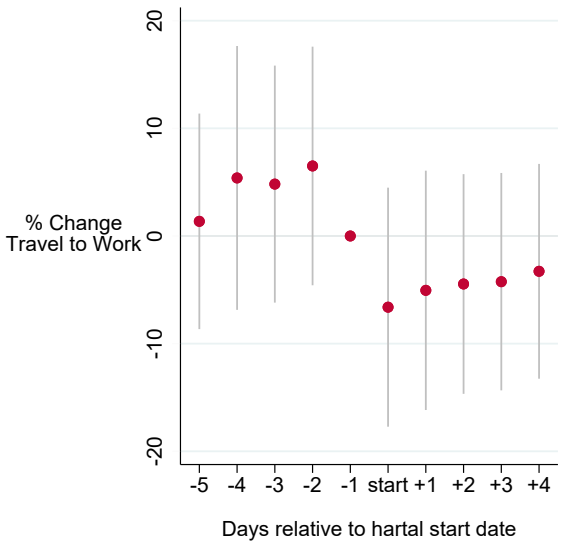
Columns (3)-(7) show that the test R^2 falls when we use the cross-validation procedure for choosing λ. In fact, test R^2 is lower than the OLS with model-predicted income. At the same time, training R^2 is higher than in columns (1) and (2), suggesting that poorer predictive performance is likely due to overfitting. Overfitting is unavoidable given the small sample

size.[4]

Hyperparameter α robustness. $\alpha = 1$ assigns all weight to the L^2 norm, which is equivalent to the ridge regression. $\alpha = 0$ assigns all weight to the L^1 norm, which is equivalent to LASSO. Test R^2 for $\alpha = 0, 0.25, 0.5, 0.75$ and 1 is 0.17, 0.24, 0.24, 0.24 and 0.23, respectively.

F Additional Results: the Impact of Hartal

Figure F.1: Impact of Hartal on Commuting to Work



Notes. This figure shows the event study impact of the onset of a hartal event on the probability to commute to work. The sample is based on all commuters whose long-term home and workplace towers are different (35% of all users), who travel at least once on hartal days, and once on non-hartal days. The sample is all days with commuting data (including stationary trips). "Trip to Work" is a dummy for making a proper trip (origin distinct from destination) to the long-term workplace location (defined based on non-Hartal days).

The event study in Figure F.1 shows that there is a fall in commuting to work at the onset of hartal strikes. The point estimates are consistent with anticipation and a partial reduction in commuting to work on the day before the onset of hartal.

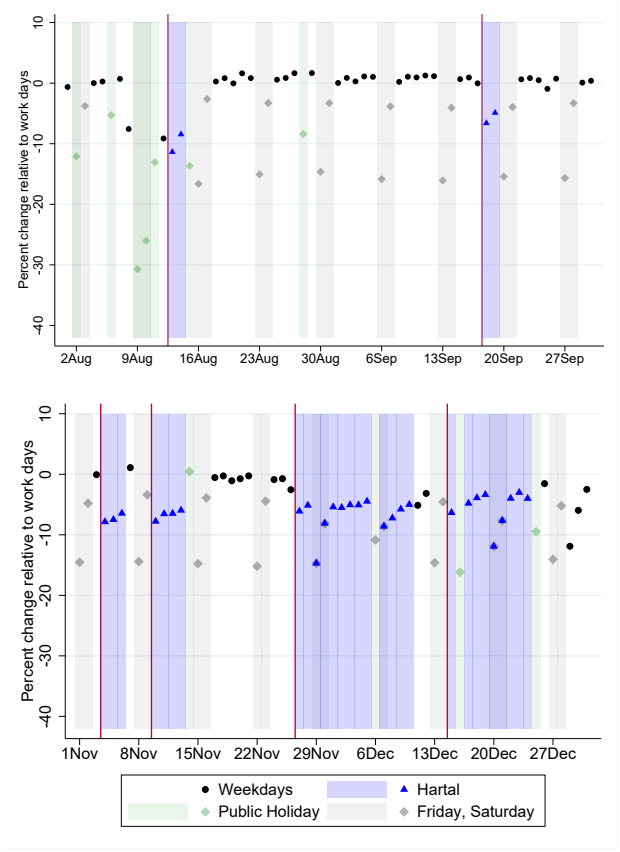
To construct this figure, we proceed as follows. First, we compute calendar date fixed effects using the regression $C_{\omega t} = \psi_t + \mu_\omega + \epsilon_{\omega t}$ where ω denotes a commuter, t denotes a calendar date, and $C_{\omega t}$ is a dummy for commuting to work. (Figure F.2 plots these fixed effects, normalized as percentage changes relative to the mean of the outcome variable on non-hartal, non-holiday workdays.) Next, we adjust the date fixed effects by the average differences on Friday (the main free day in Bangladesh) and Saturday (the other weekend day). We exclude holidays from the sample, as well as the 5 days in the sample that are both hartal and weekend. Lastly, we construct hartal "onset" events. We require at least two days between hartal events, which leads to a sample of six hartal onset events (see the thin vertical red lines in Figure F.2). We use an unbalanced panel pooling the six hartal events. For each event, we include up to 5 days prior to the first hartal day, excluding holidays. If another hartal takes place in this

[4]Indeed, for the residential asset prediction (where the sample size is over 1,000) the cross-validation and choosing λ to maximize the test R^2 perform similarly (not reported).

preceding period, we exclude it and all previous days. We include all consecutive hartal days after it starts.

Table F.1 replicates Table 4 using a sample of frequent callers. The patterns of results is very similar, thus alleviating the concern that commuting reductions during hartal may be driven by commuters making fewer calls rather than commuting to work less.

Figure F.2: Commuting by Calendar Date (Hartals, Holidays and Weekends)



Notes. This figure shows average commuting probability by calendar date. The Y axis plots the percentage change relative to the mean on non-hartal, non-holiday workdays. The sample and outcome are as in Figure F.1. The figure plots calendar date fixed effects from a regression of any trip commuting dummy 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/. The red vertical lines indicate hartal event onset date for the six hartal events. Friday is the main free day in Bangladesh, and Saturday is the other weekend day. Five days in the sample are both hartal and weekend: August 13, September 18, November 4, 10, and 27, and December 15. We drop these throughout the analysis.

Table F.1: The Heterogeneous Impacts of Hartal on Commuting: Frequent Commuter Sample

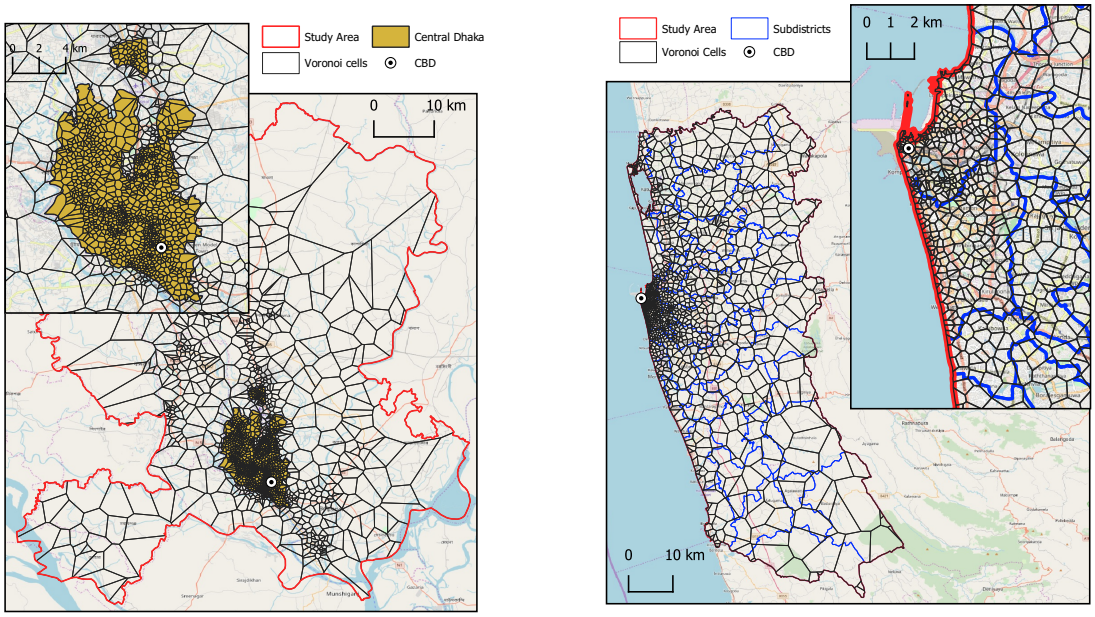
	Work Commute (% change vs weekday)			
	(1)	(2)	(3)	(4)
Hartal	-0.053***	-0.053***	-0.053***	
	(0.004)	(0.004)	(0.004)	
Interactions: Hartal ×				
(β^L) % Low Skill				-0.025**
				(0.012)
(β^H) % High Skill				-0.051***
				(0.004)
Dest. FE (z)		-0.022***	-0.019***	
		(0.005)	(0.005)	
(β^L_W) % Low Skill × Dest. FE Low Skill (z)				-0.049***
				(0.012)
(β^H_W) % High Skill × Dest. FE High Skill (z)				-0.008
				(0.006)
Log Duration (z)			-0.028***	
			(0.002)	
(β^L_D) % Low Skill × Log Duration (z)				-0.025***
				(0.008)
(β^H_D) % High Skill × Log Duration (z)				-0.029***
				(0.003)
Commuter FE	X	X	X	X
P-value $\beta^L = \beta^H$				0.05
P-value $\beta^L_W = \beta^H_W$				0.00
P-value $\beta^L_D = \beta^H_D$				0.70
Observations	3.9e+06	3.9e+06	3.9e+06	3.9e+06

Notes. This table replicates Table 4 on the sample of frequent callers, defined as those who have commuting data on at least half of all days (61 out of 122 days), who account 8.3% of all commuters.

G Additional Figures and Tables

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

(A) Dhaka (B) Colombo



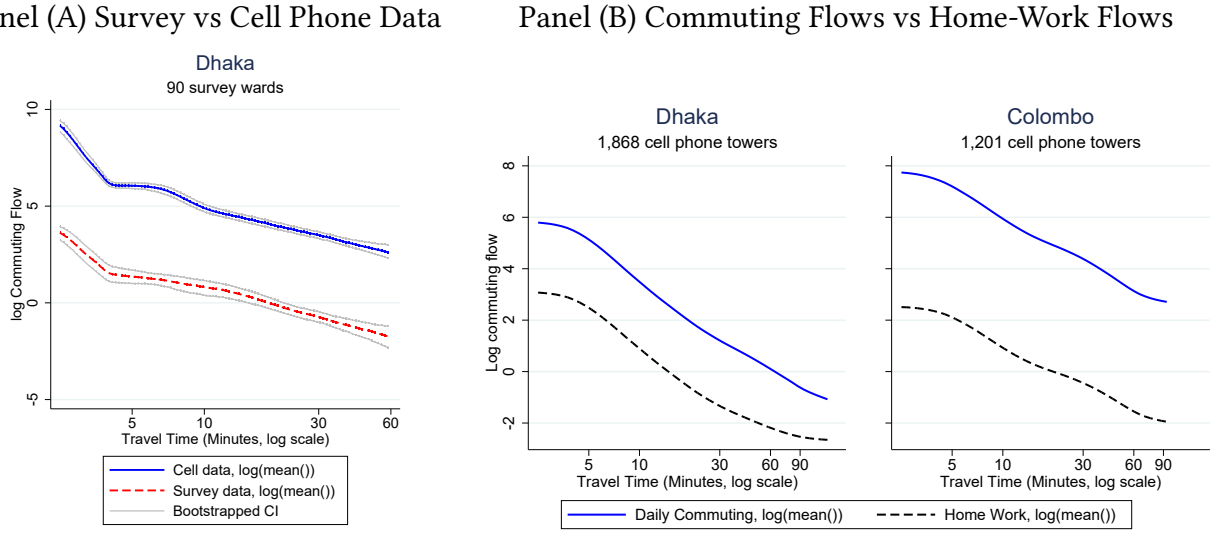
Notes. This figure shows the map of cell phone tower Voronoi cells in Dhaka, Bangladesh (Panel A), and in Colombo, Sri Lanka (Panel B). The yellow shaded area is the Dhaka City Corporation (DCC), the urban core of Dhaka, the main sample in the DHUTS transportation survey. The overall study area covers for Dhaka are three districts in Bangladesh: Dhaka, Gazipur, and Narayanganj, and 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.

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

		Dhaka, Bangladesh	Colombo, Sri Lanka
Panel A. Home-Work Commuting Flows			
(1)	Unique users	5.1e+06	3.0e+06
(2)	Users with home and work towers	4.9e+06	2.6e+06
(3)	Users (distinct home and work towers)	1.6e+06	9.9e+05
(4)	Users (gravity equation sample)	1.5e+06	9.4e+05
Panel B. Daily Commuting Flows			
(5)	Unique users	3.6e+06	3.0e+06
(6)	Weekdays in sample	87	282
(7)	All user-days possible (= (5) × (6))	3.1e+08	8.4e+08
(8)	User-days with data (daily trips)	3.8e+07	2.4e+08
(9)	Coverage rate (= (8) / (7))	12.4%	28.1%
(10)	Trips (distinct origin and destination towers)	2.1e+07	1.4e+08
(11)	Trips (gravity equation sample)	1.9e+07	1.3e+08

Notes: This table describes data coverage in the two countries. Panel A reports the number of commuters based on our home-work classification. Row 1 indicates the number of commuters with at least one home tower (based on calls between 9pm and 5am) or at least one work tower (based on calls between 10am and 3pm). Row 2 indicates the number of commuters with both home and work towers. Row 3 restricts to distinct towers, and row 4 to our baseline gravity equation estimation sample, towers more than 180 seconds away and closer than the 99th percentile of the duration distribution. Panel B reports information about daily commuting trips. A daily trip is a pair of origin and destination towers visited by the same user during a single day, in the intervals 5am-10am and 10am-3pm, respectively. Row 5 indicates the number of unique users who have at least one trip on a weekday. (We do not have this number for Sri Lanka so we use the number of users from row 1.) Row 6 is the number of calendar weekdays in the data. Row 7 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.) Row 8 describes the actual number of daily trips. Row 9 reports coverage for daily trips. Rows 10 and 11 replicate rows 3 and 4 for daily trips.

Figure G.2: Commuting Flows from Survey Data and Cell Phone Data



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 1 for the definition of home-work and daily 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 G.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.

Table G.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	0.63***	0.70***	0.30***	0.53***
	(0.020)	(0.026)	(0.059)	(0.049)
Log duration			-1.05***	-0.51***
			(0.17)	(0.11)
Origin and destination fixed effects		Yes		Yes
Observations	6026	6026	6026	6026

Notes: This table shows the relationship between commuting flows from two different data sets in Dhaka: the DHUTS transportation survey (outcome) and home-work comuting flows from cell phone data (explanatory variable). The survey sample consists of the 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. (The sample includes government workers.) 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. Standard errors are clustered at the origin survey area level. $^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

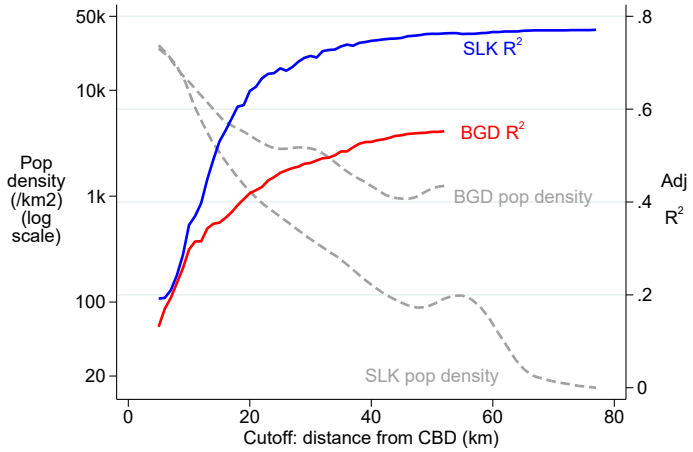


Figure G.3: Distance to CBD and R^2

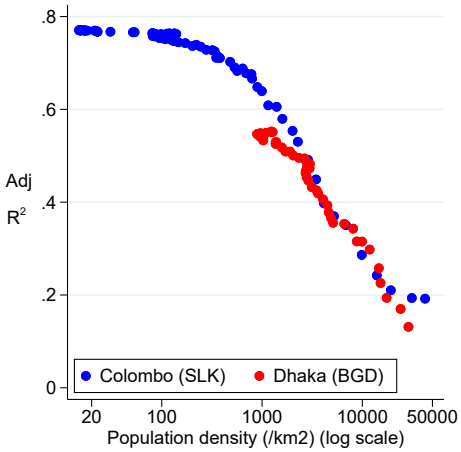
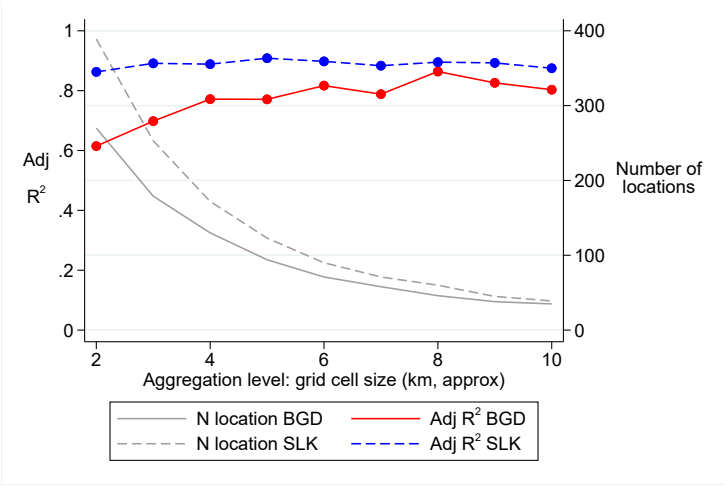


Figure G.4: Population Density and R^2

Notes. This figure shows how the R^2 of census residential income proxy depends on the sample of locations included in the analysis. In Figure G.3, we restrict the sample to cell phone towers within a certain distance to the CBD. The graph plots the adjusted R^2 when regressing census income proxy on model residential income. The figure also plots the local linear regression of population density (per square km) for towers at a given distance to the CBD. Figure G.4 shows the relation between population density at the distance cutoff and the achieved adjusted R^2. Interestingly, the relationship between the population density at the distance cutoff and adjusted R^2 is similar in the two cities, suggesting a more general relationship.

Figure G.5: Prediction R^2 and Geographic Aggregation Level



Notes. This graph shows the R^2 of the regression of census income proxy on model residential income, at different levels of aggregation. For $k = 2,\ldots,10$ we aggregate cell phone towers into grid cells of size $k \cdot 0.01$ in decimal coordinates (equal to approximately $k \cdot 1.11$ kilometers). We aggregate commuting flows and run the gravity equation at this level, and recover average residential income. (Note, we do not area adjust the destination fixed effects as grid cells have approximately equal area.) The gray lines (right Y axis) indicate the number of grid cells in the aggregated data.

Table G.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 representativeness of the cell phone data at the residential level. 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. Census residential population in a Voronoi cell is computed as the average census population in the finest available census geographic units , weighted by their area 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 comparatively lower adjusted R-squared in columns (3) and (4) may be due to the fact that 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 G.4: Gravity Equation Robustness: Destination Fixed Effects

	Destination Fixed Effects (Benchmark)										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Dest FE (Daily Flows)	0.98***					1.09***					
	(0.01)					(0.01)					
Dest FE (Full Sample)		0.95***					1.03***				
		(0.01)					(0.01)				
Dest FE (OLS with log(volume))			3.58***					3.20***			
			(0.04)					(0.04)			
Dest FE (OLS with log(volume + 1))				7.06***					5.32***		
				(0.11)					(0.12)		
Dest FE (Nonparameric Gravity Equation)					0.98***					0.98***	
					(0.003)					(0.003)	
Dest FE (Travel Time with Congestion)											0.98***
											(0.003)
Estimation Method	PPML	PPML	OLS	OLS	PPML	PPML	PPML	OLS	OLS	PPML	PPML
City	Dhaka	Dhaka	Dhaka	Dhaka	Dhaka	Colombo	Colombo	Colombo	Colombo	Colombo	Colombo
Observations	1,859	1,859	1,859	1,859	1,859	1,201	1,201	1,201	1,201	1,859	1,201
Adjusted R^2	0.92	0.88	0.81	0.68	0.98	0.92	0.87	0.82	0.62	0.98	0.99

Notes. This table compares destination fixed effects computed under different assumptions. The outcome in the first four (last five) columns is the destination fixed effects from the first (third) column in Table 1. Each row uses destination fixed effects (FE) from the gravity equation estimated differently. The (destination FE estimated in the) first row uses daily commuting flows (columns 2 and 4 in Table 1). The second row uses all tower pairs below the 99th percentile of the travel time including same-tower pairs (which account for over half of all commuting flows), with travel time censored from below at 180 seconds. The third row estimates the gravity equation by OLS dropping all tower pairs with zero commuting flows (to allow for logarithms). The fourth row estimates the gravity equation by OLS using log commuting flow plus one as outcome. The fifth row estimates the gravity equation with log travel time entering non-parametrically instead of linearly, as dummies for the deciles of log travel time. The last row uses the travel time from Google Maps query with traffic congestion taken into account. (The query for Sri Lanka was sent for 8am on Friday, August 26, 2016, one month prior to this date.) Most coefficients are close to 1 and the R^2 is above 0.8, except for the third and fourth rows. High regression coefficients of the third and fourth rows indicate that the destination effects are flatter if we estimate the gravity equation by OLS ignoring zero flows, due to sample selection. Standard errors in parentheses. $^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table G.5: Average Workplace Income: Model Prediction and Survey Data in Dhaka

	log Survey Income (workplace)				
	(1)	(2)	(3)	(4)	(5)
$\epsilon\times$ log Model Income (workplace)	0.12^{***}			0.11^{***}	0.17^{*}
	(0.03)			(0.03)	(0.09)
log Employment Density		0.11^{**}		-0.07	-0.06
		(0.06)		(0.05)	(0.05)
log Dist. to CBD			-0.18^{***}	-0.14^{***}	-0.15^{***}
			(0.03)	(0.02)	(0.03)
$\epsilon\times$log Model Income (residential)					-0.12
					(0.15)
Adjusted R2	0.25	0.06	0.33	0.42	0.42
Root Mean Squared Error	0.22	0.24	0.22	0.21	0.21
Observations	88	88	88	88	88

Notes. Robustness of Table 2 controlling for employment density, distance to CBD, and the model residential income.

Table G.6: Robustness: Average Workplace Income and Survey Income Comparison

	log Survey Income (workplace)							
	(1) Daily Flows		(2) Excluding Neighboring Towers		(3) Without Area Adjustment		(4) Include All Origins	
Panel A. Log Survey Income								
log Model Income (workplace)	0.13^{***}	0.24^{***}	0.10^{***}	0.08^{**}	0.21^{***}	0.08	0.11^{***}	0.18^{**}
	(0.03)	(0.06)	(0.02)	(0.03)	(0.05)	(0.08)	(0.03)	(0.08)
Geographic Controls		X		X		X		X
Adjusted R2	0.26	0.44	0.2	0.41	0.25	0.41	0.21	0.45
Observations	88	88	88	88	88	88	89	89
Panel B. Log Survey Income Residual on Demographics								
log Model Income (workplace)	0.07^{***}	0.13^{***}	0.05^{***}	0.05^{**}	0.11^{***}	0.03	0.06^{***}	0.08
	(0.02)	(0.04)	(0.01)	(0.02)	(0.02)	(0.05)	(0.01)	(0.05)
Geographic Controls		X		X		X		X
Adjusted R2	0.21	0.28	0.16	0.26	0.18	0.25	0.2	0.27
Observations	88	88	88	88	88	88	89	89

Notes. Robustness for Table 2 and Table G.5. Odd and even columns correspond to the specifications in columns 1 and 5 of Table G.5. The first two columns use commuting flows defined at the daily level instead of commuting flows from home and work assignment (see Section 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. (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.)

Table G.7: Individual Income: Model Predictions and Survey Data

	log Survey Income			
	(1)	(2)	(3)	(4)
Model log Income (workplace)	0.11^{***}	0.04^{***}	0.03^{***}	0.02^{**}
	(0.02)	(0.01)	(0.01)	(0.01)
log Travel Time		0.12^{***}	0.13^{***}	0.07^{***}
		(0.02)	(0.01)	(0.01)
log Dest. Dist. to CBD		-0.05^{***}	-0.05^{***}	0.01
		(0.01)	(0.02)	(0.02)
log Dest. Commuting Zone Area		-0.04^{***}	-0.06^{***}	-0.07^{***}
		(0.02)	(0.02)	(0.02)
Male				0.46^{***}
				(0.02)
Age				0.01^{***}
				(0.001)
Level of education				0.17^{***}
				(0.01)
Origin FE		X	X	X
Occupation and Sector FE				X
Government Worker	No	No	Yes	Yes
Observations	10,948	10,948	12,348	12,347
Adjusted R^2	0.02	0.03	0.03	0.28

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, 3 and 4 include origin survey area fixed effects, and column 4 includes occupation and job sector fixed effects. Conley standard errors with 5 km distance cutoff in parentheses. (For computational purposes, when including fixed effects, the standard errors are computed after residualizing the fixed effects.) $^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$

Table G.8: Average Residential Income: Model Prediction and Residential Income Proxy

	Census Residential Income Proxy			
	(1)	(2)	(3)	(4)
Panel A. Dhaka				
log Model Income (residential)	0.89^{***}			0.64^{***}
	(0.06)			(0.23)
log Residential Density		0.67^{***}		0.37^{***}
		(0.02)		(0.06)
log Dist. to CBD			-0.84^{***}	-0.02
			(0.10)	(0.11)
log Model Income (workplace)				-0.35^{***}
				(0.13)
Sub-district FE (count)				X (55)
Adjusted R2	0.54	0.63	0.33	0.74
Observations	1,844	1,844	1,844	1,844
Panel B. Colombo				
log Model Income (residential)	1.29^{***}			1.38^{***}
	(0.06)			(0.19)
log Residential Density		1.23^{***}		0.20^{***}
		(0.07)		(0.07)
log Dist. to CBD			-2.04^{***}	-0.57^{**}
			(0.22)	(0.27)
log Model Income (workplace)				-0.72^{***}
				(0.12)
Sub-district FE (count)				X (41)
Adjusted R2	0.77	0.67	0.7	0.92
Observations	1,193	1,193	1,193	1,193

Notes. Robustness of Table 3 controlling for residential density, distance to CBD, and the model workplace income.

Table G.9: Robustness: Average Residential Income and Census Income Proxy

	Census Residential Income Proxy					
	(1) Daily Flows		(2) Excluding Neighboring Towers		(3) No Area Adjustment	
Panel A. Dhaka						
log Model Income (residential)	1.08^{***}	0.37^{***}	0.93^{***}	0.82^{***}	-1.52^{***}	-0.82^{***}
	(0.08)	(0.12)	(0.06)	(0.17)	(0.11)	(0.13)
Geographic Controls		X		X		X
Sub-district FE (count)		X (55)		X (55)		X (55)
Adjusted R2	0.47	0.7	0.56	0.74	0.42	0.74
Observations	1,821	1,821	1,866	1,866	1,866	1,866
Panel B. Colombo						
log Model Income (residential)	1.69^{***}	0.68^{***}	1.48^{***}	1.00^{***}	-1.52^{***}	-0.62^{***}
	(0.08)	(0.14)	(0.08)	(0.33)	(0.31)	(0.16)
Geographic Controls		X		X		X
Sub-district FE (count)		X (41)		X (41)		X (41)
Adjusted R2	0.82	0.91	0.82	0.91	0.08	0.91
Observations	1,188	1,188	1,197	1,197	1,197	1,197

Notes. Robustness for panel (A) in Tables G.8. Odd and even columns correspond to the specifications in columns 1 and 4 in Tables G.8. The first two columns use daily commuting flows instead of home-work commuting flows (see Section 1 for definitions). 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 last two columns use destination fixed effects not adjusted for Voronoi cell tower area.