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

Table of Contents

A. Conventional and Modern Data Availability in Developing Countries
B. Model Extension and Estimation: Worker Skill Heterogeneity
C. Model Extension and Estimation: Preference or Productivity Shocks
D. Robustness: Gravity Equation Over-identification
E. Supervised-Learning Method Details
F. Additional Results: the Impact of Hartal
G. Additional Figures and Tables

A Conventional and Modern Data Availability in Developing Countries

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

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

Of these, 16 ever had an economic census, 11 covered informal firms. However, at most 4 included wage data, which accounts for between 5.6 and 8.6% of the urban population of all countries in the sample. (The 2014 Ghana and 2015 Zimbabwe censuses included wage data, while for the ongoing censuses in Mali and Togo we do not know if wage data was collected.)

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

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

B Model Extension and Estimation: Worker Skill Heterogeneity

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

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

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

present results from the first model.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

C Model Extension and Estimation: Preference or Productivity Shocks

In the main analysis, we assume that an agent earns income directly proportional to her wage. Formally, the Fréchet shocks $Z_{ij\omega}$ and travel time D_{ij} affect utility but not income. Here, we relax this assumption and allow $Z_{ij\omega}$ and D_{ij} to affect income instead. We show that the model income continues to be correlated with survey income. Furthermore, we develop a method to estimate *the degree* to which $Z_{ij\omega}$ and D_{ij} affect income instead of preferences.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table C.2 reports the estimates of α_z, α_d, and ϵ based on estimating equation (C.4) with OLS, and using transformation (C.5). We report two types of standard errors: based on the Delta method (in round parentheses) and based on bootstrapping at the origin survey area level (in square parentheses). In columns 1-2, we estimate the full equation (C.4), and we find that $\hat{\alpha}_d$ is close to zero with a small and insignificant negative value, and the other parameters are imprecisely estimated when using bootstrapped standard errors. Given that the model restricts $\rho_3 \geq 0$ (from $\alpha_d \in [0,1]$), in columns 3-4 we restrict the coefficient on travel time to be equal to zero ($\rho_3 = 0$) and estimate the other two parameters. This increases the point estimate for $\hat{\alpha}_z$ and slightly lowers that for $\hat{\epsilon}$ while improving precision.

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

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

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

Notes. This table reports estimates of the structural parameters that control the degree to which idiosyncratic shocks affect income (α_z), travel time affects income (α_d), and the Fréchet shape parameter ϵ, using the procedure described in Appendix C. We estimated equation (C.4) by regressing individual log survey income from the DHUTS survey on the three model-predicted terms. In columns 4 and 5, we restrict the third coefficient that corresponds to travel time to be zero ($\rho_3 = 0$). The estimates for α_z, α_d and ϵ in this table are transformations of the estimated OLS coefficients as detailed in equation (C.5). Columns 1 and 4 report standard errors computed using the Delta method. Columns 2, 3, and 4 report results from 100 bootstrap runs where we cluster at the origin survey area level (70 survey areas with at least one out-commuter in DHUTS survey). The coefficient is the median estimate and standard errors in square parentheses. Column 3 censors $\hat{\rho}_1 \geq 0$ and $\hat{\rho}_2 \geq 0$.

D Robustness: Gravity Equation Over-identification

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

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

Panel B repeats the validation exercise in Bangladesh. Column 1 repeats the analysis Panel A in Table 2, while the next two columns use model income computed using destination fixed effects from the gravity equation estimated on the "close" and on the "far" samples, respectively. In both cases, the model income measure is predictive of log survey income, with a similar slope. The explanatory power is higher when using the "far" sample.

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

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

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

E Supervised-Learning Method Details

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

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

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

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

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

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

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

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

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

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

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

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

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

Elastic Net Regularization for Relevant Feature Selection

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

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

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

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

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

Additional Robustness Results with DHUTS Survey Workplace Income

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

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

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

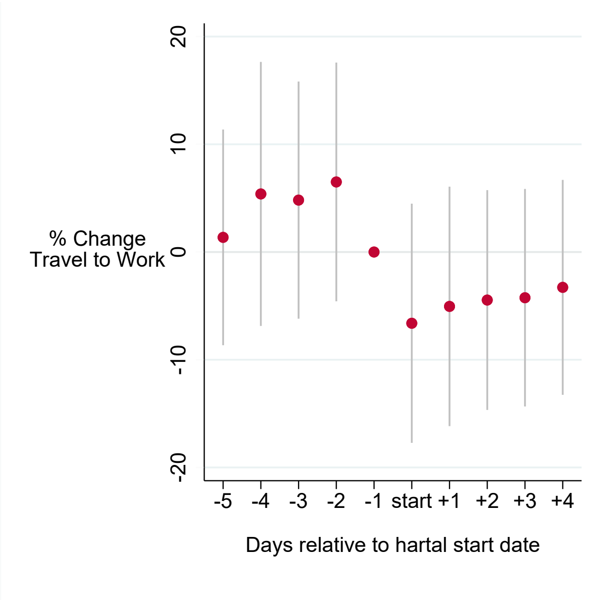
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	OLS	Elastic Net					
	(log Model Income)	(All CDR Features)					
		Maximize Test R^2	CV	CV	CV	CV	CV
Training R^2	0.26	0.44	0.44	0.48	0.50	0.51	0.53
Test R^2	0.22	0.24	0.19	0.18	0.13	0.16	0.12
Number of Folds for CV			3	5	10	20	44
Observations	88	88	88	88	88	88	88

Columns (3)-(7) show that the test R^2 falls when we use the cross-validation procedure for choosing λ. In fact, test R^2 is lower than the OLS with model-predicted income. At the same time, training R^2 is higher than in columns (1) and (2), suggesting that poorer predictive performance is likely due to overfitting. Overfitting is unavoidable given the small sample size.[4]

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

F Additional Results: the Impact of Hartal

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



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

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

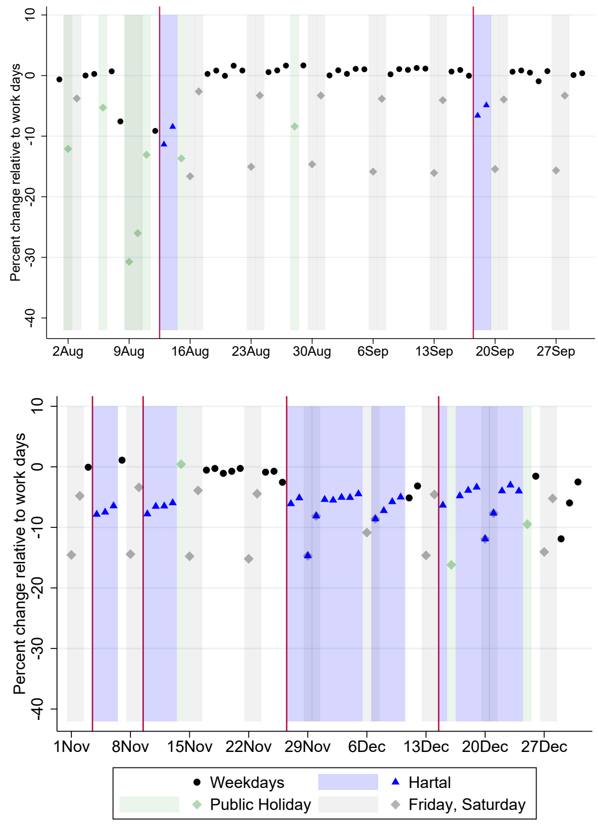
To construct this figure, we proceed as follows. First, we compute calendar date fixed effects using the regression $C_{\omega t} = \psi_t + \mu_\omega + \epsilon_{\omega t}$ where ω denotes a commuter, t denotes a calendar date, and $C_{\omega t}$ is a dummy for commuting to work. (Figure F.2 plots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

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

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

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

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



Notes. This figure shows average commuting probability by calendar date. The Y axis plots the percentage change relative to the mean on non-hartal, non-holiday workdays. The sample and outcome are as in Figure F.1. The figure plots calendar date fixed effects from a regression of any trip commuting dummy on commuter and calendar date fixed effects. Hartal dates are from Ahsan and Iqbal (2015) and public holidays from https://www.timeanddate.com/holidays/bangladesh/. The red vertical lines indicate hartal event onset date for the six hartal events. Friday is the main free day in Bangladesh, and Saturday is the other weekend day. Five days in the sample are both hartal and weekend: August 13, September 18, November 4, 10, and 27, and December 15. We drop these throughout the analysis.

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

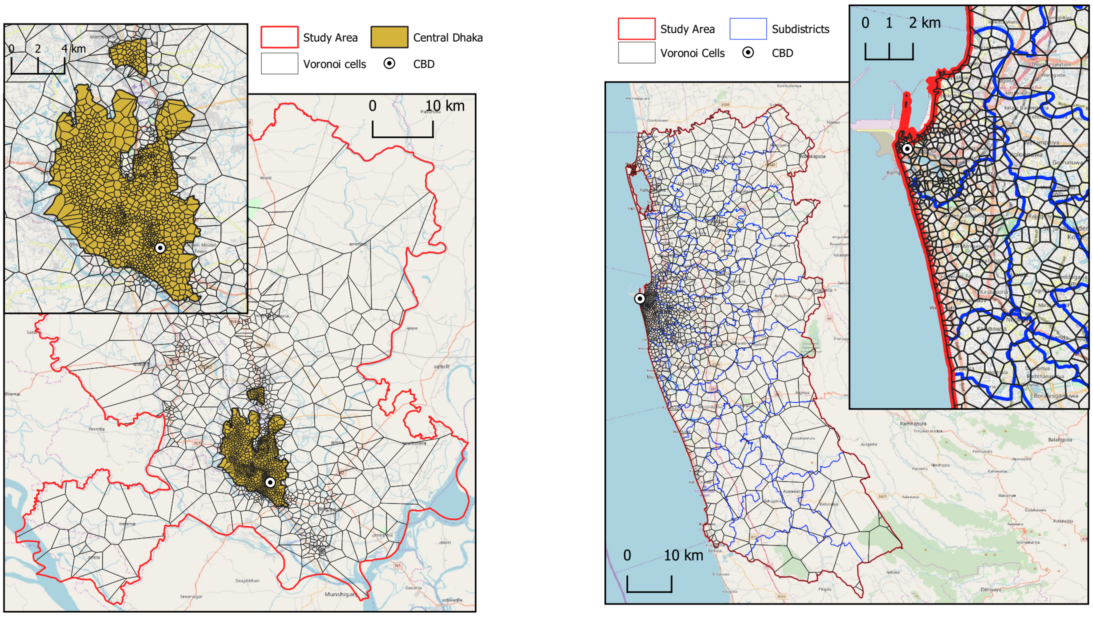
	Work Commute (% change vs weekday)			
	(1)	(2)	(3)	(4)
Hartal	-0.053***	-0.053***	-0.053***	
	(0.004)	(0.004)	(0.004)	
Interactions: Hartal $\times$				
(β^L) % Low Skill				-0.025**
				(0.012)
(β^H) % High Skill				-0.051***
				(0.004)
Dest. FE (z)		-0.022***	-0.019***	
		(0.005)	(0.005)	
(β^L_W)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

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

G Additional Figures and Tables

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

(A) Dhaka (B) Colombo



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

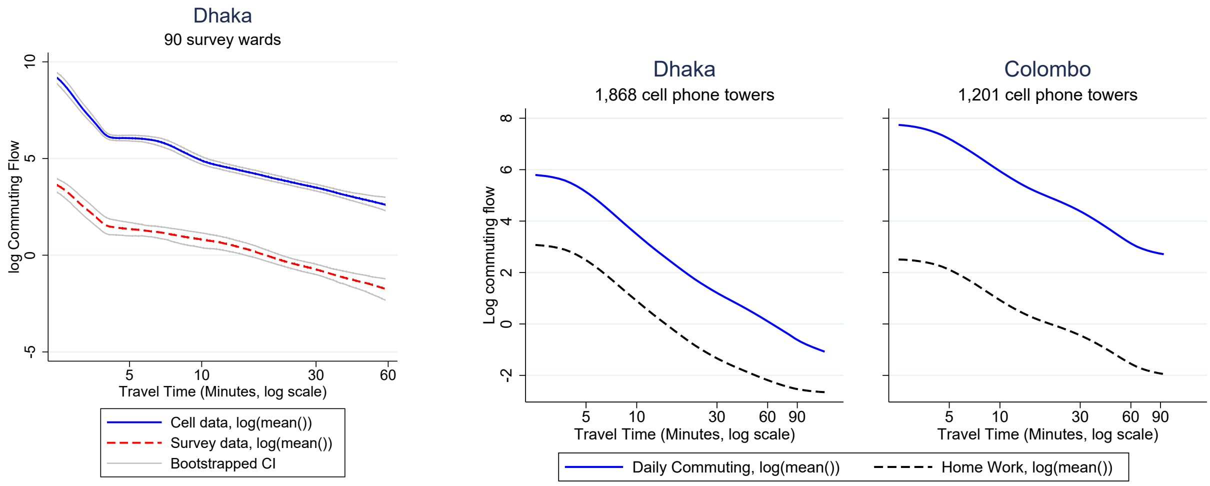
Table G.1: Cell Phone Data Coverage at User-Day Level

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

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

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

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 1 for the definition of home-work and daily commuting trips. In each graph, commuting flows are first averaged within each of 100 equal bins of log travel time below the 99th percentile, and the plot shows the local linear regression of log mean commuting flow on log travel time. This procedure avoids the bias due to zero commuting flows, which is important for survey and home-work commuting data. The DHUTS sample (described in Table G.2) has 12,510 commuters. The cell phone data sample has $18 \cdot 10^6$ trips in Panel A, and $38 \cdot 10^6$ daily trip and $5.2 \cdot 10^6$ for home-work trips in Dhaka, and $237 \cdot 10^6$ daily trips and $2.6 \cdot 10^6$ home-work trips in Colombo, in Panel B. In Panel A, pointwise bootstrapped 95% confidence intervals clustered at the origin survey area shown in gray.

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

	Flow survey data (DHUTS)			
	(1)	(2)	(3)	(4)
Log flow cell phone data	0.63***	0.70***	0.30***	0.53***
	(0.020)	(0.026)	(0.059)	(0.049)
Log duration			-1.05***	-0.51***
			(0.17)	(0.11)
Origin and destination fixed effects		Yes		Yes
Observations	6026	6026	6026	6026

Notes: This table shows the relationship between commuting flows from two different data sets in Dhaka: the DHUTS transportation survey (outcome) and home-work comuting flows from cell phone data (explanatory variable). The survey sample consists of the 12,510 commuters who live and work within the 90 survey areas inside the DCC and who report positive income from work, excluding students, homemakers, and the unemployed. (The sample includes government workers.) An observation is a pair of survey areas from the DHUTS survey. The coefficients show the estimates from the Poisson pseudo-maximum-likelihood (PPML) estimation of DHUTS commuting flow on log flows from cell phone. We use PPML to deal with the presence of zeros in DHUTS commuting flows (Silva and Tenreyro 2006). If cell phone commuting flow data is a perfect measure of commuting flows, one would expect coefficients equal to one. Standard errors are clustered at the origin survey area level. $^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

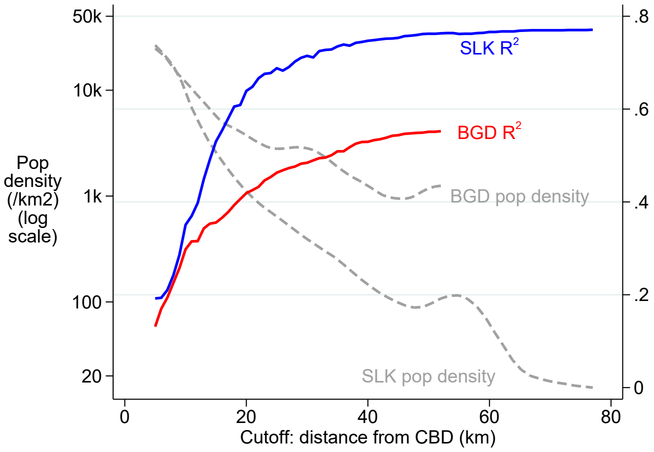


Figure G.3: Distance to CBD and R^2

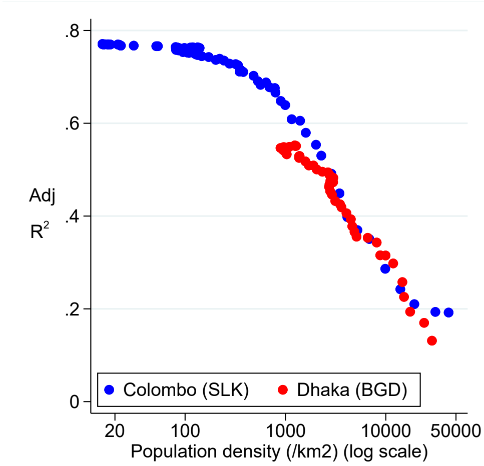
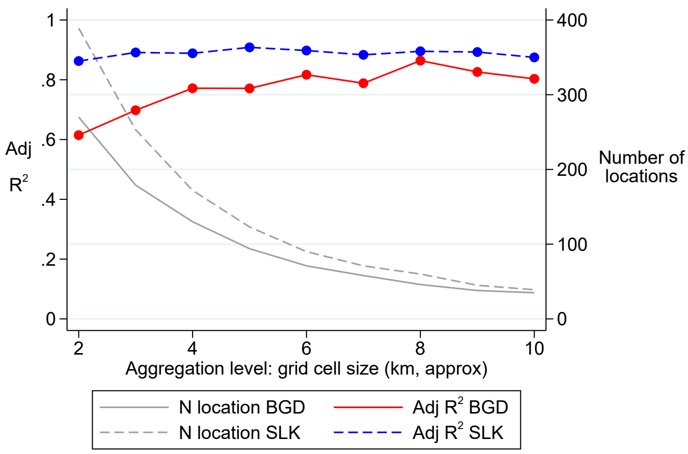


Figure G.4: Population Density and R^2

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

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



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

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

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

Notes: This table shows the representativeness of the cell phone data at the residential level. The unit of analysis is a Voronoi cell around each cell phone tower in the greater metropolitan area of each city (Dhaka, Gazipur, and Narayanganj districts in Bangladesh, and Western Province in Sri Lanka). In cell phone data, residential population is defined as out-commuting flow. Census residential population in a Voronoi cell is computed as the average census population in the finest available census geographic units , weighted by their area overlap with the Voronoi cell. The high adjusted R-squared in columns (1) and (2) indicates a strong association between the geographic density from the two data sources. The comparatively lower adjusted R-squared in columns (3) and (4) may be due to the fact that cell phone operators tend to assign cell phone towers to equalize the subscriber coverage per tower. Conley standard errors with 5 km distance cutoff shown in parentheses. $^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$.

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

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

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

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

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

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

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

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

Notes. Robustness for Table 2 and Table G.5. Odd and even columns correspond to the specifications in columns 1 and 5 of Table G.5. The first two columns use commuting flows defined at the daily level instead of commuting flows from home and work assignment (see Section 1 for the definition). The next two columns define workplace income at the survey-area level excluding commuters whose origin towers are within 180 seconds of the destination cell tower, when we aggregate up from cell tower level. The next two columns use destination fixed effects not adjusted for Voronoi cell tower. The last two columns include commuters from DHUTS survey whose origin locations are outside the DCC area. (In the main analysis, we exclude households outside of DCC, because the 18 corresponding survey areas are significantly coarser and detailed information on sampling is not available.)

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

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

Notes: This table regresses log income from the DHUTS survey on model-predicted income and controls. The unit of observation is a survey respondent in the sample described in Table 2. Model-predicted income for a pair of origin and destination survey areas is the weighted average of tower-pair model income, with weights given by tower-to-tower commuting flows. Formally, for survey areas a and b, $y_{ab} \equiv \sum_{i \in a, j \in b} V_{ij} / V_{ab} \cdot y_j$, where $i \in a$ and $j \in b$ index towers, $y_j = \hat{\psi}_j^R$ is the area-adjusted destination fixed effect at j, and $V_{ab} \equiv \sum_{i \in a, j \in b} V_{ij}$ is the total flow between a and b. We assign to each survey respondent the predicted income between his or her home and work survey areas. Columns 2, 3 and 4 include origin survey area fixed effects, and column 4 includes occupation and job sector fixed effects. Conley standard errors with 5 km distance cutoff in parentheses. (For computational purposes, when including fixed effects, the standard errors are computed after residualizing the fixed effects.) $^{*}p \leq 0.10$, $^{**}p \leq 0.05$, $^{***}p \leq 0.01$

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

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

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

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

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

Notes. Robustness for panel (A) in Tables G.8. Odd and even columns correspond to the specifications in columns 1 and 4 in Tables G.8. The first two columns use daily commuting flows instead of home-work commuting flows (see Section 1 for definitions). The next two columns define workplace income at the survey-area level excluding commuters whose origin towers are within 180 seconds of the destination cell tower, when we aggregate up from cell tower level. The last two columns use destination fixed effects not adjusted for Voronoi cell tower area.