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

Gabriel E. Kreindler[†] Yuhei Miyauchi[‡]

April 8, 2020

JEL Codes: C55, E24, R14

Abstract

We show how commuting flows can be used to infer the spatial distribution of income within a city. We use a simple workplace choice model, which 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. Unlike machine learning approaches, our method does not require training data, yet achieves comparable predictive power. In an application, we show that hartals (transportation strikes) in Dhaka lower commuting, leading to 5-8% lower predicted income.

[*] 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-benot 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.

[†] Harvard University. Email: gkreindler@fas.harvard.edu

[‡] Boston University. Email: miyauchi@bu.edu

1 Introduction

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. For example, less than 10% of the urban population in sub-saharan African countries is covered by a census of firms with wage data.[1] At the same time, comprehensive new data sources on urban behavior, especially individual mobility and commuting, are becoming available across the world.

In this paper, we provide a theory-based method to predict the spatial distribution of urban economic activity from commuting choices. 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 average wage at a location based on how "attractive" it is as a commuting destination. We use tools from recent urban economics models to formalize this intuition. In the model, work location decisions aggregate up to a gravity equation on commuting flows, and destination fixed effects are proportional to log wages. This property holds for a general class of models developed to evaluate urban policies and transport infrastructure (Redding and Turner 2015, Redding and Rossi-Hansberg 2017).[2]

Our approach differs from previous studies that use machine learning techniques to empirically predict wealth and consumption at the individual and geographic area level (Blumenstock et al. 2015, Jean et al. 2016, Glaeser et al. 2017). First, our primary focus is the prediction of income *within a city*, which is more challenging than at a wider spatial scale. Furthermore, the distribution of income changes during the day, as income "moves" around the city due to commuting. The model and data we use are explicit about this link. Second, our approach is grounded in a simple and general theory of behavior; as such, it may be more transferable across settings compared to data-driven prediction methods. Third, we will show that despite not using any training data,

[1] Authors' calculation (Appendix A.1).

[2] Other contributions include Ahlfeldt et al. 2015, Heblich et al. 2018, Tsivanidis 2018, Severen 2019.

our approach performs comparably to machine learning techniques, explaining 70-90% as much income variation within cities as those methods.

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 (phone call or text message). We construct individual home and work locations by observing a user's location at different times of the day over time. We show that commuting flows constructed this way correlate strongly with commuting flows from a transportation survey from Dhaka, while additionally offering very fine geographic resolution. We use this data to estimate the gravity equation implied by the model. We return to the high-frequency temporal aspect of the data in an application below, where we use the cell phone data to construct daily individual commuting trips.

Estimated log wages by location are derived solely from observed commuting decisions and data on travel times, without any model training with actual wage data. We next assess how well this simple measure captures real differences in wages, using two income proxy data sources.

First, model *workplace* income is significantly positively correlated with *workplace* commuter income data from a large transportation survey in Dhaka. This also holds after controlling for employment density and distance to the central business district (CBD), and when repeating the exercise after projecting out demographic and occupation covariates from the survey. Second, in both cities, model-predicted *residential* income is a robust predictor of the census *residential* income proxy. This relationship remains stable within sub-districts and after controlling for residential density and distance to the CBD.

A key advantage of the model is that we can compute how income "moves" around the city. As a further check of the model fit, we perform a horse-race between residential- and workplace-income. While the two measures are highly correlated, we find suggestive evidence that model *workplace* income better correlates with *workplace* survey income data, and model *residential* income better correlates with the *residential* income proxy.

In each validation exercise, the model measure, computed without any training data, explains between 70% and 90% as much income variation compared to a trained machine learning method that uses a rich set of cell phone features (Zou and Hastie 2005, Blumenstock et al. 2015). Hence, the simple revealed preference logic in the model captures a significant part of all the information contained in the cell phone data, and the destination fixed effects function as a near "summary statistic."

The overall explanatory power (R^2) varies for the workplace and residential exercises described above. Prediction is significantly more difficult (using any explanatory variable) when focusing only on areas within the urban core of Dhaka ($R^2 \approx 0.25$), compared to an exercise that covers a wider geographical region with both urban and peri-urban areas ($R^2 \approx 0.5$). Overall, comparisons of predictive power are more informative within a given setting rather than across applications. We further discuss the comparison with previous work using model-based or machine learning-based approaches in section 4.1.

The ideal application of our income-prediction method and of the high-frequency commuting data is to trace out the spatial and temporal impact of urban events and policies. This may be particularly useful in the case of acute events when there is no time for conventional data collection, such as lock-downs due to quarantines for pandemics or terrorist attacks, floods, etc. To illustrate this potential, we estimate a measure of the economic costs of *hartals*, a type of strike intended to disrupt transportation and economic activity in Bangladesh. We use daily commuting flows constructed by observing each user's location at different times during a single day. The onset of hartal lowers commuting by 5-8%. Assuming wages throughout the city are unchanged on hartal days, our accounting exercise implies a decrease in take-home income of around the same magnitude. While precisely estimated, these changes are relatively small, in line with previous studies of hartal.

2 Cell-Phone Data and Commuting Flows

2.1 Data Sources

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

We construct commuting trips by 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, to incorporate the possibility that some users do not have fixed work locations.[4] 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 B.4).

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, and 1,716 households outside the DCC. Home and work locations are at the level of 108 "survey areas." Our main analysis sample covers 12,510 commuters who live and work within the 90 survey areas inside the DCC, with positive income from work, excluding students, homemakers, and the unemployed. In the main analysis, we exclude households outside of DCC, because the 18 corresponding survey areas are significantly coarser and detailed information on sampling is not available. Results are robust to including commuters who live outside DCC (Appendix Table B.5).

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,

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

[4] 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 B.1).

water and electricity connection) at the finest geographic unit available.[5] The residential income proxy at the cell tower level is the average across overlapping census units, weighted by overlap area with the tower.

2.2 Representativeness of Commuters in Cell Phone Data

Here we explore to what extent cell phone data is representative of urban commuters. 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 (Appendix Table B.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 (Appendix Figure B.2, Panel A).

Residential population density from cell phone data correlates well with census population density at the level of 1,866 and 1,201 cell phone towers in the two cities (Appendix Table B.3), with $R^2 = 0.61$ in Dhaka and $R^2 = 0.49$ in Colombo. 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 we show in the next section, our approach infers wages based on how commuters in a given location choose between different workplace locations. However, this method may still be biased if, for example, workers in high-density locations choose their work locations based on different criteria compared to other workers (e.g., they place different weight on commuting distances and wages).

3 Model: Commuting Flows, Gravity, and Wages

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

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

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

3.1 Workplace Choice Model

Space is partitioned into a finite set of locations, which may serve as both residential and work locations. In our application, locations correspond to Voronoi cells around cell phone towers (Appendix Figure B.1).

There is a unit mass of workers, and each worker ω sequentially decides where to live, and then where to work. We do not impose restrictions on the home location choice.[6] Given her residential location (or origin) i, the worker chooses her work location (or destination) j. The utility of worker ω residing in location i if she chooses destination j is:

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

W_j is the wage per effective unit of labor supply at location j (all firms at location j offer the same wage), D_{ij} is the travel time between i and j, and $Z_{ij\omega}$ is an idiosyncratic utility shock that is i.i.d. following the Fréchet distribution, with scale parameter T and shape parameter ϵ. We assume that each worker supplies one unit of labor, and hence earns income W_j if she works in location j. We abstract from heterogeneity due to skill or other worker attributes.[7]

Each worker observes the shocks $Z_{ij\omega}$ and chooses the work location j where $U_{ij\omega}$ is maximized. The probability that a worker commutes to j conditional on residing in i is given by

[6] Assuming joint home and work location choice leads to the same gravity equation (Ahlfeldt et al. 2015). However, if workers choose their workplace first and then the home location (as perhaps in the case of new migrants), we would obtain a different gravity equation.

[7] We model and investigate empirically two extensions where labor supply varies across individuals. First, in Appendix A.2, labor supply (and hence income) depends on observable demographics. Second, in Appendix A.3, $Z_{ij\omega}$ and D_{ij} partly affect labor supply, rather than only affecting utility, as in the main analysis. We develop a method to estimate how much $Z_{ij\omega}$ and D_{ij} affect income using survey income data. The results are consistent with D_{ij} being a pure utility shock, and $Z_{ij\omega}$ partly affecting income (Appendix Table A.1).

$\pi_{ij} = (W_j / D_{ij}^{\tau})^{\epsilon} / \sum_s (W_s / D_{is}^{\tau})^{\epsilon}$. Taking logs, and denoting log quantities by lowercase letters:

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

3.2 Estimating the Gravity Equation

We estimate equation (2) using the empirical Poisson pseudo-maximum likelihood (PPML) method with two-way fixed effects:

$$\log(E[\pi_{ij}]) = \psi_j - \beta d_{ij} + \mu_i \tag{3}$$

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).[8,9]

Importantly, 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.[10]

To obtain relative wage *levels*, we further require knowing ϵ, the Fréchet parameter, which in the model governs the variance of idiosyncratic preferences shocks. This is identified, for example, from the overall variance of wages in the city (Ahlfeldt et al. 2015, see Section 3.4).

Lacking detailed bilateral commuting flow data, some authors estimate log wages with an exactly identified procedure only using residential and employment populations and separately calibrated parameter β (Ahlfeldt et al. 2015, Tsivanidis 2018). Our approach using commuting flows is more robust against noise in a particular subset of tower pairs in the gravity equation (3). In fact, we explore robustness to including or excluding flows between nearby tower pairs and within-tower flows.

[8] Fally (2015) shows that the Poisson regression estimator asymptotically satisfies the structural relationship between ψ_j and μ_i in equation (2).

[9] Log travel time as a measure of distance offers a good fit (Appendix Figure B.2).

[10] 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 4 address empirically the extent to which our measure is correlated with wages.

3.3 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, the choice of spatial scope of location may not be innocuous for inferring wages. Larger Voronoi cells may mechanically yield larger destination fixed effects.

When the true model has independent shocks at sub-locations within a given location, commuting flows defined at the larger location level still follow gravity equation (2) approximately using an "effective" wage at that location. Assume that location j is divided into N_j smaller areas with independent shocks, and all areas have the same "true" wage W_j^R. By standard Fréchet properties, the commuting probability to j is approximately equivalent to a model with a single shock at j and "effective" wage $W_j = N_j^{1/\epsilon} W_j^R$.[11] (The approximation comes from assuming that the smaller areas are located at exactly the same location.) From equation (3) we estimate $\psi_j = \epsilon \log W_j$ and we recover the true wage as the *area-adjusted* destination fixed effect

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

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

3.4 Estimation Results: Gravity and Wages

We estimate gravity equation (3) using cell phone commuting flows and Google Maps travel times. Our goal is to recover the destination fixed effects, which in the model are proportional to workplace log wages. The estimation sample is non-holiday weekday commuting trips between pairs of towers excluding nearby and very distant towers.[12]

Table 1 reports the results, based on commuting flows between 1,859 locations in Dhaka (columns 1-2) and between 1,201 locations in Colombo (columns 3-4). The gravity equation is estimated with commuting flows constructed from assigned home and work locations for 1.5 and

[11] See Appendix A.4. Redding and Weinstein (2019) prove a related result for gravity models in trade.

[12] 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 (Appendix Table B.4). Towers over the 99th percentile of the travel time distribution are also excluded (137 and 96 minutes in Dhaka and Colombo, respectively).

1 million commuters in the two cities (columns 1 and 3), and using the commuting flows identified at the daily level (columns 2 and 4), which number 20 and 130 million in the two cities.

Commuting probability decreases strongly with travel time. Interestingly, although the average commuting trip is 25% longer on average in Sri Lanka, once we adjust for residential locations, the coefficients become similar, -2.44 and -2.19. This is a substantive finding, as the two cities differ in terms of economic development, population, and urban structure (mono- vs poly-centric).

Figure 1 displays smoothed estimated wages in Dhaka and Colombo using choropleth maps. Our estimated measure ψ_j^R is proportional to log wages, with factor ϵ (the Fréchet shape parameter). If we know ϵ (e.g., from other data or estimate using ground-truth wage data), we can also recover wage levels up to a multiplicative constant. Ahlfeldt et al. (2015) estimate $\epsilon = 6.83$ in Berlin, while in Dhaka we find $\epsilon = 9.09$ (Appendix A.3). In Figure 1 we use $\epsilon = 9.09$. 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, 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 (Appendix Table B.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).

4 Validation using Survey Income and Census Residential Income Proxy

The method above infers wages based on observed commuting choices. This section investigates to what extent this approach predicts within-city income patterns, using independent data on workplace and residential income.

4.1 Model-Predicted and Survey Workplace Income in Dhaka

Our first validation exercise compares income from the model and survey income from the DHUTS survey (Section 2.1). We compute average income at the workplace level in each survey area in the DCC, the finest geographic location available in the DHUTS survey.[13]

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

In our main exercise, we investigate whether model-predicted income predicts survey income at the workplace level. We benchmark the statistical significance and the predictive power in two ways. First, we compare with other simpler measures: employment density (from cell phone data) and distance to the Central Business District (CBD), established indicators of spatial economic activity within cities (Duranton and Puga 2015). Second, we compare the predictive performance with that from a supervised-learning approach (elastic net regularization) using hundreds of features from cell-phone data, trained using survey income data on a subset of locations.

Table 2 presents the main results. Column 1 in panel A shows that model-predicted income explains 25 percent of the variation in average income at the survey area level, and the coefficient implies a Fréchet shape parameter of $\hat{\epsilon} = 8.3$, similar to estimates in the urban economics literature (6.83 in Ahlfeldt et al. 2015). In columns 2-3, employment density and distance have slightly lower and slightly higher predictive power, respectively.[14] The coefficient on model-predicted income is almost unchanged when controlling for these variables (column 4), showing that the model contains information not available in these other measures. In column 5, we include model-predicted *residential* income. While the two model measures are highly correlated, the positive correlation with survey *workplace* income is loaded onto model *workplace* income. The coefficient on residential income is negative and less precisely estimated.

Panel (A) shows that our model-predicted income is a statistically significant predictor of survey income. This is a substantial result, given that our model-predicted income only requires commuting matrix data extracted from cell-phone data. In some settings, wage data may be available at some point in time or for a subset of locations within a city. In such cases, a supervised-learning

[13] 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 (Appendix Table B.7).

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

approach using the high-dimensional information contained in cell phone data is also possible.

We implement the supervised learning approach as follows. We randomly select half of all survey areas as "training data," and predict survey income in the other half as "test data." We compare two models: first, using OLS with model-predicted workplace income. Second, we use elastic net regularization using 498 features extracted from cell-phone data (Blumenstock et al. (2015) uses a similar method). (See Appendix A.5 for more details.)

Panel (B) of Table 2 reports the results. Test R^2 and Training R^2 indicate the average R^2 in the training data and test data 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).[15] 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). 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.

Here, we compare the model measure with alternate predictors within the same setting. This helps hold constant factors such as how much underlying variation in the outcome variable there is across locations, as well as the level of aggregation, which affect the value of R^2 across the board, for any explanatory variable. In this setting, we focus on the urban core of Dhaka (due to DHUTS data availability), where prediction is significantly more difficult than if we were to include peri-urban areas (as we do in the next section).[16]

Results are robust to several alternate gravity equation specifications (Appendix Table B.5), and to replacing log income from the survey with the residual after partialing out age, gender, years of education, occupation and job sector (Appendix Table B.6). Appendix Table B.7 uses an individual-level specification and shows that our main result is robust to controlling for workplace sorting along observable worker characteristics: origin survey area fixed effects, geographic area

[15] Cell phone operators tend to locate more towers in locations with high activity. See Appendix Figure B.1 for maps of tower density.

[16] Indeed, the use of wider areas and more aggregate units helps explain the level of predictive power in previous studies. Using fine divisions (census tracts), Severen (2019) finds that model wages estimated using commuting flows barely predict tract-level wages in Los Angeles. Tsivanidis (2018) calibrates and estimates a general equilibrium model in Bogotá, and finds that model-predicted wages across 19 urban areas predict survey wages, with an $R^2 = 0.36$. For machine learning methods, Blumenstock et al. (2015) finds $R^2 \approx 0.35$ when restricting to 37 urban DHS clusters in Rwanda. Jean et al. (2016) do not report results separately by urban areas. However, within entire countries, DHS-cluster level predicted consumption explains between 0.37 and 0.55 of the variation in measured consumption.

of destination location, travel time, and including government workers.

4.2 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

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

where j indexes workplace towers, $\hat{\psi}_j^R$ is the area adjusted destination fixed effect at j, V_i^H is total residential population at i, and V_{ij} is the commuting volume from i to j.

Table 3 shows the results in Bangladesh. Model residential income is strongly related to the income proxy at the cell tower level (panel A). The $R^2 = 0.54$ is high, partly because of the coverage of suburban areas.[17] Residential density and distance to CBD are also highly correlated with residential income (columns 2-3).

Model income performs well at fine spatial resolution. The coefficient on model-predicted residential income remains large when including sub-district fixed effects (55 units in Dhaka), and when controlling for residential density, distance to CBD, as well as the model-predicted workplace income (column 4). Residential income remains significant, while workplace income is negative and significant. Once again, this indicates that residential and workplace model income are correlated, yet it is encouraging that the positive correlation loads onto the residential measure.

We next benchmark the predictive power to a supervised learning method (panel (B) Table 3). The procedure is similar to Panel (B) in Table 2. Test R^2 is 0.53 when using model-predicted income alone (column 1). The test R^2 when using the cell phone tower Voronoi cell area alone is 0.68 (column 2), and the supervised-learning method using all features increases it to 0.71. Model-predicted residential income alone achieves about 75% of the predictive power of using all the cell-phone data metrics.

Appendix Table B.8 repeats the same exercise in Colombo, Sri Lanka. Results are similar and model income has slightly better predictive performance ($R^2 = 0.77$). Results are robust to using

[17] The analysis covers three districts of Dhaka, Narayanganj, and Gazipur, which include suburban areas outside the DHUTS survey areas investigated in Section 4.1.

daily commuting flows and to excluding neighboring towers in the definition of residential income. Not area-adjusting destination fixed effects reverses the sign of the correlation with census income proxy, because of very large cells far away from the city center (Appendix Table B.9).

As a last robustness exercise, we explore the residential income validation exercise using survey income. We do not find any statistically significant correlation between model income and survey income at the residential level. Residential population density and distance to CBD also have very low R^2 (≤ 0.05). This may be due to lower underlying differences in average income at the *residential* level (compared to the workplace level) in the urban core area, and hence a more noisy measure. Indeed, residential model income also has lower explanatory power for the census income proxy when we restrict to towers inside the DCC urban core in Dhaka ($R^2 = 0.15$).

5 Application: The Impact of Hartal on Commuting and Forgone Income

We illustrate how high-frequency commuting data and the detailed model-predicted income measure can be used to measure the economic impact of sudden urban shocks.

Hartals are a form of political strike that involves a partial shutdown of urban transportation and businesses. They are common in South Asia, and especially in Bangladesh (UNDP 2005). On hartal days, typically announced a few days in advance by unions or political groups, groups of people enforce the transportation shutdown, especially on major roads and in certain locations. However, the ultimate impact of hartals on travel is an empirical question, as commuters may defy disruptions, change routes, or take advantage of the lower traffic congestion.

We use the cell-phone data and model-predicted income to quantify how the onset of hartal affects short-run commuting behavior and predicted take-home income.

Hartal dates in Dhaka are from Ahsan and Iqbal (2015). They identify 33 hartal days and we code 6 hartal events over the 4 months in our sample.[18] The study period preceded parliamentary elections and was marked by general instability, and hence hartals were more frequent than in previous years. Hence, our results may not directly generalize to periods with lower hartal intensity.

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

[18] Unfortunately, spatial information on hartal location is not available.

2.1, accounting for 35% of all users in the data.[19] We only observe travel behavior if a user makes calls on a given day, and call behavior itself may differ on hartal days. Hence, we include commuter fixed effects to ensure that our results are not driven by selection across different types of commuters. Moreover, restricting to frequent callers are almost identical (Appendix A.6).

Figure 2 shows the impact of hartal onset on the probability to commute (red, solid dots). Hartals have a sudden negative impact, reducing commuting probability by approximately 5% relative to the days just before. Appendix Figure A.3 shows for each calendar date, the change in probability to commute relative to workdays. Commuting probability is lower on all hartal days, yet not as much as on Fridays or some important holidays. Longer hartals spells have lower impacts on average.

Figure 2 also shows the impact of hartal onset on the income forgone due to lower commuting on hartal days. For each individual trip we assign an income as follows. From non-hartal days, we obtain destination log wages $\hat{\psi}_j^R$, which we assume do not change during the study period. In other words, we assume that workers earn a daily wage if they show up to work, and zero otherwise, and that market wages do not change given short-term fluctuations due to hartal or other events. Our empirical strategy does not quantify direct impacts of hartals on worker productivity, nor long-term adaptation costs. In Figure 2, the drop in predicted take-home income is slightly larger yet around the same magnitude as the drop in commuting (5-8%). The reduction in predicted income is driven primarily by the extensive margin, namely fewer trips. However, the difference is also statistically significant, as the commuting reduction is stronger for commuters who work in locations with high predicted wage (Appendix Table A.4).

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

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

6 Conclusion

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

Big data, such as cell phone or smartphone mobility records, credit card transactions, or user-generated reviews, are rapidly gaining popularity due to their ability to *predict* behavior, individual characteristics and economic conditions (Blumenstock et al. 2015, Jean et al. 2016, Glaeser et al. 2017, Björkegren and Grissen 2018).

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

References

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

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

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

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

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

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

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

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.*

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.

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

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.

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

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

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.

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

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

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.

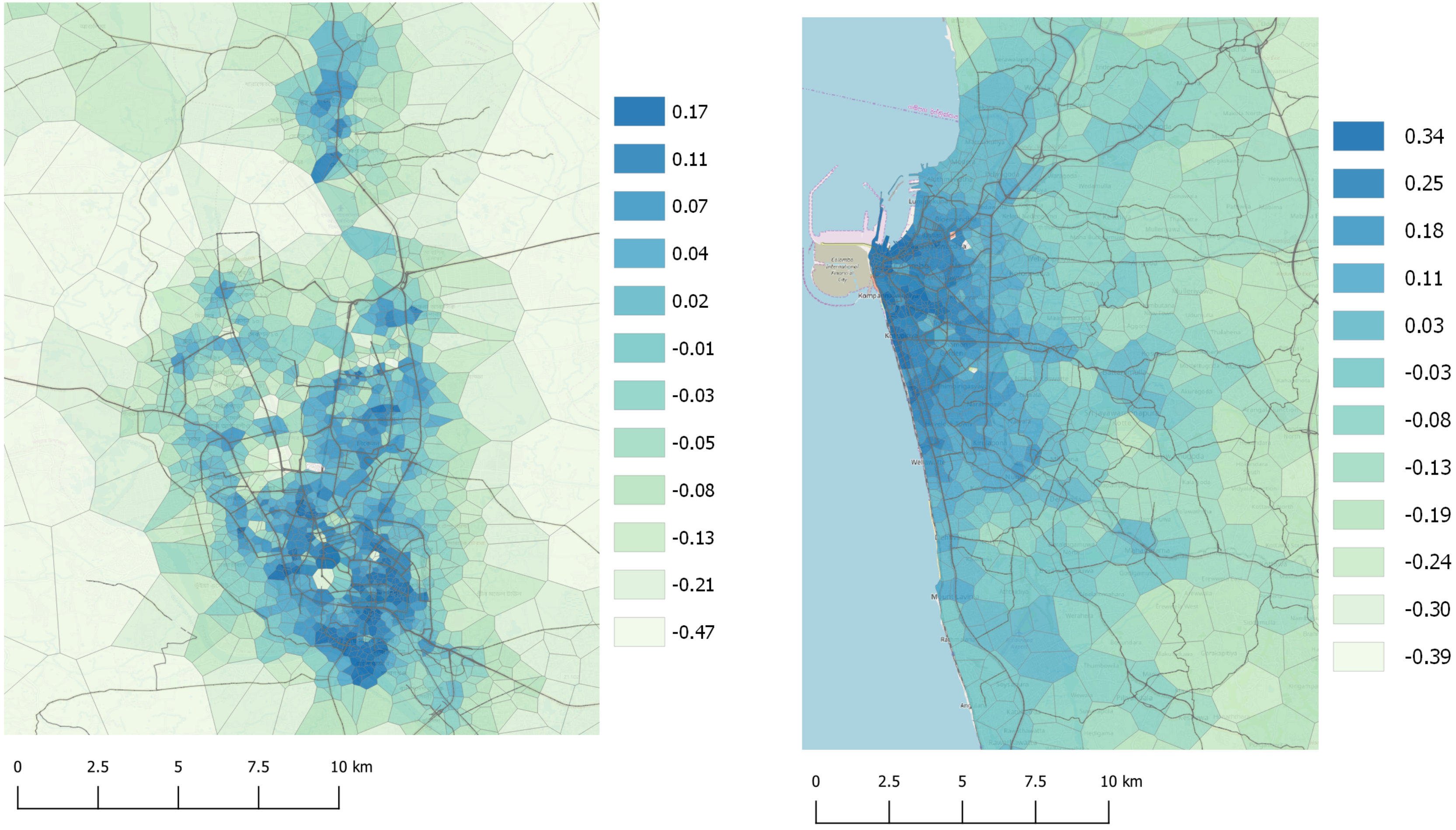
Figures and Tables

Table 1: Gravity Equation Estimation Results

	Commuting Probability			
	(1)	(2)	(3)	(4)
log Travel Time	-2.44	-2.55	-2.19	-2.22
	(0.0011)	(0.0003)	(0.0015)	(0.0001)
City	Dhaka	Dhaka	Colombo	Colombo
Commuting Measure	Home-Work	Daily	Home-Work	Daily
Number of Destination FE	1859	1868	1201	1201
Number of Trips	1.5e+06	1.9e+07	9.4e+05	1.3e+08
Observations	3.4e+06	3.4e+06	1.3e+06	1.3e+06
Pseudo R^2	0.67	0.82	0.66	0.88

Notes. This table reports estimates of the gravity equation (3) by Poisson pseudo-maximum likelihood (PPML) method with two-way fixed effects. The outcome variable is total commuting probability (π_{ij}) between a pair of cell phone towers, 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 and 3) and using the commuting flows identified at the daily level (columns 2 and 4) using cell phone data as described in Section 2.1. 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. 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 $\epsilon = 9.09$, as estimated in Appendix A.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) Comparison with other economic indicators

	log Survey Income (workplace)				
	(1)	(2)	(3)	(4)	(5)
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)
log Model Income (residential)					-0.12
					(0.15)
Adjusted R^2	0.25	0.06	0.33	0.42	0.42
Observations	88	88	88	88	88

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

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

Notes. This table compares survey and model predictions of average workplace income. The unit of analysis is a survey area from the DHUTS survey. The survey sample is 11,006 commuters who live and work inside the Dhaka City Corporation, who report positive income, excluding students, homemakers, the unemployed, and government workers. The outcome variable is the average income of survey respondents who work in a survey area, using log income truncated at the 99th percentile. 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). Regressions in both panels are weighted by survey area employment population (from the DHUTS survey). In Panel (A), the Central Business District (CBD) is Shapla Chatter in Motijheel. Conley standard errors with 5 km distance cutoff shown in parentheses. $^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

In Panel (B), Test R^2 and Training R^2 indicate the average R^2 in the training data and test data over 100 random splits. See Appendix A.5 for the description of the supervised learning method (elastic-net regularization) and cell phone data feature construction.

Appendix Table B.6 repeats the analysis using the residual of survey income on demographic and job characteristics.

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

(A) Comparison with other economic indicators

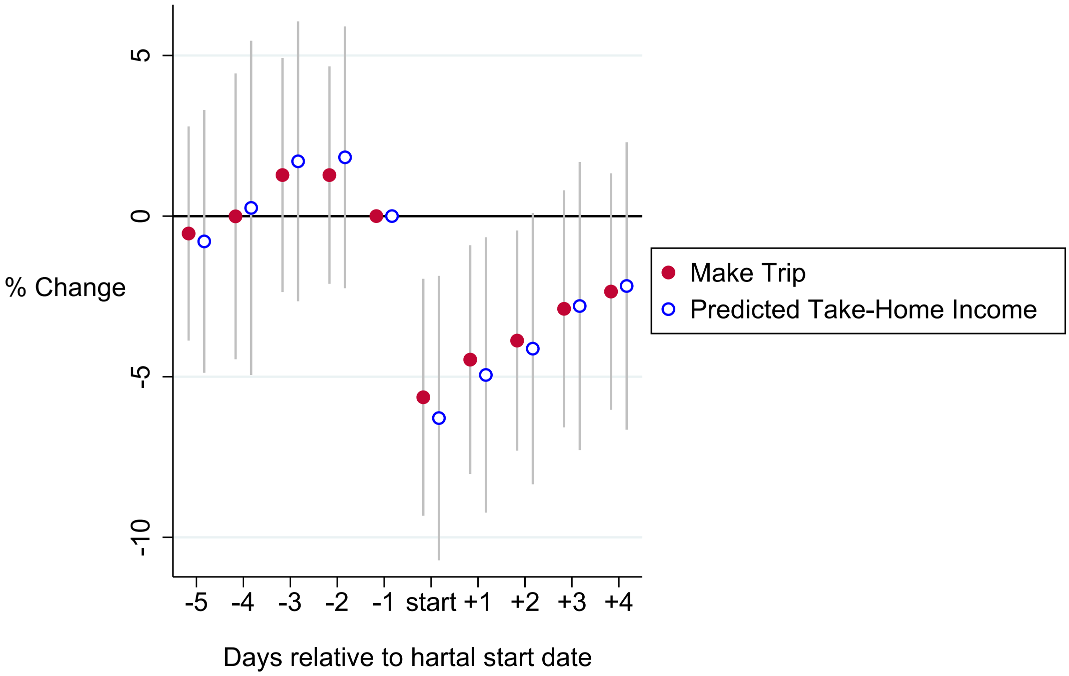
	Census Residential Income Proxy			
	(1)	(2)	(3)	(4)
log Model Income (residential)	0.89^{***}			0.64^{***}
	(0.06)			(0.23)
log Residential Density		0.67^{***}		0.37^{***}
		(0.02)		(0.06)
log Dist. to CBD			-0.84^{***}	-0.02
			(0.10)	(0.11)
log Model Income (workplace)				-0.35^{***}
				(0.13)
Sub-district FE (count)				X (55)
Adjusted R2	0.54	0.63	0.33	0.74
Observations	1,844	1,844	1,844	1,844

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

	(1)	(2)	(3)	(4)
Features	log Model Income (residential)	log Tower Area	All CDR Features	(3) + log Model Income (residential)
Training R^2	0.54	0.69	0.78	0.78
Test R^2	0.53	0.68	0.71	0.71
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 in the greater metropolitan area of Dhaka. 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 $\sum_j y_j V_{ij}/V_i^H$ where j indexes workplace towers, $y_j = \hat{\psi}_j^R$ is the area adjusted destination fixed effect at j, V_i^H 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). In Panel (A), the Central Business District (CBD) is Shapla Chatter in Motijheel. Column 4 controls for 55 sub-district (thana) fixed effects. Conley standard errors with 5 km distance cutoff shown in parentheses.$^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$
Panel (B) repeats the analysis in Table 2 panel (B). See Appendix A.5 for details.
Appendix Table B.8 repeats the analysis in Panel A for Colombo, Sri Lanka.

Figure 2: Impact of Hartal on Travel Behavior and Predicted Take-Home Income



Notes. This figure shows the event study impact of the onset of a hartal event on the probability to commute and on model-predicted income. 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 and on non-hartal days. The base analysis sample is all days with commuting data (including stationary trips). "Make trip" is a dummy for making a proper trip (origin distinct from destination). Predicted model income is $\exp\left(\hat{\psi}_j^R/\epsilon\right)$ for a trip to destination j, where $\hat{\psi}_j^R$ is the (area-adjusted) estimated destination fixed effect at j (our measure of log wages), and $\epsilon = 9.09$ is the Fréchet shape parameter. Predicted model income is set to zero when "Make Trip" is zero. To construct the figure, we first obtain calendar date fixed effects from a regression that also includes commuter fixed effects. We then adjust the date fixed effects for the average effect on Friday and Saturday, and create an unbalanced panel over six hartal events. Finally, we regress the date fixed effect on hartal event study time dummies. The bars represent 95% confidence intervals from robust standard errors. See section A.6 for details. Appendix Table A.4 reports corresponding regression results.

A Appendix

A.1 Availability of Conventional Data on Economic Activity 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.

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

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.

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

In Section 3, we assumed that workers are ex-ante identical. However, in appendix table B.6, we measure the model's predictive power *after* netting out individual demographic characteristics from survey income. Here, we show how this validation regression arises directly in a specific model with worker heterogeneity.

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

However, the average ξ_ω of commuters working in j affects average income at that location. Hence, the correct validation regression should control for average ξ_ω at location j from individual income. To the extent that ξ_ω depends on observable characteristics (gender, age, education level, occupation, job sector), this is exactly what the specification in Appendix Table B.6 achieves.

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

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

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

$$\begin{aligned}
\mathbb{E}y_{ij\omega}^{0,0} &= w_j \\
\mathbb{E}y_{ij\omega}^{0,1} &= w_j - \tau d_{ij} \\
\mathbb{E}y_{ij\omega}^{1,1} &= \frac{1}{\epsilon}\log\left(\sum_s \exp\left(\epsilon w_s - \epsilon\tau d_{is}\right)\right) - \frac{K}{\epsilon} \text{ for some absolute constant } K \\
\mathbb{E}y_{ij\omega}^{1,0} &= \mathbb{E}y_{ij\omega}^{1,1} + \tau d_{ij}
\end{aligned} \tag{5}$$

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

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

$$y_{ij\omega}^{\alpha_z,\alpha_d} = \alpha_z\alpha_d \cdot y_{ij\omega}^{1,1} + \alpha_z\left(1-\alpha_d\right) y_{ij\omega}^{1,0} + \left(1-\alpha_z\right)\alpha_d \cdot y_{ij\omega}^{0,1} + \left(1-\alpha_z\right)\left(1-\alpha_d\right) y_{ij\omega}^{0,0}. \tag{6}$$

Using (5) and dropping the constant K, this simplifies to

$$\mathbb{E}y_{ij\omega}^{\alpha_z,\alpha_d} = \frac{\alpha_z}{\epsilon}\left[\log\left(\sum_s \exp\left(\epsilon w_s - \epsilon\tau d_{is}\right)\right) + \epsilon\tau d_{ij}\right] + \frac{1-\alpha_z}{\epsilon}\left[\epsilon w_j\right] + \frac{\alpha_d}{\epsilon}\left[-\epsilon\tau d_{ij}\right] \tag{7}$$

The intuition of this expression is as follows. For the third term, if travel time affects income, we expect that people who commute further away have lower income. The difference between the first two terms is more subtle. If Fréchet shocks affect income, then the first term is the best explanatory variable for income.[20] If shocks do not affect income, the wage at the destination should be the best predictor of income.

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

$$y_{ij\omega}^S = \rho_1 \hat{X}_{ij}^1 + \rho_2 \hat{X}_j^2 + \rho_3 \hat{X}_{ij}^3 + \varepsilon_{ij\omega}^S, \tag{8}$$

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

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

$\hat{X}^1_{ij} = \log\left(\sum_s \exp\left(\hat{\psi}_s - \hat{\beta} d_{ij}\right)\right) + \hat{\beta} d_{ij}$, $\hat{X}^2_j = \hat{\psi}_j$ and $\hat{X}^3_{ij} = -\hat{\beta} d_{ij}$ are estimators of the three terms in square brackets in (7), computed using the gravity equation estimates. (Recall that $\hat{\psi}_j$ is a consistent estimator for ϵw_j, and $\hat{\beta}$ is a consistent estimator for $\epsilon \tau$.) Asymptotically, we have

$$\hat{\alpha}_z = \frac{\hat{\rho}_1}{\hat{\rho}_1 + \hat{\rho}_2},\ \hat{\alpha}_d = \frac{\hat{\rho}_3}{\hat{\rho}_1 + \hat{\rho}_2},\ \text{and } \hat{\epsilon} = \frac{1}{\hat{\rho}_1 + \hat{\rho}_2}. \tag{9}$$

Table A.1 reports the estimates of α_z, α_d, and ϵ based on estimating equation (8) with OLS, and using transformation (9). 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 (8), 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 A.1: 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 A.3. We estimated equation (8) 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 (9). 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 $\hat{\rho}_2 \geq 0$.

A.4 Model: Approximate Invariance to Aggregation Level

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

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

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

A.5 Details of Supervised-Learning Approach in Section 4

In Section 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.

A.5.1 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.[21] To capture the nonlinear patterns, for each variable described

[21] Note that, our cell phone data from Bangladesh only record outgoing calls, hence transactions refer to all outgoing calls only.

below, we include both the variable and its logarithm. Altogether, we have 498 tower-level features from this procedure.

User-level characteristics averaged at home and work locations. The first set of features are constructed as the average statistics of users at the identified home and work location level. 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 × 2 (home and work)).

Hourly statistics at the tower level. The second set of features are constructed for each hour of the day and tower level. 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 × 24 hours × 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 × (8+240+1)).

A.5.2 Elastic Net Regularization for Relevant Feature Selection

Given the large number of features (or variables) relative to the size of the data, 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):

$$\lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|) \tag{10}$$

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).

A.5.3 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 A.2 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 A.2: 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.[22]

Hyperparameter α robustness. Table A.3 shows the results where we choose different weights α of the L^1 and L^2 penalty regularization terms. $\alpha = 1$ 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. Note that our baseline result in Panel (B) of Table 2 was based on $\alpha = 0.5$. The results indicate that the predictive performance is lower for $\alpha = 0$, but stays the same for all other values of α.

Table A.3: Predicting Workplace Income: Different Weights for L^1 and L^2 Penalty Terms

	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	Elastic Net				
	(log Model Income)	(All CDR Features)				
Training R^2	0.26	0.61	0.53	0.44	0.50	0.45
Test R^2	0.22	0.17	0.24	0.24	0.24	0.23
α		0	0.25	0.5	0.75	1
Observations	88	88	88	88	88	88

A.6 Additional Results for the Impact of Hartal on Commuting and Forgone Income

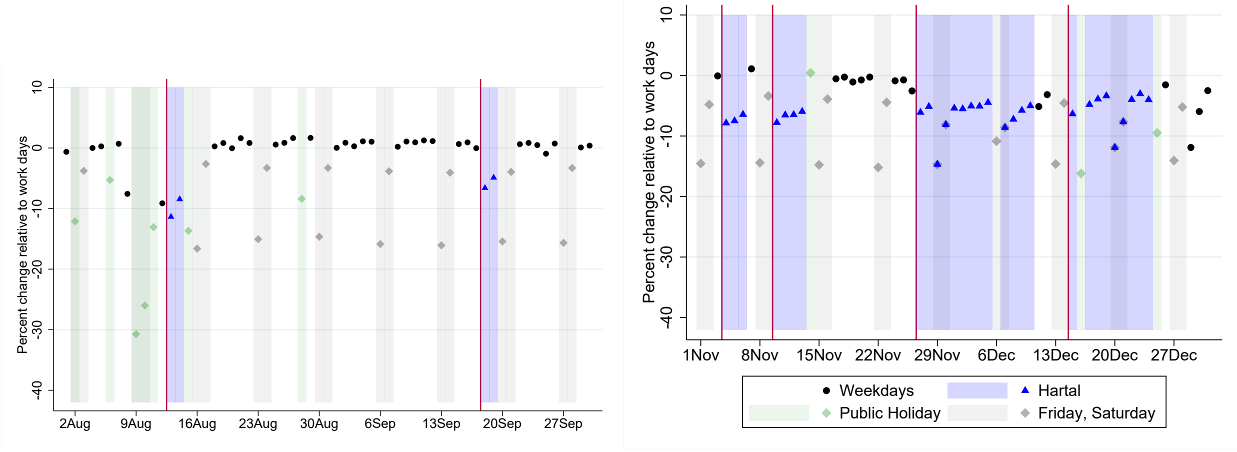
In order to construct the predicted income (up to scale), we need to choose the Fréchet parameter ϵ. We use $\epsilon = 9.09$, our estimate from the structural estimation method (Appendix A.3). The

[22] 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).

regression coefficient from Table 2 of log survey income (measuring ω_j) on the destination fixed effect ($\epsilon\omega_j$) implies a very similar number ($\hat{\epsilon} = 8.3 = 0.12^{-1}$). Since we are interested in *changes* in income during hartal days, the scale of $\hat{\psi}_j^R$, which is not identified, does not matter for this exercise. Finally, note that results are not particularly sensitive to the value of ϵ. Indeed, they are very similar using $\epsilon = 4.65$, which sets the variance of average income at the commuting zone level to the value in the DHUTS data (results not reported).

The event study in Figure 2 is constructed as follows. First, we compute calendar date fixed effects using the regression $X_{ct} = \psi_t + \mu_c + \epsilon_{ct}$ where c denotes a commuter, t denotes a calendar date, and X_{ct} is the outcome of interest. (Appendix Figure A.3 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 A.3). 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 preceding period, we exclude it and all previous days. We include all consecutive hartal days after it starts.

Figure A.3: 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 Panel A, Column 1 in Table A.4. 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. August 2 is Jumatul Bidah, August 6 is Shab-e-qadr, August 9-12 is the Eid ul-Fitr (end of Ramadan), August 15 is the National Mourning Day, August 28 is Janmashtami, November 14 is Ashura, December 16 is Victory Day, and December 25th is Christmas Day. The last week in December preceded the General Election of January 5, 2014. 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 A.4, panel A reports the average effect of hartal and heterogeneity by high-wage workplace and commute duration. Given that users may travel to different destinations on different days, in this table we use two definitions of commuting. The outcome in odd columns is a dummy for any proper trip (a trip with different origin and destination towers), while in even columns it is a dummy for proper trip going to the commuter's long-term workplace. To facilitate interpretation, all coefficients indicate proportional changes relative to the outcome mean on workdays.

The specification in the first two columns is:

$$y_{ct} = \beta^H Hartal_t + \beta^F Friday_t + \beta^S Saturday_t + \beta^{Ho} Holiday_t + \mu_c + \gamma_{Month(t)} + \varepsilon_{ct} \qquad (11)$$

for commuter c and calendar date t, where $Hartal_t$, $Friday_t$, $Saturday_t$ and $Holiday_t$ are date type dummies, and μ_c and $\gamma_{Month(t)}$ are commuter and month fixed effects. Throughout Table A.4, standard errors are clustered at the level of calendar dates.

Hartal days reduce any trip by around 5%, compared to a 14% reduction on Fridays. Work trips (daily trips where the destination corresponds to the user's long-term workplace) account for around 40% of all trips, and they decrease by 8% on hartal days and by 42% on Fridays. Hence, work trips are disproportionately affected on Fridays. This suggests a limited "destination selection" effect of hartals; commuters do not switch a lot to traveling to lower-wage destinations. In columns 3-4, we fully interact the model with an indicator for whether the commuter's long-term workplace location is below median in the predicted wage distribution. High-income commuters see large decreases in trips, both on hartal days and especially on Fridays. For work days, the hartal effect for high-income commuters is almost twice as large as for low-income commuters. Columns 5-6 document that this heterogeneity is not driven by heterogeneity in commute duration.

We now conduct an accounting exercise to estimate the income forgone due to lower commuting on hartal days. To do so, for each individual trip we assign an income as follows. First, we run our procedure on non-Hartal days and obtain predicted destination log wages $\hat{\psi}_j^R$, which we assume do not change during the study period. In other words, we assume that workers earn a daily wage if they show up to work, and zero otherwise, and that market wages do not change given short-term fluctuations due to hartal or other events. Hence, our empirical strategy does not quantify direct impacts of hartals on worker productivity, nor long-term adaptation costs.

As before, we assign income in two different ways. With the first method ("all trips"), for a proper trip from i to $j \neq i$, the commuter "earns" the wage $\exp\left(\hat{\psi}_j^R/\epsilon\right)$, and zero for trips with $i = j$. This is meant to capture the fact that workers may earn income from different destinations on different days. In the second method ("work trips"), the commuter "earns" $\exp\left(\hat{\psi}_j^R/\epsilon\right)$ only if the destination j is her long-term workplace location. The assigned income is zero when $i = j$ or when $i \neq j$ but j is not her long-term workplace.

The results show that the drop in predicted take-home income is around the same magnitude as the drop in commuting (5-8%). These results show that most of the reduction in predicted income is driven primarily by the extensive margin, namely fewer trips.

As robustness, restricting the sample to 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 in the sample, does not substantially change the results. The results from Table A.4 change as follows: first, the number of observations becomes ≈ 6 million. Second, in panel A, the coefficients on Hartal become -0.040 and -0.049 in the first two columns. Third, in panel B, the coefficients on Hartal become -0.046 and -0.051 in the first two columns.

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

	(1)	(2)	(3)	(4)	(5)	(6)
	All Coefficients: Proportional Change From Workday Mean					
	All Trips	Work Trips	All Trips	Work Trips	All Trips	Work Trips
Panel A. Make a Trip						
Hartal	-0.049^{***}	-0.081^{***}	-0.054^{***}	-0.102^{***}	-0.053^{***}	-0.100^{***}
	(0.007)	(0.016)	(0.008)	(0.019)	(0.008)	(0.019)
Friday (free day)	-0.142^{***}	-0.423^{***}	-0.173^{***}	-0.542^{***}	-0.175^{***}	-0.558^{***}
	(0.010)	(0.023)	(0.011)	(0.029)	(0.011)	(0.030)
Hartal x Low Income			0.011^{***}	0.045^{***}	0.011^{***}	0.046^{***}
			(0.003)	(0.010)	(0.003)	(0.010)
Friday x Low Income			0.067^{***}	0.253^{***}	0.066^{***}	0.247^{***}
			(0.005)	(0.016)	(0.005)	(0.016)
Hartal x Short Commute					-0.008^{***}	-0.014^{***}
					(0.001)	(0.003)
Friday x Short Commute					0.010^{***}	0.121^{***}
					(0.001)	(0.005)
Observations	26,165,887	26,165,887	26,165,887	26,165,887	26,165,887	26,165,887
Workday Mean	0.74	0.32	0.74	0.32	0.74	0.32
Panel B. Predicted Income						
Hartal	-0.056^{***}	-0.084^{***}				
	(0.009)	(0.016)				
Friday (free day)	-0.180^{***}	-0.439^{***}				
	(0.012)	(0.023)				
Observations	26,165,887	26,165,887				

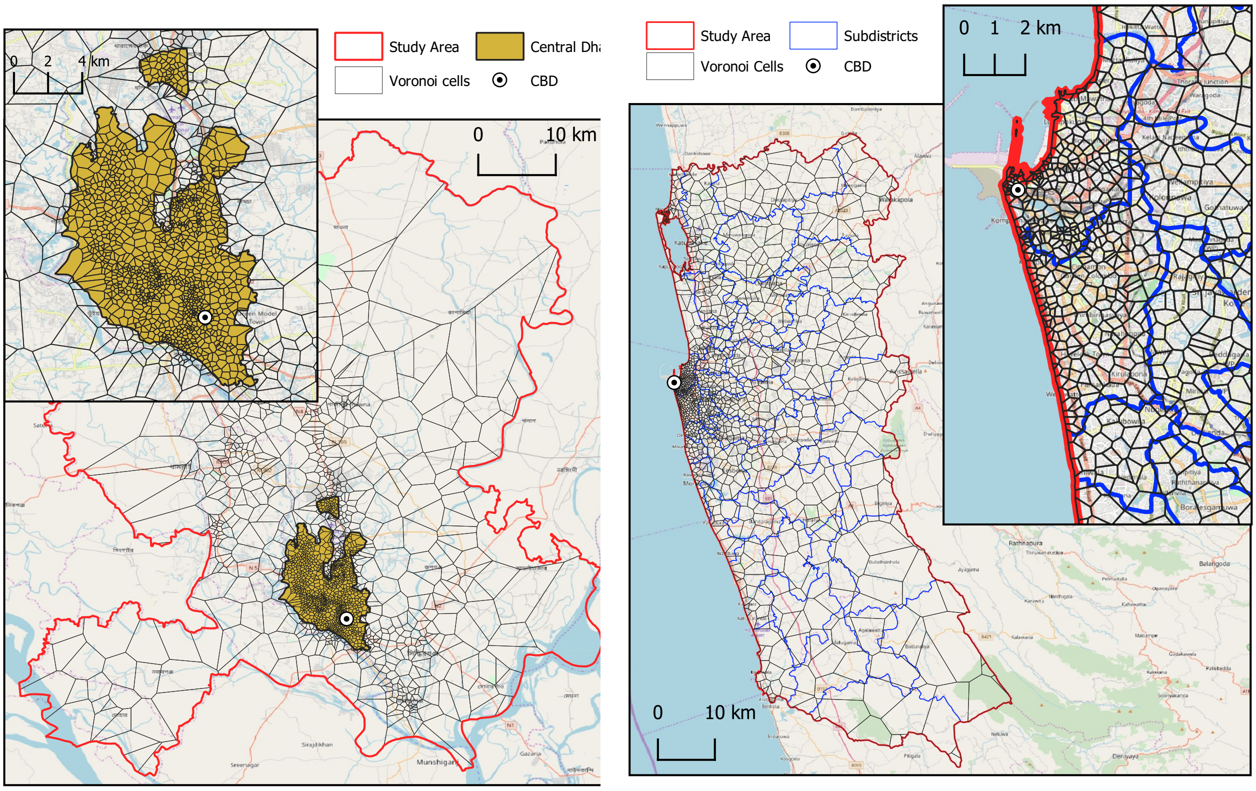
Notes. This table shows differences in travel probability and predicted income on hartal days and Fridays relative to workdays. All coefficients show proportional changes relative to the outcome mean on non-hartal, non-holiday workdays. The sample is all days with commuting data (including stationary trips) for commuters whose long-term residential and workplace towers are different (35% of all users). For commuter c on calendar date t, denote their trip origin by i_{ct}, destination by j_{ct}, and c's long-term workplace by j_c^W. In panel A, the outcome is a dummy for proper trip ($j_{ct} \neq i_{ct}$) in odd columns, and a dummy for proper *workplace* trip ($j_{ct} = j_c^W \neq i_{ct}$) in even columns. In panel A columns 3-6, we fully interact the model with dummies for low-wage commuters (c's long-term workplace wage $\hat{\psi}^R_{j_c^W}$ is below-median) and short-commute commuters (c's travel time between long-term home and work is below-median). In panel B, the outcome is predicted income; in the first column, commuters earn the destination wage $\exp\left(\hat{\psi}^R_{j_{ct}}/\epsilon\right)$ for any proper trip and zero otherwise. In the second column, commuters earn positive income only when $i_{ct} \neq j_{ct} = j_c^W$ and zero otherwise. In both cases, the gravity equation is estimated on non-hartal weekdays, and we use the Fréchet shape parameter set to $\epsilon = 9.09$ (see section A.6 for details). All regressions include commuter and month fixed effects, and dummies for Saturday and holidays. Standard errors clustered at the calendar date level in parentheses. ${}^{*}p \leq 0.10$, ${}^{**}p \leq 0.05$, ${}^{***}p \leq 0.01$

B Additional Figures and Tables

B.1 Cell-Phone Data and the Validation of Commuting Flows

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

(A) Dhaka (B) Colombo



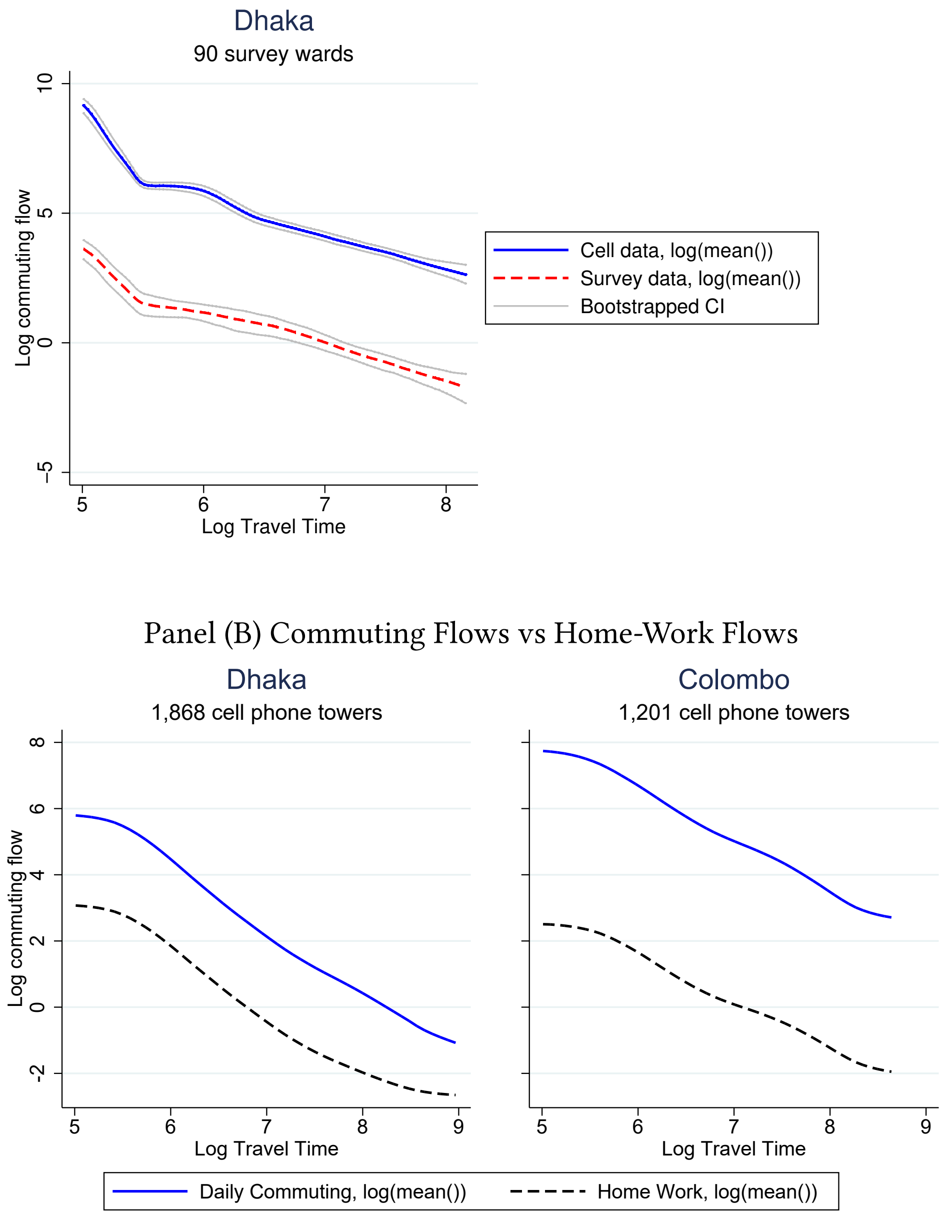
Notes. This figure shows the map of cell phone tower Voronoi cells in Dhaka, Bangladesh (Panel A), and in Colombo, Sri Lanka (Panel B). The yellow shaded area is the Dhaka City Corporation (DCC), the urban core of Dhaka, the main sample in the DHUTS transportation survey. The overall study area covers 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 B.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 (= $(5) \times (6)$)	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 B.2: Commuting Flows from Survey Data and Cell Phone Data



Panel (A) Survey vs Cell Phone Data

Panel (B) Commuting Flows vs Home-Work Flows

Notes. This figure compares the decay of commuting flows with travel time in survey and cell phone data. The unit of analysis is 7,836 survey area pairs in Panel A, and $1.6 \cdot 10^6$ and $1.4 \cdot 10^6$ tower pairs in Dhaka and Colombo in Panel B, respectively. Panel A compares commuting flows from the DHUTS survey (red, dash) and from cell phone data (blue, solid) in Dhaka. Panel B compares daily commuting trips (blue, solid) and home-work commuting trips (black, dash). See Section 2.1 for the definition of home-work 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 B.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 B.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. $^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

Table B.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, namely the total number of commuting trips from a given origin excluding stationary trips (including them yields virtually identical results). Census residential population in a Voronoi cell is computed as the average census population in census geographic units (Mauza for Bangladesh, Grama Niladhari for Sri Lanka), weighted by their 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 slope above one indicates that the cell phone data slightly over-represents residential population in denser areas. The comparatively lower adjusted R-squared in columns (3) and (4) 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. $^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

B.2 Estimation of Gravity Equation

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

	Destination Fixed Effects (Benchmark)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
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 (Travel Time with Congestion)									0.98***
									(0.003)
Estimation Method	PPML	PPML	OLS	OLS	PPML	PPML	OLS	OLS	PPML
City	Dhaka	Dhaka	Dhaka	Dhaka	Colombo	Colombo	Colombo	Colombo	Colombo
Observations	1,859	1,859	1,859	1,859	1,201	1,201	1,201	1,201	1,201
Adjusted R^2	0.92	0.88	0.81	0.68	0.92	0.87	0.82	0.62	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 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. This leads to a flatter profile of destination fixed effects. Omitting zero flows results in overestimation of destination fixed effects for locations with low wages (in third row). Incorporating the zero flows by arbitrarily adding one does not solve this issue. Standard errors in parentheses. $^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

B.3 Validation of Workplace Income with DHUTS Survey Workplace Income

Table B.5: 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 (panel A) and B.6 (panel A). Odd and even columns correspond to the specifications in columns 1 and 5 of Panel A of Table 2. The first two columns use commuting flows defined at the daily level instead of commuting flows from home and work assignment (see Section 2.1 for the definition). The next two columns define workplace income at the survey-area level excluding commuters whose origin towers are within 180 seconds of the destination cell tower, when we aggregate up from cell tower level. The next two columns use destination fixed effects not adjusted for Voronoi cell tower. The last two columns include commuters from DHUTS survey whose origin locations are outside the DCC area (see Section 2.1).

Table B.6: Average Workplace Income: Survey Income Residualized by Demographic Characteristics

	log Survey Income (workplace, residual)				
	(1)	(2)	(3)	(4)	(5)
log Model Income (workplace)	0.06^{***}			0.06^{***}	0.10^{*}
	(0.01)			(0.02)	(0.05)
log Employment Density		0.06^{**}		-0.04	-0.03
		(0.03)		(0.03)	(0.03)
log Dist. to CBD			-0.08^{***}	-0.05^{***}	-0.06^{***}
			(0.02)	(0.02)	(0.02)
log Model Income (residential)					-0.06
					(0.10)
Adjusted R2	0.2	0.05	0.17	0.27	0.26
Observations	88	88	88	88	88

Notes. This table replicates Panel A, Table 2 replacing log survey income by the residual of log income on gender, age, years of education, occupation and job sector dummies. See the footnote of Table 2 for the specification.

Table B.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.) $^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$

B.4 Residential Income Validation

Table B.8: Average Residential Income: Model Prediction and Residential Income Proxy in Colombo, Sri Lanka

	Census Residential Income Proxy			
	(1)	(2)	(3)	(4)
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. This table repeats the analysis in Table 3, panel (A), in Colombo, Sri Lanka. Column 4 controls for 41 sub-districts (Divisional Secretariat) fixed effects. The Central Business District (CBD) is Colombo Fort. For Sri Lanka, beyond commuting flows, we do not have access to the cell phone data necessary to construct the features used in the supervised learning method in Bangladesh.

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

(A) Dhaka

	Census Residential Income Proxy					
	(1) Daily Flows		**(2) Excluding Neighboring Towers**		**(3) Without Area Adjustment**	
log Model Income (residential)	1.08^{***} (0.08)	0.37^{***} (0.12)	0.93^{***} (0.06)	0.82^{***} (0.17)	-1.52^{***} (0.11)	-0.82^{***} (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

(B) Colombo

	Census Residential Income Proxy					
	(1) Daily Flows		**(2) Excluding Neighboring Towers**		**(3) Without Area Adjustment**	
log Model Income (residential)	1.69^{***} (0.08)	0.68^{***} (0.14)	1.48^{***} (0.08)	1.00^{***} (0.33)	-1.52^{***} (0.31)	-0.62^{***} (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 3 and B.8. Odd and even columns correspond to the specifications in columns 1 and 4 in Tables 3 and B.8. The first two columns use daily commuting flows instead of home-work commuting flows (see Section 2.1 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.