Traffic congestion and labor supply: Evidence from smartphone data

(Job Market Paper)

Minoru Higa*

November 30th, 2022

[Click here for the latest version]

Abstract

Does traffic congestion affect time allocation? I use highly granular smartphone data from Mexico City to empirically study how traffic congestion affects work-time allocation. I find that traffic increases hours worked. The effect is driven by workers leaving work later, rather than by changes in arrival time. I show modest evidence that labor income does not increase despite the increase in total hours worked. These results highlight an avoidance mechanism (consistent with bottleneck models) that has been previously overlooked when estimating the costs of congestion.

Keywords: traffic congestion, labor supply, big data

JEL Classifications: Q5, J22, R41, O18

* *Contact Information:* Department of Economics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada (e-mail: minoru_higa@sfu.ca)

1 Introduction

Traffic congestion has become one of the "plagues of modern life" in most cities worldwide, depleting the benefits that cities offer (Arnott and Small, 1994). Congestion contributes to air pollution, increases crime, and wastes valuable leisure time spent seated in traffic. However, traffic congestion may also distort work-time allocation decisions, which would have important welfare implications if (for example) changes in hours worked are not compensated by changes in income. Likewise, this may lead to reconsidering the way we measure congestion externalities, a crucial factor in the calculus of congestion pricing and assessing the benefits of different urban transit policies.

This paper examines the effect of traffic congestion on work-time allocation. Because the relationship between these two variables is theoretically ambiguous, researchers have sought empirical methods to identify the causal effects. However, the lack of data directly measuring work time and traffic has prevented researchers from doing so up to this point, despite the high degree of policy relevance. The current debates over automobile use in the developed world (e.g., investment in electric vehicles or autonomous vehicles) and trends in developing countries (e.g., rapid growth in urban population and private vehicle ownership) indicate that congestion will likely increase over time.

The main contribution of the paper is to identify and quantify an unintended externality of traffic congestion previously overlooked when estimating the costs of congestion. Existing estimates of welfare loss from traffic congestion only consider the time lost on congested roads (Akbar et al., 2020, Kim, 2019). However, this may underestimate the real costs of congestion in two ways: (i) by missing the costly avoidance behavior of staying at work longer without receiving additional compensation, and (ii) as these measures of time lost include commuting times already reduced by the avoidance behavior.

To answer this question, I use data for one of the most congested cities in the world, Mexico City (Akbar et al., 2020, INRIX, 2019). I build a unique longitudinal dataset with individual daily hours spent at work and daily exposure to district-level traffic congestion for 2019. The smartphone data allows me to track the daily work-time allocation of individuals and to identify where individuals work (once combined with geocoded establishment-level data). I exploit the richness of the smartphone data to recover each individual's arrival and departure time for work. Traffic congestion is measured using GPS sensors installed in vehicles circulating around the city, and proxied by the inverse of the average speed. My identification strategy exploits within-district daily variation in traffic congestion. I complement this approach using road accidents as an exogenous shifter of traffic congestion.

I find that traffic congestion increases the time workers spend at work. The magnitude is economically relevant. In a single day, doubled traffic congestion lengthens the workday by one hour. This effect is driven by congestion during the afternoon rush hour. I also find that workers adapt to traffic congestion in the sense that individuals working in more congested areas are less affected than individuals working in less congested areas.

The positive effect of traffic congestion on hours at work is robust. Replacing hours at work using smartphone data with self-reported hours worked from household surveys does not affect the results. This finding is also robust to using road accidents as a measure of exogenous variation in traffic congestion.

I find that individuals stay longer at work primarily because they delay their departure time. These results are consistent with the bottleneck model where one may choose when to start their commute in response to congestion. These results may also suggest the presence of labor market frictions that prevent workers from arriving late to work or departing earlier. Hence, a potential mechanism for this effect is that workers respond to traffic congestion by departing later from work, despite starting at the same time or earlier.

Even though workers stay longer at work, labor income does not seem to increase. I find suggestive evidence that workers are not paid more. One potential explanation may be that workers stay longer one day at work in response to congestion, but they compensate by leaving earlier another day, therefore creating minimal change to the total hours of work over a given week or month. However, I do not find evidence of that compensating behavior in the short run. Alternatively, workers may be rewarded in the future for their longer hours today. However, this cannot be explored in this study due to data limitations.

This study contributes to a broader literature analyzing the effects of commuting costs on labor supply. In these studies, commuting costs are usually measured by changes in distance (Fu and Viard, 2019, Gutiérrez-i-Puigarnau and van Ommeren, 2010) or changes in commuting time (Black et al., 2014, Gutiérrez-i-Puigarnau and van Ommeren, 2015). However, we cannot attribute results from those studies to changes in traffic congestion, while the work presented in this study can. This study also contributes to the literature on environmental outcomes and labor supply. Previous literature indirectly addresses the relationship between traffic congestion and work-time allocation, investigating how driving restriction policies affect leisure time (Viard and Fu, 2015). One contribution of the present study is to add traffic congestion as a new variable of interest. Second, this paper use novel "big data" from smartphones to track individuals' daily time allocation, particularly, the number of hours at work, and work arrival and departure times. With these new sources of data, I can directly study the relationship between traffic congestion and work-time allocation.

The remainder of the paper is organized as follows. Section 2 describes the conceptual framework. Section 3 describes the data used to measure hours spent at work and traffic congestion. Section 4 discusses the empirical approach and identification concerns. Section 5 describes the results. Section 6 presents the discussion. Finally, section 7 concludes.

2 Conceptual Framework

There are two main models to understand the relationship between traffic congestion and labor supply: (i) the bottleneck model, and (ii) the standard neoclassical model of labor-leisure choice with commuting costs.

The bottleneck model (Arnott et al., 1990, 1993, Noland and Small, 1995, Small, 1982, Vickrey, 1969) allows individuals to choose when to start their commute to respond to congestion. Hence, individuals may choose to leave earlier from home to avoid the morning rush hour or delay their departure time from work to avoid the afternoon rush hour. Consequently, congestion may change the number of hours allocated to work. However, this model has not been yet used to study the effect of congestion on labor supply. It is focused mainly on the morning commute and on the "schedule delay" which is the difference between arrival time to work and some ideal time that usually coincides with the time work starts (*e.g.* 9 am). The model uses the schedule delay to measure the social welfare loss due to congestion (Kim, 2019).

On the other hand, traffic congestion can be seen as a shifter of commuting costs. Black et al. (2014) introduce commuting time costs in the labor supply model. In this model, traffic congestion increases commuting time costs, and this increases the value of leisure relative to the value of working. This effect may push some individuals to work fewer hours or to exit the labor force. However, in a two-person household, if the labor supply of one of the members is negatively affected by the increase in commuting costs, the household will face a negative income shock. Then the other member increases the time allocated given that leisure is assumed to be a normal good. The effect on the overall labor supply is ambiguous, but the negative effect on labor force participation is unambiguous. In a similar fashion, Gutiérrez-i-Puigarnau and van Ommeren (2010) develop a labor supply model with both time and monetary commuting costs. These are variable costs when choosing workdays, but fixed costs when deciding the number of work hours within a day. Given an increase in commuting costs, workers may respond by working fewer days to avoid extra commuting costs but may increase the number of hours worked per day to mitigate a reduction in income. It is again ambiguous which of these effects dominates.

3 Data

This paper aims to estimate the effect of traffic congestion on work-time allocation. This requires longitudinal information that links individual hours worked with traffic congestion on a daily basis. Ideally, traffic congestion should be measured on the individual's commuting route considering their preferred mode of transport. Data with such granularity is not available yet. I, therefore, construct a novel longitudinal dataset combining smartphone data that allows me to track the time allocation of individuals with daily traffic congestion that comes from GPS sensors installed in vehicles. The unit of observation is the owner of the smartphone device. I restrict the sample to manufacturing and office workers. My final dataset consists of an unbalanced panel of 6,709 observations, representing 1,262 devices for all sixteen districts in Mexico City (CDMX) in 2019.

Tables 1 shows the description of the main variables. I approximate hours worked with the number of hours spent at work when using the smartphone data. Table 2 shows summary statistics. The number of hours spent at work using the smartphone data is higher than self-reported hours worked using household surveys, on average. The average worker arrives to work around 9 am and departs from work at approximately 7 pm. This pattern occurs either in high- or low-congested districts.

Table 1: Variable description and data source

Variable	Description	Data Source
Hours worked	Number of hours spent at workplace	Quadrant
Hours worked	Number of self-reported hours worked	ENOE (INEGI)
Arrival time	Device's first time at work (in 24h format)	Quadrant
Departure time	Device's last time at work (in 24h format)	Quadrant
Traffic congestion	Inverse of average speed (h/km)	Dat's why
Accidents	Number of confirmed road incidents by CDMX 911	Gobierno CDMX
Temperature	Average temperature (in Celsius)	CONAGUA
Precipitation	Rain (in mm)	CONAGUA
Humidity	Relative humidity (in %)	CONAGUA
Daylight hours	Difference between sunshine and sunset times.	CONAGUA

Notes: This table presents the description of the main variables and their corresponding source. All variables are available from January-December 2019.

Smartphone data. This data is provided by Quadrant, a private organization specializing in high-quality mobile location-based data. The raw data consists of pings (i.e. the time and location of a given smartphone) collected from applications installed in deidentified smartphones. A ping is recorded every time the location of the smartphone is requested by the applications installed on the device. This data provides representative information on the population in Mexico City. Figure A.1 shows that the total population at the district level according to the Census 2020 is correlated with the total number of smartphone devices with an R^2 of 0.62.

I use this data to build a panel of individuals and identify where they work and live, and the number of hours they stay at work. I follow individuals for several days within a week, during all weeks in 2019 except for the first and the last weeks of the year given that patterns in working hours and congestion may be particularly unusual in these two weeks. I combine this data with geocoded establishment-level information to identify workplaces, and with residential areas from census data to identify homes. See the data appendix for details about the algorithm used for this purpose. The richness of this data allows me to know both the time individuals arrive and depart from work. Using this information, I estimate the

Table 2: Summary Statistics

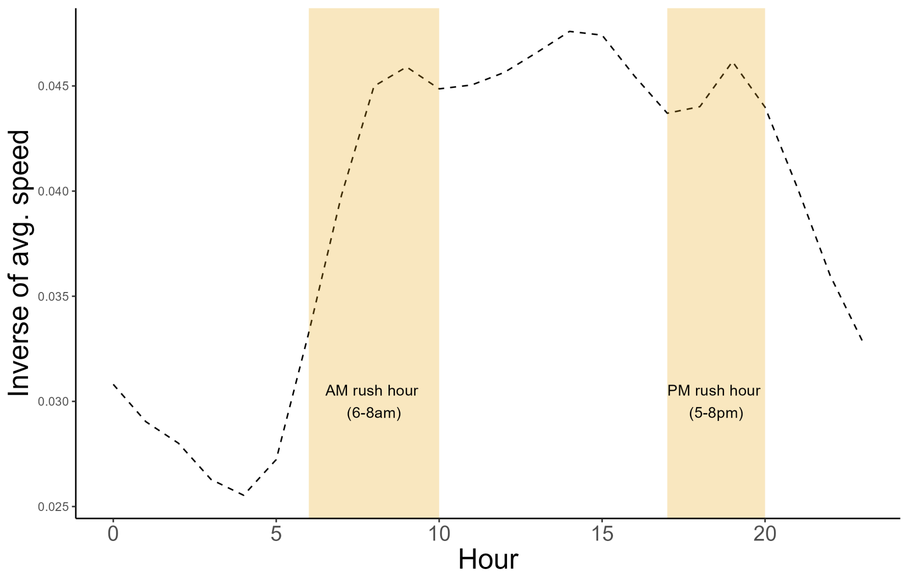
	All	High-congestion districts	Low-congestion districts
	(1)	(2)	(3)
Panel A. Labor outcomes			
Hours worked (daily, mobile data)	10.47	10.45	10.50
Hour worked (daily, survey data)	8.36	8.36	8.37
Arrival time	8.84	8.89	8.79
Departure time	19.32	19.34	19.29
Panel B. Traffic congestion			
Inverse of avg. speed (h/km)	.042	.044	.039
Panel C. Weather			
Temperature (C)	19.96	19.97	19.94
Humidity (%)	63.02	63.06	62.98
Rain (mm)	0.02	0.02	0.02
Daylight (hours)	12.07	12.05	12.10
No. of Smartphones	1,262	671	591
No. of Observations	6,709	3,723	2,986

Notes: This table presents mean values for the main variables. Arrival and departure time are in 24 hours format.

number of hours individuals stay at work which I use as a proxy of hours worked.

This dataset has three main limitations: (i) It becomes sparse very quickly. The raw semi-unstructured data contains billions of pings per month. However, most of the devices are observed either once or multiple times within a single day. For instance, imposing the structure described in the data appendix to identify workplaces reduces drastically the number of observations. Hence, there is a trade-off between the number of observations and the reliability of the statistics. (ii) This data does not provide information regarding the demographics of the owners of the devices such as gender, age, etc. Socio-economic characteristics can be inferred from the neighborhood of residence or points of interest (POI) visited regularly. (iii) This data does not provide work and home locations, which then need to be inferred using supplemental data.

Figure 1: Distribution of traffic congestion



Notes: The figure depicts the distribution of traffic congestion per hour using data from Dat's Why for Mexico City in 2019. Morning (6 am-10 am) and afternoon (5 pm-8 pm) rush hours are highlighted in yellow.

Traffic congestion data. This data is provided by Dat's Why, a private company with the largest real-time Big Data network of smartphones, vehicles, and sensors in Mexico to monitor traffic congestion. The raw data consists of hourly average speed measures at the street segment level in Mexico City for every day of 2019. I use this data to build a district-level panel of daily average speed.

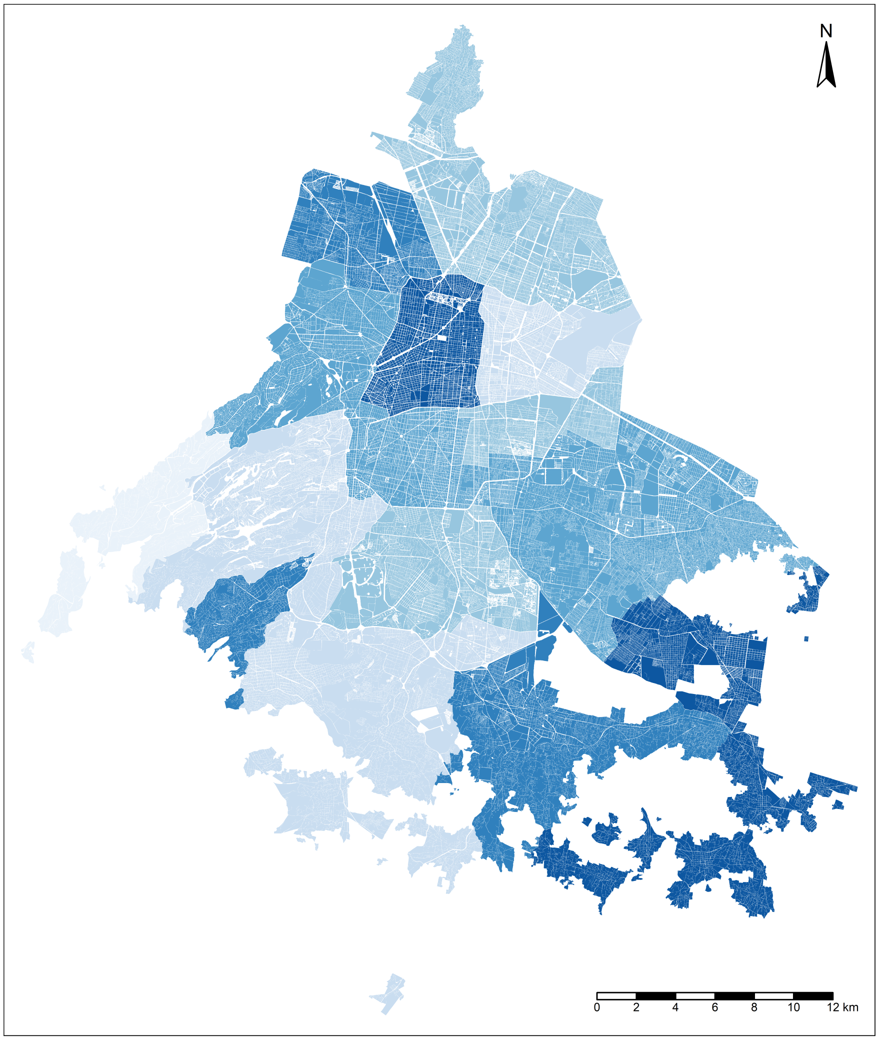
I use the inverse of average speed as a proxy of traffic congestion as in Hanna et al. (2017). In addition to the daily average traffic congestion, I use this data to calculate the traffic congestion during the morning (6 am-10 am) and afternoon (5 pm-8 pm) rush hours for Mexico City. Figure 1 shows the distribution of congestion per hour using data from Dat's Why.

This data is representative of the traffic congestion in Mexico City. Figure A.3 shows the distribution of congestion per hour using aggregated data from Waze reported in Calatayud et al. (2021) and provided by the corresponding authors. We can observe that both distributions in Figure 1 and Figure A.3 show a similar pattern of traffic congestion in the city. Both distributions capture the morning and afternoon rush hours for similar hours during the day. The correlation between them is 0.93.

Supplemental data. I complement the smartphone and traffic congestion data with information for the year 2019 about establishments, residential venues, self-reported income and hours worked, weather, daylight hours, and road accidents. I use the National Statistical Directory of Economic Units (DENUE) to obtain the latitude and longitude coordinates for the location of the establishments, the size of the firm, and the economic sector. See the distribution of establishments with 50 workers or more in Figure A.5 in the appendix. I use information from the 2020 Census and the National Geostatistical Framework (MGN) to identify residential venues in the city. To address identification concerns regarding omitted variables related to weather I use information about temperature, precipitation, and relative humidity from monitoring stations. Likewise, I include daylight hours calculated by taking the difference between sunset and sunrise times. I use self-reported hours worked from household surveys (ENOE) as an alternative outcome variable to the hours worked built using the smartphone data. Finally, I use road accidents from administrative records as an exogenous source of variation in traffic congestion to address different identification concerns. See the appendix for more details about these data sources, and about Dat's Why and Quadrant.

Location This study uses information from Mexico City (CDMX) in 2019. Mexico City is one of the most congested cities in the world. It is more congested than cities such as Mumbai and Delhi in India, and New York in the US (Akbar et al., 2020). For instance, it was the third most congested city in the world in 2019 (INRIX, 2019). Also in 2019, residents lost more than 600 million hours due to congestion representing a cost of more than twice

Figure 2: Spatial distribution of congestion in Mexico City



Notes: Figure depicts a map of Mexico City with the average annual traffic congestion per district in 2019. The darker the more congested is the district.

the budget assigned for education in the city (Calatayud et al., 2021). Figure 2 displays a map of Mexico City with the average congestion per district.

Mexico City is an ideal setting to study the impacts of traffic congestion on our well-being. It is ideal not only because it is one of the most congested cities in the world (Akbar et al., 2020, INRIX, 2019). Studying the context of a city in a developing country is relevant given current trends in urban population and motorization rates (Akbar et al., 2020, Calatayud et al., 2021, Kreindler, 2022). First, the urban population is growing rapidly. By 2050, approximately 2.5 billion people will migrate to cities in developing countries. This may pressure cities in the developing world where the transportation infrastructure is already outdated to the current population size. Second, private vehicle ownership is also growing rapidly. This is because of increasing motorization rates due to economic growth.

4 Empirical Approach

Baseline regression To explore the effect of traffic congestion on work-time allocation, I estimate the following regression model:

$$y_{ijt} = \delta_t + W_{jt} + \beta \times \ln(\text{Traffic Congestion})_{jt} + \epsilon_{ijt}, \tag{1}$$

where the unit of observation is individual i working in district j in day t. y_{ijt} represents the labor outcome variables such as hours worked, and arrival and departure times from work. Traffic Congestion$_{jt}$ is proxied by the inverse of the average speed. W_{jt} is a set of weather variables that include temperature, precipitation, humidity, and daylight hours. δ_t includes day of the week and month fixed effects. Once we divide it by 100, β can be interpreted as the unit change in the outcome variable when traffic congestion increases by 1%. I estimate the model using OLS and clustering standard errors at the week-district level. Identification comes from assuming that within-district daily variation in traffic congestion is exogenous conditional on weather and fixed effects or from quasi-random (temporal) variation in (demean) traffic congestion across days.

Identification concerns To study the effect of traffic congestion on work-time allocation, ideally, we would like to observe how many hours a person works where there is and there is no traffic congestion on a given day. However, we cannot observe the counterfactual for each person. We can only observe the hours worked either when there is or there is no traffic congestion, but not the hours worked under both scenarios. Alternatively, we can design a randomized controlled trial where, ceteris paribus, we randomly assign traffic congestion to a group of workers (treated group) and no traffic congestion to another group of workers (control group) on a given day. We can then compare the average hours worked between groups to find the average treatment effect. However, traffic congestion cannot be randomly assigned.

I include weather controls and time fixed effects in the baseline panel regression to reduce concerns regarding omitted variables bias. Traffic Congestion is not randomly assigned and confounders elements in ϵ_{ijt} may be correlated with both traffic congestion and our outcome y_{ijt}. For example, rainy days may be positively correlated with both congestion and hours worked, or darkness of the day may be positively correlated with congestion, and negatively correlated with hours worked. Alternatively, a higher temperature may be negatively correlated with congestion and hours worked. Hence, I control for temperature, precipitation, humidity, and daylight hours in W_{jt}. Likewise, Fridays may be positively correlated with congestion, but negatively correlated with hours worked, or a particular month may experience a decline in business activity that both affect congestion and work-time allocation. Thus, I control for day of the week and month fixed effects. I also address individual time-invariant unobservables by individual fixed effects as part of the robustness checks.

It is likely that my measures of hours worked and traffic congestion contain measurement error. As described in the data section, my measure of traffic congestion seems to represent the patterns regularly observed on the streets of Mexico City. Regarding hours worked, I use self-reported hours worked from household surveys (ENOE) as an alternative outcome variable.

Reverse causality is unlikely in a context where I only follow individuals during a week and estimate short-run effects. Changes in traffic congestion patterns may affect the spatial distribution of economic activities. In response to these changes, residents may re-optimize their decision of where to live, work or consume. Hence, there may be sorting. However, changes in the spatial distribution of economic activities may affect patterns in traffic congestion (Allen and Arkolakis, forthcoming). This can be a problem for the long-run effects of congestion on work-time allocation, but not for the short run. Given the focus on the short run, all the analysis is conditional on sorting (i.e. sorting already took place). It is unlikely that we see people shifting residential areas or workplaces across days during the short period of analysis.

Instrumental variable I complement the baseline approach with an instrumental variable approach to address concerns regarding potential omitted variables, measurement error and reverse causality that may persist. For example, there may be time-variant unobservables that are correlated with congestion and that also affect hours worked. Hence, I complement the identification strategy in the baseline regression with an instrumental variable design. I use road accidents as my instrumental variable as in Beland and Brent (2018). This instrument is relevant, as good as random, and, I argue, it satisfies the exclusion restriction that accidents are only affecting hours worked via changes in traffic congestion. In this context, accidents introduce exogenous variation in traffic congestion to lessen concerns regarding omitted variable bias, measurement error, and reverse causality.

5 Results

5.1 Main results

Table 3 shows the main results. Column (1) shows the results for the baseline model in Equation 1 estimated using OLS. The outcome variable is hours worked approximated by the daily number of hours spent at work constructed based on the smartphone data. Traffic congestion is measured as the inverse of the daily average speed in the district where individuals work. Column (1) indicates that a ten percent increase in traffic congestion increases time at work by 0.13 hours. The estimated coefficient is statistically significant at the five percent level of significance. Identification in Equation 1 may be affected by individual time-invariant unobservables. Column (2) shows that results are robust to adding individual fixed effects. The estimated coefficient is similar in magnitude and statistically significant at the five percent level of significance. It indicates an effect of traffic congestion on hours worked in the same direction as in column (1), and that omitting individual fixed effects was generating a negative bias on the estimated coefficient.

Column (3) shows results using a 2SLS approach where road accidents act as an instrumental variable for traffic congestion. These results reduce concerns regarding omitted variables, measurement error in traffic congestion, and reverse causality. The instrumental variable satisfies the relevance condition. The first-stage estimated coefficient is 0.01 and it is statistically significant at the one percent level. Regarding the exclusion restriction, it is unlikely that individuals change the hours worked in response to road accidents for different reasons than the traffic congestion generated by the accidents. Column (3) indicates that a ten percent increase in traffic congestion increases time at work by 0.18 hours. The estimated coefficient is statistically significant at the five percent level of significance. It is higher than the estimated coefficient in column (1) probably because the instrumental variable is cor-

Table 3: The effect of traffic congestion on hours worked

	Dependent variable: Hours worked		
	(1)	(2)	(3)
Traffic Congestion (log)	1.277**	1.744**	1.769**
	(0.518)	(0.778)	(0.894)
Method	OLS	OLS	2SLS
Individual FE	No	Yes	No
Observations	6,333	6,271	6,307
R-squared	0.087	0.763	0.004

Notes: Standard errors clustered at week-district level in parentheses. Statistical significance at the one, five, and ten percent levels is indicated by ***, **, and *, respectively. Traffic congestion is measured as the inverse of the daily average speed at the district level. All regressions use smartphone data and include weather (average daily temperature, precipitation, humidity), daylight hours, and day of the week and month fixed-effects. The sample considers manufacturing and office workers only. In column (3), the Kleibergen-Paap rk Wald F-statistic is 157.96 and the first-stage coefficient is 0.01 and statistically significant at the one percent level.

recting for measurement error. I obtain qualitatively the same result when adding individual fixed effects. Even though the sign of the coefficient is the same, the estimates lose a lot of precision (see details in column (2) in Table B.1).

Previous studies have not always found a positive effect of commuting costs on labor supply. Commuting costs are usually measured by changes in distance or changes in commuting time. Some studies find that either increase in commuting distance (Fu and Viard, 2019) or commuting time (Black et al., 2014) reduces labor supply in China or reduces female labor force participation in the US. Other studies find that either increase in commuting distance (Gutiérrez-i-Puigarnau and van Ommeren, 2010) or commuting time (Gutiérrez-i-Puigarnau and van Ommeren, 2015) have a positive effect on hours worked in Germany and in the UK. Another study indirectly addresses the relationship between traffic congestion and labor supply investigating how changes in driving restriction policies affect changes in leisure time (Viard and Fu, 2015). They find a positive effect of driving restriction on leisure time for self-employees and a negative effect for hourly-wage workers.

The magnitude of estimated coefficients in Table 3 is not small. Gutiérrez-i-Puigarnau and van Ommeren (2010) find that doubling the commuting distance increases labor supply by approximately 15 minutes per week, which is equivalent to 13 hours per year. However, I find that doubling traffic congestion increases hours worked in one hour per day, which is equivalent to five hours per week or 260 hours per year. One important difference with respect to the commuting costs literature is that we cannot attribute results from those studies to changes in congestion since changes in commuting time can be a result of a less direct commute or route, or there may be a change in the commuting distance but without changes in traffic congestion.

Table 4: Robustness checks

	Dependent variable: Hours worked		
	(1)	(2)	(3)
Traffic Congestion (log)	1.375***	0.120**	1.238***
	(0.487)	(0.049)	(0.425)
Change in specification	Week FE	Outcome in log	All sectors
Observations	6,333	6,333	15,870
R-squared	0.110	0.076	0.082

Notes: Standard errors clustered at week-district level in parentheses. Statistical significance at the one, five and ten percent levels is indicated by ***, ** and *, respectively. All regressions use the baseline model and smartphone data.

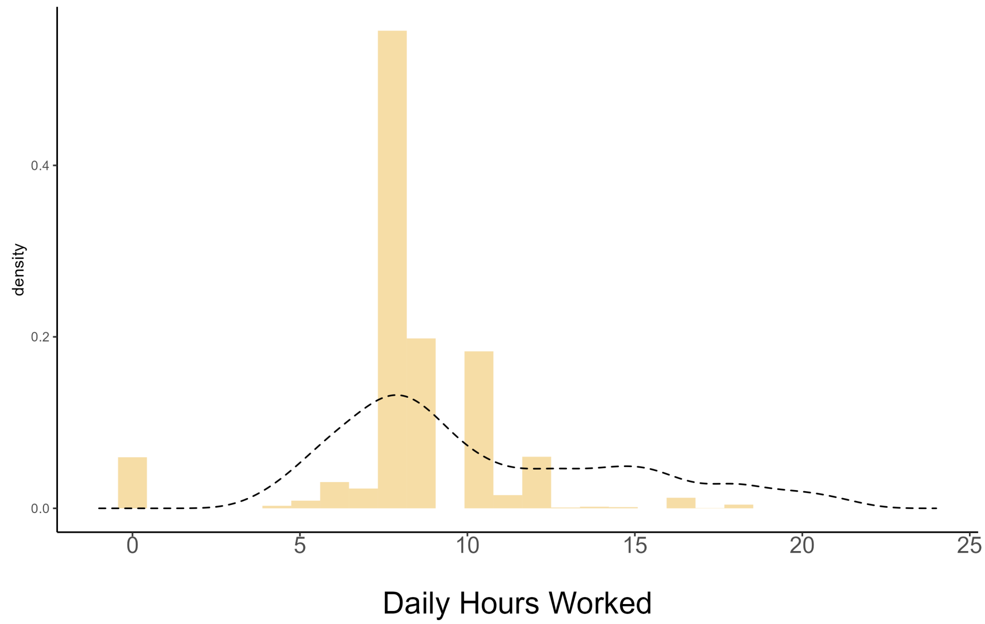
Robustness checks Table 4 displays a robustness analysis of the main results to alternative specifications. Column (1) replaces month fixed effect with week fixed effects in the baseline specification. Controlling for seasonality at a finer level provides similar results. Column (2) replaces the outcome variable hours worked in levels from equation 1 for its log-transformed version. Now the coefficient in column (2) represents an elasticity. A one percent increase in congestion increases hours worked by 0.12 percent. If congestion doubles in a day, then hours worked increase by 12 percent. This is equivalent to an increase in one hour considering that the average individual works around ten hours as reported in Table 2. Column (3) considers all sectors in the economy as opposed to only the manufacturing and office workers. I restrict the sample to manufacturing and office workers in the baseline model to reduce measurement error in the outcome variable. In several economic sectors such as retail and services (e.g. leisure, health, and education), it is likely to confound workers with clients. However, column (3) shows that results are robust to include all economic sectors. Additional robustness analyses are performed in Table B.1 in the appendix.

5.2 Using labor household surveys

This section presents results replacing the outcome variable in Equation 1 by self-reported hours worked from household labor surveys (ENOE). ENOE is the main labor market household survey in Mexico providing monthly and quarterly information. Information about individuals aged 15 years or more is collected by the National Statistics Office (INEGI) on a continuous basis throughout the year. It has a rotating panel design where one household can be followed for five consecutive quarters. The quarterly sample size is around 126,000 housing units. ENOE is representative of the country and cities such as Mexico City.

Figure 3 compares the distribution of hours worked between the official household labor survey (ENOE) and the smartphone data. The sample consists of manufacturing or office workers in firms with 50 or more employees, and the first week of January and the last week of December are excluded, as well as Saturdays and Sundays. Both distributions seem to represent the same patterns of work time. However, I performed a Kolmogorov-Smirnov test and rejected the null of equality of distributions. Hours worked from ENOE is relatively

Figure 3: Distribution of daily hours worked: ENOE vs Smartphone



Notes: The figure depicts self-reported daily hours worked from the labor household survey ENOE (in yellow) and daily hours spent at workplace from the smartphone data (dashed line) both conditional on being manufacturing or office workers in firms with 50 or more employees. The first week of January and the last week of December are excluded, as well as Saturdays and Sundays.

highly concentrated around the 8 hours compared to the distribution from the smartphone data. ENOE reports zero hours worked which contrasts with the smartphone data where all individuals work a positive number of hours by design. These differences may suggest the presence of measurement error in hours worked from the smartphone data. Hence, I replace the smartphone data with self-reported hours worked from ENOE.

Table 5 shows that results are similar after replacing the smartphone data with self-reported hours worked from household labor surveys (ENOE). Column (1) contains the baseline results. Column (2) shows estimated coefficients replacing the outcome variable from Equation 1 for daily hours worked reported in ENOE. Another key difference between columns (1) and (2) is where traffic congestion is measured. In column (1), traffic congestion is measured in the district where individuals work. In column (2), it is measured instead at the district where they work. This is relevant since 41.6% of individuals in Mexico City work in a district different than their residential district in Mexico City according to the intercensal survey in 2015. The estimated coefficients in both columns look alike in terms of magnitude and statistical significance. This result reduces measurement error concerns in hours worked using the smartphone data.

Table 5: The effect of traffic congestion on hours worked using smartphone and labor survey data

	Dependent variable: Hours worked	
	(1)	(2)
Traffic Congestion (log)	1.277**	1.018***
	(0.518)	(0.386)
Method	OLS	OLS
Individual FE	No	No
Labor data source	Phone data	ENOE survey
Congestion measured in:	Workplace	Residence
Observations	6,333	7,219
R-squared	0.087	0.014

Notes: Standard errors clustered at week-district level in parentheses. Statistical significance at the one, five, and ten percent levels is indicated by ***, ** and *, respectively. All regressions use the baseline model.

5.3 Rush hour and bottleneck model

In this section, I explore whether the main results are consistent with the bottleneck model. In this model, the departure time decision is endogenous. Individuals may choose when to start their commute in response to congestion. Hence, individuals may choose to leave earlier from home to avoid the morning rush hour or delay their departure time from work to avoid the afternoon rush hour. To conduct this exploration, I estimate the baseline model breaking traffic congestion into congestion in the morning and afternoon rush hours. The morning rush hour is from 6 am to 10 am, and the afternoon rush hour is from 5 pm to 8 pm.

Table 6 shows that the positive effect of traffic congestion on hours worked is driven by traffic congestion in the afternoon rush hour. In column (1), the outcome variable is daily number of hours spent at work constructed based on the smartphone data. We can observe a positive and statistically significant effect of the traffic congestion during the afternoon rush hour on hours worked, but a negative and not statistically significant estimated coefficient for traffic congestion during the morning rush hour.

Table 6: The effect of rush hour traffic congestion on hours worked, arrival time to work, and departure time from work

	Dependent variable:		
	Hours worked	Arrival time	Departure time
	(1)	(2)	(3)
Traffic congestion (log)			
AM rush hour	-0.274	-0.189	-0.462
(6-10am)	(0.572)	(0.297)	(0.380)
PM rush hour	1.564***	-0.407	1.157***
(5-8pm)	(0.536)	(0.254)	(0.378)
Observations	6,243	6,243	6,243
R-squared	0.089	0.101	0.041

Notes: Standard errors clustered at week-district level in parentheses. Statistical significance at the one, five, and ten percent levels is indicated by ***, **, and *, respectively. All regressions use the baseline model and smartphone data.

Columns (2) and (3), in Table 6, report that individuals are spending more time at work due to delaying their departure time from work. In column (2), the outcome variable is the arrival time to work. In column (3), the outcome variable is the departure time from work. Both columns use smartphone data. Results suggest that traffic congestion has a negative effect on the time workers arrive to work. However, the estimated coefficients are not statistically significant. Instead, traffic congestion, particularly during the afternoon rush hour, increases the time individuals depart from their jobs. The estimated coefficient is statistically significant at the 1 percent level.

These results are consistent with the bottleneck model. I find evidence that workers are spending more time at work because they are delaying the time they leave work to avoid the congestion in the afternoon rush hour.

These results may also suggest the presence of labor market frictions that prevent workers from arriving late to work or departing earlier. Hence, a potential mechanism that workers are using to respond to traffic congestion is departing later from work and, therefore, staying longer hours.

5.4 Mitigation and adaptation

Table 7 shows that workers do not mitigate the effect of traffic congestion through intertemporal labor substitution. One way to investigate this is to regress weekly hours worked using the smartphone data on weekly traffic congestion. Column (1) shows the estimated coefficient using OLS exploiting cross-sectional variation across weeks and districts. Evidence of intertemporal labor substitution would be supported with an estimated coefficient close to zero. This would support the idea that the extra time workers stay longer one day is then compensated another day by staying less time at work. However, I do not find evidence of that. I find a positive and statistically significant effect of weekly traffic congestion on weekly hours worked. The size of the coefficients suggests a cumulative effect of the single-day effect across business days. This result also goes in line with the presence of labor rigidities stated above. In a context where there are frictions to leave earlier from work, then it is unlikely that we can observe intertemporal substitution of time allocated to work.

Table 7 also shows that workers adapt to traffic congestion. Columns (2) and (3) use the

Table 7: Mitigation and adaptation to traffic congestion

	Hours worked (weekly)	Hours worked (daily) High-congestion districts	Hours worked (daily) Low-congestion districts
	(1)	(2)	(3)
Traffic congestion (weekly, log)	6.446***		
	(1.778)		
Traffic congestion (daily, log)		-0.525	2.804***
		(1.382)	(0.650)
Observations	2,671	3,508	2,825
R-squared	0.074	0.058	0.150

Notes: Standard errors clustered at week-district level in parentheses. Statistical significance at the one, five, and ten percent levels is indicated by ***, **, and *, respectively. All regressions use the baseline model and smartphone data. High-congested districts have traffic congestion above the median for all of Mexico City.

smartphone data and variables as in our baseline model. The only difference in Table 7 is that I separate districts by their intensity of traffic congestion. In column (2), I consider only the sub-sample of individuals working in high-congested districts. In column (3), I include only individuals working in low-congested districts. A district is highly congested if traffic congestion in its jurisdiction is above the median traffic congestion in the entire Mexico City. Results indicate that traffic congestion has no effect on hours worked for individuals working in high-congested areas. On the other hand, traffic congestion has a positive and statistically significant effect on hours worked for individuals working in low-congested areas. I interpret these results as evidence of adaptation. Individuals working in high-congested areas are less sensitive to shocks in traffic congestion. Instead, individuals working in low congested areas are more affected by shocks in traffic congestion in terms of the increase in work time.

Table 8: The effect of traffic congestion on labor income (ENOE)

	Labor income (monthly, log)		
	(1)	(2)	(3)
Traffic congestion (monthly, log)	-0.506**	-1.689	-1.389**
	-0.217	-1.85	-0.688
Method	OLS	OLS	2SLS
Fixed effects	Month	Month, individual	Month
Observations	684	684	684
R-squared	0.035	0.982	-0.015

Notes: Standard errors clustered at month-district level in parentheses. Statistical significance at the one, five, and ten percent levels is indicated by ***, **, and *, respectively. Traffic congestion is measured as the inverse of the daily average speed at the district level. All regressions include monthly weather (temperature, precipitation, humidity), monthly daylight hours, month fixed-effects, and exclude income from the top 1% earners. In column (3), the Kleibergen-Paap rk Wald F-statistic is 10.77 and the first-stage coefficient is 0.0001 and statistically significant at the one percent level.

5.5 Labor income

Table 8 reports the effect of traffic congestion on labor income. The outcome variable, monthly self-reported labor income, is measured in logs and it comes from ENOE household surveys. Traffic congestion is aggregated at the monthly level, and it is also measured in logs. Estimated coefficients reported in the table are elasticities. The number of observations decreases due to the monthly aggregation, but also because a large proportion does not provide income information in the survey. Column (1) shows OLS estimates pooling all individuals and controlling for month fixed effects. Column (2) is similar to column (1) but exploits the fact that some individuals were interviewed in multiple months adding individual fixed effects to the control variables. Column (3) presents results using 2SLS and instrumenting traffic congestion with monthly road accidents.

If individuals are staying longer hours at work, are they getting paid more as well? I do not find evidence of this. All results from Table 8 suggest, if anything, that traffic congestion is not increasing labor income. Hence, individuals are staying longer at work, but suggestive evidence implies that they are not earning more income. Note that one potential explanation is that workers are rewarded in the future. However, this cannot be explored in this study due to data limitations.

6 Discussion

Road accidents Figure A.7 in the appendix suggests that workers react to shifters of congestion, namely road accidents, that happen before the afternoon rush hour. It displays the coefficient estimates and 90% confidence intervals of reduced form regressions. Estimated coefficients are reported in Table B.3. Panel (a) in Figure A.7 shows the results of regressing the number of accidents on the number of hours spent at work. I split the number of accidents into three categories: (i) before the morning rush hour, (ii) during the morning rush hour and before the afternoon rush hour, and (iii) during and after the afternoon rush hour. Panel (b) shows similar results using departure time as the outcome variable instead. We can observe that the estimated coefficient associated with the morning rush hour and after the afternoon rush hour is not statistically significant. This suggests that workers react to accidents that happen before the afternoon rush hour. However, the magnitude of all coefficients is similar. Nonetheless, another piece of evidence suggests that workers can react to accidents during the day. In 2019, the navigation company Waze was already operating in the city and had two million active users monthly.

Labor supply To what extent is the time spent at work measured using smartphone data capture labor supply? Time at work does not always equal work time. For example, individuals may stay one hour longer at work, but they may be partially working that hour or not working at all. Moreover, individuals may leave work and continue working at home.

Hence, in principle, it seems that "hours spent at the workplace" is far from capturing labor supply. Moreover, we observe only an equilibrium outcome of the supply and demand in the labor market.

Table 5 suggests the "hours spent at the workplace" are not far from capturing hours worked. Columns (1) and (2) present the main results using smartphone data and ENOE household surveys. We can observe that both estimated coefficients are similar in magnitude, and the coefficient using ENOE is slightly lower. We can take this as evidence that hours stay at work slightly overestimates hours worked. Workers may be staying extra time at work, and part of that time is allocated to work. Figure A.6 may also indicate that hours spent at work are related to hours worked. If output per labor hour is increasing with congestion, and hours worked are increasing with congestion, this means that output should also be increasing in congestion. If people were only shirking at work, then we would not observe an increase in output. If the extra hours at work were artificial, then we should not see an increase in productivity.

In our short-run setting, it is unlikely that labor demand factors play a role. Therefore, we could use changes in hours worked to approximate shifts in labor supply. However, differences between the number of hours spent at work and hours worked reported above suggest being cautious and interpreting the results as changes in work-time allocation instead of labor supply.

Labor productivity If individuals are staying longer hours at work, are they producing more output per labor hour? Figure A.6 presents suggestive evidence that this may be the case. It displays the correlation between monthly traffic congestion and monthly labor productivity. Labor productivity is calculated as total output value divided by total hours worked using information from manufacturing firm surveys (EMIM) in 2019. However, this evidence should be taken with caution given that we cannot find a causal relationship between traffic congestion and labor productivity from Figure A.6. The positive correlation may be explained by other factors such as seasonality. Labor productivity and traffic congestion may be higher in particular months of the year (*e.g.* December). EMIM collects more granular firm-level data. Unfortunately, only aggregate numbers used in Figure A.6 are publicly available. This exercise can be replicated in the future when access to the fully EMIM microdata is provided.

Welfare and Inequality I find evidence that workers are staying more hours at work due to congestion. However, suggestive evidence implies that workers are not earning more for this extra hour. Regardless of whether individuals are conducting actual work or shirking during that extra hour, there is evidence that this time is not being remunerated. This extra hour represents 10 percent of the average shift reported in Table 2. Hence, traffic congestion may be reducing the welfare that individuals obtain from participating in the labor market by 10 percent. However, we first need to test whether this extra hour at work is not rewarded in the medium or long term. On the other hand, traffic congestion may be a shifter of inequality. Workers are staying more time at work without being compensated for it, but firms are enjoying more labor productivity.

7 Conclusion

Traffic congestion is a major and yet unsolved concern in most cities in the world. This paper studies the effect of traffic congestion on work-time allocation, a previously unquantified externality. I exploit highly granular smartphone data to measure daily work-time allocation, including arrival and departure times from work. I combine these data with daily exposure to traffic congestion measured using GPS sensors installed in circulating vehicles in Mexico

City in 2019. My identification strategy exploits within-district daily variation in traffic congestion. I complement this approach using road accidents as an exogenous shifter of traffic congestion.

The results suggest that traffic congestion increases time allocated to work. Facing twice as much traffic congestion leads to an additional hour spent at work. This finding is robust to using self-reported hours worked from household surveys, as well as to using road accidents as an instrumental variable for congestion. I find that workers stay longer largely because they delay their departure time from work to avoid traffic congestion during the afternoon rush hour. Moreover, workers seem to respond to congestion shifters (i.e, accidents) that occur before the afternoon rush hour. I do not find evidence that workers mitigate the effect of traffic congestion through intertemporal labor substitution; for example, a worker who stays longer today does not compensate by leaving work early tomorrow. I do find evidence of adaptation in the sense that individuals working in high-congestion areas are less affected than individuals working in low-congestion areas. I also find suggestive evidence that workers are not earning more even though they are staying longer hours at work, but labor productivity is increasing. However, this study has some limitations. It is focused only on the short-run effects of traffic congestion on work-time allocation. The findings are also silent on modes of transportation, which may be another channel that individuals use to avoid traffic congestion.

Staying longer hours at work has detrimental effects on well-being, with wide implications for human health, productivity, and the quality of leisure time (e.g., time spent on hobbies and with those we love). By prompting people to stay longer at work, traffic congestion may be mitigating the substantial benefits that cities offer to workers. In addition, rescheduling the timing of activities has important welfare effects (Small, 1982). Time is the ultimate finite resource, which puts time allocation at the heart of the human experience. In our setting, doubled traffic congestion may reduce the welfare that individuals obtain from participating in the labor market by 10 percent. Hence, the externality of traffic congestion on work-time allocation likely has major impacts on broader well-being.

This study is an example of the use of smartphone data to study human behavior. Smartphone data have huge potential. More research is needed to investigate the representativeness of these data to the whole population, their statistical reliability, and possible synergies with household- and firm-level surveys to learn more and better our behavioral patterns and new developments in the labor markets, such as the great resignation. Further research is also needed to understand long-term effects and to explore the role of modes of transportation. Likewise, future research should address the effect of traffic congestion on productivity across all sectors of the economy.

References

Akbar, Prottoy, Victor Couture, Gilles Duranton, and Adam Storeygard, "Mobility and Congestion in World Cities: Evidence from Google Maps," 2020, *work in progress.*

Allen, Treb and Costas Arkolakis, "The welfare effects of transportation infrastructure improvements," *The Review of Economic Studies*, forthcoming.

Arnott, Richard and Kenneth Small, "The economics of traffic congestion," *American scientist*, 1994, *82* (5), 446–455.

_ **, André De Palma, and Robin Lindsey**, "Economics of a bottleneck," *Journal of urban economics*, 1990, *27* (1), 111–130.

_ **, Andre De Palma, and Robin Lindsey**, "A structural model of peak-period congestion: A traffic bottleneck with elastic demand," *The American Economic Review*, 1993, pp. 161–179.

Beland, Louis-Philippe and Daniel A Brent, "Traffic and crime," *Journal of Public Economics*, 2018, *160*, 96–116.

Black, Dan A, Natalia Kolesnikova, and Lowell J Taylor, "Why do so few women work in New York (and so many in Minneapolis)? Labor supply of married women across US cities," *Journal of Urban Economics*, 2014, *79*, 59–71.

Calatayud, Agustina, Santiago Sanchez Gonzalez, Felipe Bedoya Maya, Francisca Giraldez Zuniga, and Jose Maria Marquez, *Congestion urbana en America Latina y el Caribe: Caracteristicas, costos y mitigacion*, IDB, 2021.

Couture, Victor, Jonathan Dingel, Allison Green, Jessie Handbury, and Kevin Williams, "Measuring movement and social contact with smartphone data: a real-time application to COVID-19," *Journal of Urban Economics*, 2022, *127.*

Fu, Shihe and Brian Viard, "Commute costs and labor supply: evidence from a satellite campus," *Journal of Economic Geography*, 2019, *19* (3), 723–752.

Gutiérrez-i-Puigarnau, Eva and Jos N van Ommeren, "Labour supply and commuting," *Journal of Urban Economics*, 2010, *68* (1), 82–89.

_ **and** _ , "Commuting and labour supply revisited," *Urban Studies*, 2015, *52* (14), 2551–2563.

Hanna, Rema, Gabriel Kreindler, and Benjamin A. Olken, "Citywide effects of high-occupancy vehicle restrictions: Evidence from "three-in-one" in Jakarta," *Science*, July 2017, *357* (6346), 89–93.

INRIX, "Scorecard, Global Traffic," 2019.

Kim, Jinwon, "Estimating the social cost of congestion using the bottleneck model," *Economics of Transportation*, 2019, *19*, 100119.

Kreindler, Gabriel, "Peak-hour road congestion pricing: Experimental evidence and equilibrium implications," *Working paper*, 2022.

Noland, Robert and Kenneth A Small, "Travel-time uncertainty, departure time choice, and the cost of morning commutes," *Transportation research record*, 1995, *1493*, 150–158.

Small, Kenneth A, "The scheduling of consumer activities: work trips," *The American Economic Review*, 1982, *72* (3), 467–479.

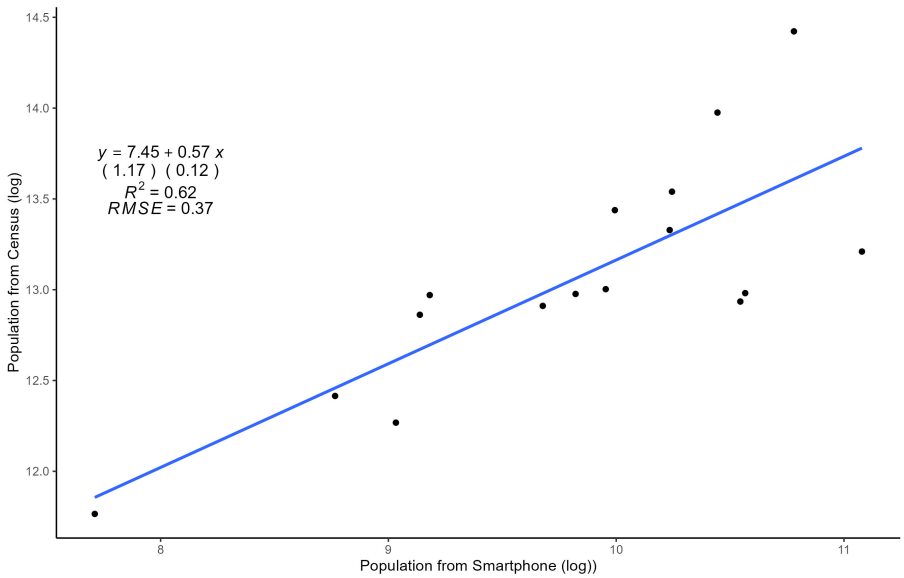
Viard, Brian and Shihe Fu, "The effect of Beijing's driving restrictions on pollution and economic activity," *Journal of Public Economics*, 2015, *125*, 98–115.

Vickrey, William, "Congestion theory and transport investment," *The American Economic Review*, 1969, *59* (2), 251–260.

APPENDIX

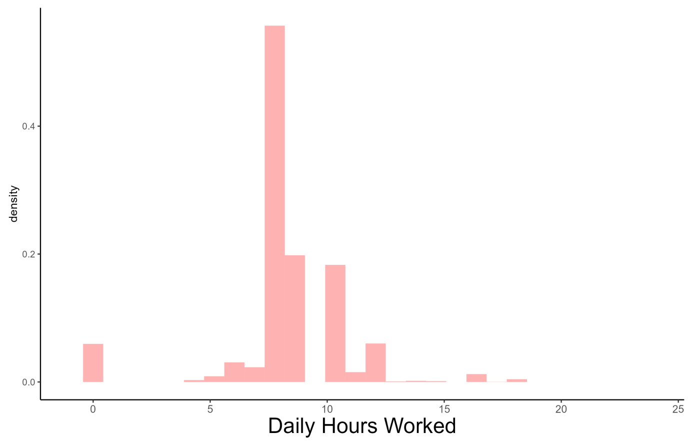
A Additional figures

Figure A.1: Correlation of population size from smartphone and census data

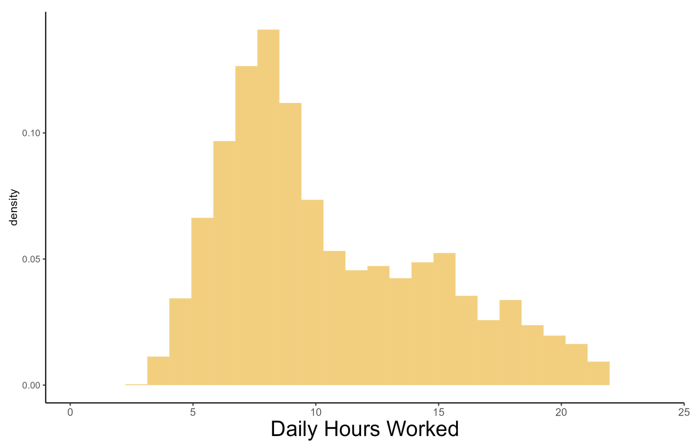


Notes: The figure depicts a comparison of the total population in Mexico City according to the Census 2020 (vertical axis) with the total number of smartphone devices in 2019. Each dot is one of the 16 districts. Linear regression line in blue.

Figure A.2: Distribution of daily hours worked: ENOE vs. Smartphone



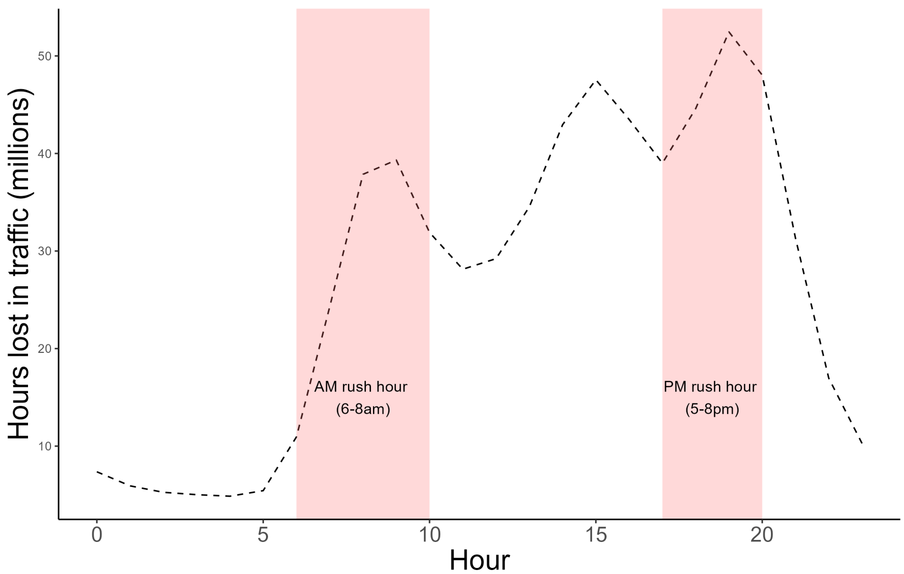
(a) Labor survey (ENOE)



(b) Smartphone

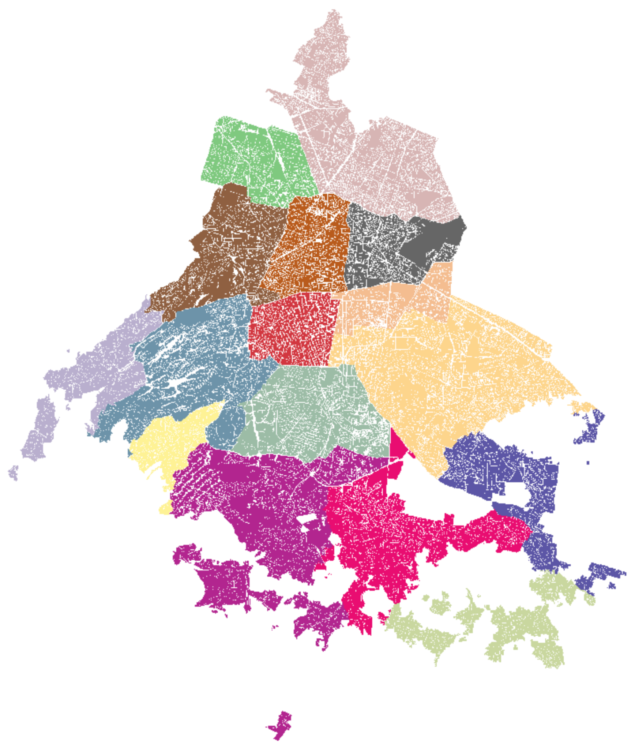
Notes: The figure depicts self-reported daily hours worked from the labor household survey ENOE (panel a) and daily hours spent at the workplace from the smartphone data (panel b) both conditional on being manufacturing or office workers in firms with 50 or more employees. The first week of January and the last week of December are excluded, as well as Saturdays and Sundays.

Figure A.3: Distribution of traffic congestion using data from Waze



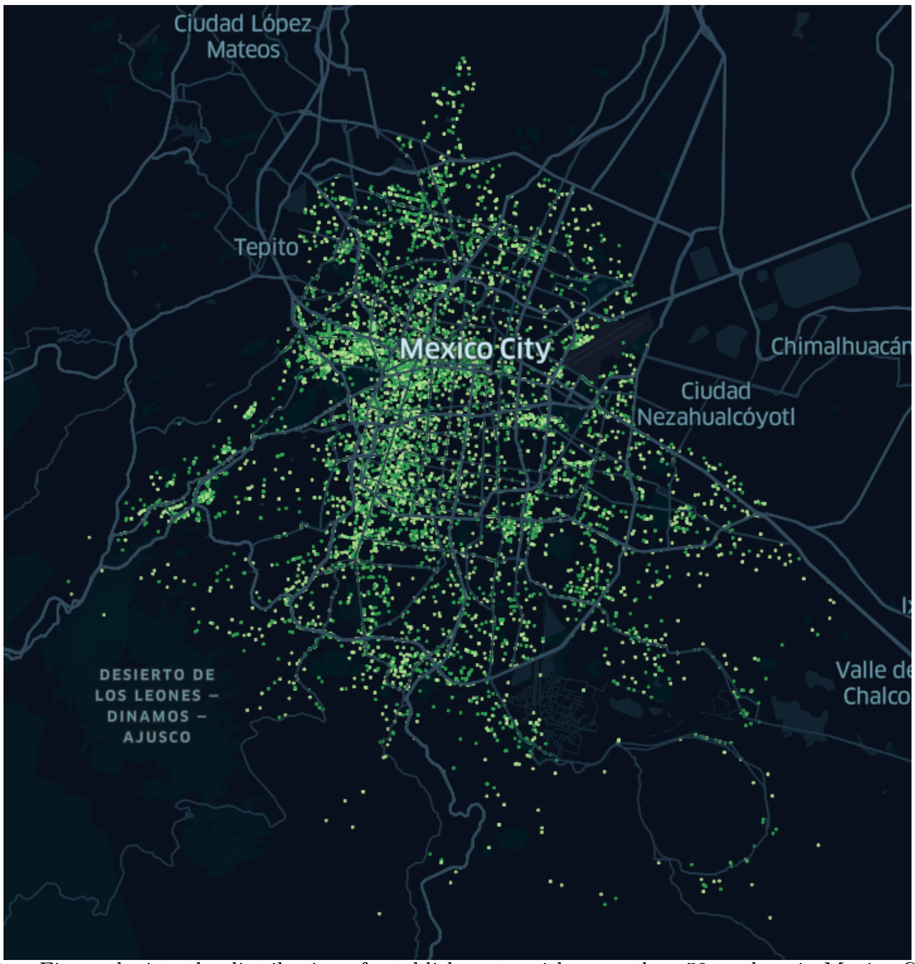
Notes: The figure depicts the distribution of traffic congestion per hour using data from Waze for Mexico City in 2019. Data provided by Calatayud et al. (2021). Morning (6 am-10 am) and afternoon (5 pm-8 pm) rush hours are highlighted in yellow.

Figure A.4: Map of Mexico City



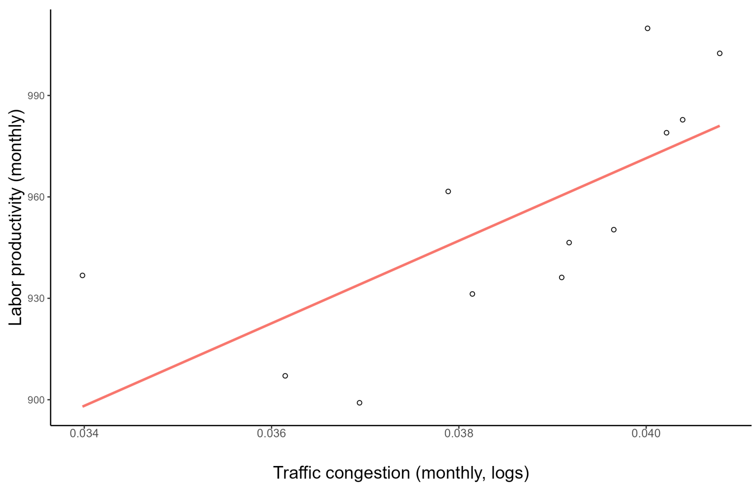
Notes: Figure depicts the map of Mexico City. Each color represents one of the 16 districts.

Figure A.5: Distribution of establishments



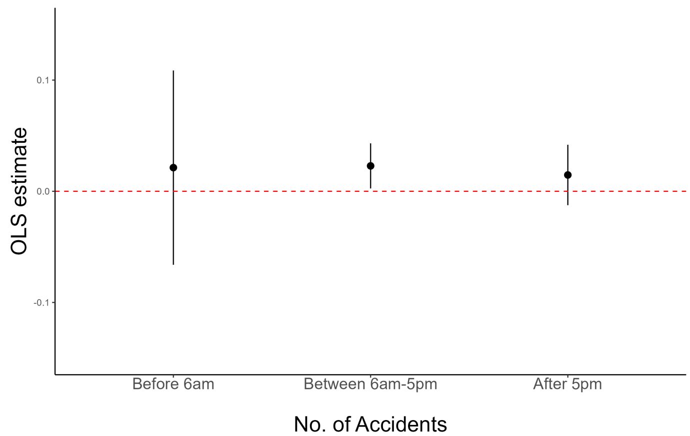
Notes: Figure depicts the distribution of establishments with more than 50 workers in Mexico City. The location of establishments is provided by the National Statistical Directory of Economic Units (DENUE) 2019.

Figure A.6: Correlation between traffic congestion and labor productivity

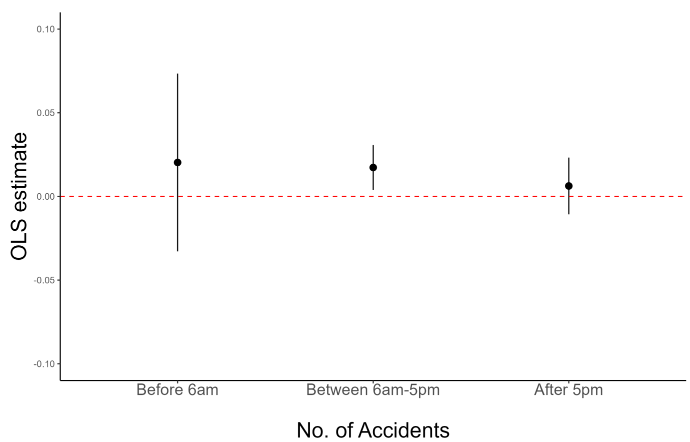


Notes: Figure depicts the correlation between monthly traffic congestion and monthly labor productivity (total output value divided by total hours worked). Each dot represents information for a month in 2019. Labor productivity comes from monthly manufacturing firm surveys (EMIM). Linear regression line in red.

Figure A.7: The effect of road accidents on hours worked and departure time from work



(a) Hours worked



(b) Departure time

Notes: The figure depicts see OLS estimates and 90% confidence intervals. See regressions output in Table B.3.

B Additional tables

Table B.1: Additional robustness analysis for main results

	Dependent variable: Hours worked	
	(1)	(2)
Traffic Congestion (log)	0.756*	3.483
	(0.435)	(6.579)
Change in specification	Distance >4km	2SLS + individual FE
Observations	1,159	6,242
R-squared	0.035	0.022

Notes: Standard errors clustered at week-district level in parentheses. Statistical significance at the one, five, and ten percent levels is indicated by ***, **, and *, respectively. All regressions use the baseline model and smartphone data.

Table B.2: Additional robustness analysis for mitigation and adaptation

	Hours worked (weekly)		Arrival time		Departure time	
	High	Low	High	Low	High	Low
	(1)	(2)	(3)	(4)	(5)	(6)
Congestion (weekly, log)	-3.694	10.296***				
	(5.685)	(2.258)				
Congestion (log)			0.977	-1.740***	0.452	1.065***
			(0.676)	(0.344)	(0.899)	(0.406)
Observations	1,474	1,197	3,508	2,825	3,508	2,825
R-squared	0.059	0.108	0.070	0.167	0.026	0.077

Notes: Standard errors clustered at week-district level in parentheses. Statistical significance at the one, five, and ten percent levels is indicated by ***, **, and *, respectively. All regressions use the baseline model and smartphone data.

Table B.3: The effect of road accidents on hours worked and departure time from work

	Hours worked	Arrival Time	Departure time	Congestion (log)
	(1)	(2)	(3)	(4)
No. of accidents				
before AM rush hour	0.021	-0.001	0.020	0.010***
(before 6am)	(0.053)	(0.029)	(0.032)	(0.003)
during and after AM rush hour	0.023*	-0.006	0.017**	0.010***
(6am-5pm)	(0.012)	(0.006)	(0.008)	(0.001)
during and after PM rush hour	0.015	-0.008	0.006	0.008***
(5pm-midnight)	(0.016)	(0.009)	(0.010)	(0.001)
Observations	6,709	6,709	6,709	6,333
R-squared	0.088	0.104	0.038	0.433

Notes: Standard errors clustered at week-district level in parentheses. Statistical significance at the one, five, and ten percent levels is indicated by ***, **, and *, respectively. All regressions use the baseline model and smartphone data.

C Data appendix

C.1 Supplemental Data

Weather data I use hourly data about temperature, precipitation, and relative humidity recorded by ground stations in Mexico City. The data was provided by the national meteorological agency (CONAGUA).

Daylight hours I calculate daylight hours for each day of the working week by taking the difference between the sunset and sunshine time at each district location. Sunset and sunshine times were calculated using astronomical algorithms taking as input the date each individual worked and the geographic coordinates of the districts where they work. I use the package suncalc in R to perform these calculations.

ENOE It is the main labor market household survey in Mexico providing monthly and quarterly information. It is conducted by the National Statistics Office (INEGI) and it collects information from individuals aged 15 years or more continuously every week from Monday to Sunday throughout the year. It has a rotating panel design where every five quarters 20% of the sample is replaced. The quarterly sample size is around 126,000 housing units. It is representative at the national level, and also at the level of cities such as Mexico City.

Accidents I use administrative records about road accidents collected by the centralized emergency center in Mexico City under the supervision of the local government in Mexico City. The administrative records contain information about the location (latitude and longitude coordinates), the date, and the time of the road accident. It also provides information about the type of accidents and whether the accident involved victims among other details.

EMIM It is a monthly establishment-level survey representative of the manufacturing sector in Mexico. All establishments report information about the number of employees, earnings, output value, and sales among other economic characteristics. They report this information every month of the year. The sample size for 2019 is 10,447.

C.2 Smartphone and Traffic Congestion Data Providers

Quadrant It is a global leader in mobile location data, POI data, and corresponding compliance services. Quadrant provides anonymized location data solutions that are fit for purpose, authentic, easy to use, and simple to organize. They offer data for almost all countries in the world, with hundreds of millions of unique devices and tens of billions of events per month, allowing our clients to perform location analyses, derive location-based intelligence, and make well-informed business decisions. Their data is gathered directly from first-party opt-in mobile devices through a server-to-server integration with trusted publisher partners, delivering genuine and reliable raw GPS data, unlike other location data sources. Their consent management platform, QCMP, ensures that their data is compliant with applicable consent and opt-out provisions of data privacy laws governing the collection and use of location data. More information about the company can be found here: https://www.quadrant.io/

Dat's Why It is a leading mobility intelligence platform with +70M smartphones, vehicles and sensors collecting in real time +40B data points annually in Latin America. Using its real-time Big Data network of Geobehavior, the largest in Mexico, monitors various traffic parameters and creates smart mobility solutions and analytics. More information about the company can be found here: https://datswhy.com/

C.3 Identify Work Location

- Step 1: Initial sample selection
  - Using SQL in Amazon AWS, select those devices with error in location accuracy < 50m, which is approximately half of a street block.
  - Select those devices observed more than seven days in a month to avoid tourists or sporadic users.
  - Select those devices observed at least twice a day to potentially know the arrival and departure time from a location (e.g. home or work).
- Step 2: Location of establishments
  - Use geocoded establishment-level data from DENUE that provides the latitude and longitude coordinates for each establishment. Use the location of the establishment as a point of interest (POI).
  - Using Python, draw a circular geofence of radius 50m around the POI.
  - Using Python, convert the POI with circular geofence into geohash grids (precision 8, +- 20m).
- Step 3: Combine the smartphone data with the establishment-level data using the geohash grids.
- Step 4: Use algorithm inspired in Couture et al. (2022)
  - Use pings observed from Monday to Friday and between 9 am and 5 pm. These are the days and times for regular daytime work. Note that the sample here is restricted only for the purposes of finding the work location. For the statistical analysis, we use all pings observed during the entire day.

- Calculate how much time each device spends at workplaces (i.e circular geofences around the POI). As a result, one device may have more than one candidate as a potential work location given that individuals move around.
- Assign the device to the workplace venue with the longest duration.
- If the duration is 0, then the workplace is the work location with the most daytime visits.
- Finally, the device needs to visit this workplace venue at least 3 times a week.

C.4 Identify Home Location

- Step 1: Initial sample selection
  - Using SQL in Amazon AWS, select those devices with error in location accuracy < 50m, which is approximately half of a street block.
  - Select those devices observed more than seven days in a month to avoid tourists or sporadic users.
  - Select those devices observed at least twice a day to potentially know the arrival and departure time from a location (e.g. home or work).
- Step 2: Location of residential areas
  - Using Python, convert the block from the National Geostatistical Framework (MGN) into polygon geofences.
  - Use Census to identify which blocks contain residential places.
  - Using Python, convert the polygon geofence into geohash grids (precision 8, +-20m).

- Step 3: Combine the smartphone data with the residential data using the geohash grids.

- Step 4: Use algorithm inspired in Couture et al. (2022)

  - Select blocks with inhabited private homes regardless of whether there are establishments as well in the block.

  - Use pings observed between 9 pm and 5 am. Presumably, these are the times when daytime workers are at home. Note that the sample here is restricted only for the purposes of finding the work location. For the statistical analysis, we use all pings observed during the entire day.

  - Calculate how much time each device spends at potential home locations. As a result, one device may have more than one candidate as a potential home location given that individuals move around.

  - Assign the device to the residential venue with the longest duration.

  - If the duration is 0, then the residential place is the home location with the most daytime visits.

  - Finally, the device needs to visit this residential venue at least 3 times a week.