Policy Research Working Paper 9308

Big Data in Transportation

An Economics Perspective

Harris Selod
Souleymane Soumahoro

WORLD BANK GROUP
Development Economics
Development Research Group
June 2020

Policy Research Working Paper 9308

Abstract

This paper reviews the emerging big data literature applied to urban transportation issues from the perspective of economic research. It provides a typology of big data sources relevant to transportation analyses and describes how these data can be used to measure mobility, associated externalities, and welfare impacts. As an application, it showcases the use of daily traffic conditions data in various developed and developing country cities to estimate the causal impact of stay-at-home orders during the Covid-19 pandemic on traffic congestion in Bogotá, New Dehli, New York, and Paris. In light of the advances in big data analytics, the paper concludes with a discussion on policy opportunities and challenges.

This paper is a product of the Development Research Group, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at hselod@worldbank.org and ssoumahoro@worldbank.org.

Big Data in Transportation: An Economics Perspective

Harris Selod and Souleymane Soumahoro*

Keywords: Traffic, Congestion, Covid-19, Bayesian Structural Time-Series.

JEL Classification Numbers: R41, C55, and D62.

*The World Bank 1818 H Street, NW Washington, DC 20433, USA. The authors are grateful to the big data analytics firm QuantCube (www.q3-technology.com/), and especially to its CEO, Thanh-Long Huynh, and its Lead Data Scientist, Alice Froidevaux, for collaboration on the empirical application in Section IV-B. They also thank Gilles Duranton, Alejandro Molnar, and participants to the workshop "Leveraging new data for better urban management and policies" and the 6th World Bank/GWU/IDB/IGC Urbanization and Poverty Reduction Research Conference (Washington, D.C., September 9-10, 2019) for useful comments. Funding from the Knowledge for Change Program is gratefully acknowledged.

I. Introduction

The paucity of economically relevant city-level data can be a significant hindrance to research on urbanization, especially in the context of less developed economies. Even in the circumstances where the desired data are available, users of such data (researchers and policy makers) may have to rely on small-scale surveys and coarse statistics. Such data constraints make it difficult to measure important urban phenomena. They also reduce the scope for economic analyses and policy design.

Over the past decade, cities have increasingly become a reservoir for billions of digital footprints that can be used to measure and assess mobility patterns with high time frequency and high spatial resolution. Combined with new methods of analysis, including artificial intelligence approaches (such as machine learning), this has led to a "big data" revolution in transportation. Researchers, experts and policy makers are increasingly exploring the wealth of mobility and city data from various sources—including digital repositories from the public and private sectors, remote and in-situ sensors, mobile phones, among others—to examine traffic flow, urban congestion and its economic, social, and environmental externalities (e.g., pollution, health hazards, or accidents). In the present review, we seek to examine how new data sources have stimulated innovative approaches to measure and predict road traffic flows, and to assess the economic and social costs of traffic externalities.

Our review of this emerging literature indicates that big data has a strong potential to more accurately measure social costs in a variety of contexts, including in developing countries where traditional transport data are lacking and where data "leapfrogging" is occurring.[1] In fact, unlike traditional household travel surveys alone, new sources of data such as Google Maps, GPS traces and sensor networks can be used to track individual travel demand patterns and estimate private and social trip costs, two important ingredients of a welfare analysis. They also make it possible to design policies

[1] For a survey of big data uses in an urban context more broadly, see Glaeser et al. (2018).

(e.g., Pigouvian pricing of externalities) and characterize their impacts. For example, call detail records (CDR) from mobile phones have become a compelling and cost-effective way to measure traffic flows and, combined with other data, can be used to assess the links between congestion and other social, economic, and environmental outcomes (e.g., emissions measured from remote sensing or in situ devices).

These improvements in outcome measurements have generated new venues for research, promising policy opportunities (e.g., the implementation of real-time traffic management systems) and, more generally, stimulated evidence-based policy designs. There are, however, multifaceted challenges that are inherent to the collection and use of big data, including technical barriers, privacy concerns, and political risks. It is also worth mentioning that although big data makes it possible to have better measurements of phenomena, track wider social and economic outcomes, and improve predictive modeling, it does not automatically establish causal relationships. There is thus a space for economic analysis to make use of big data in order to infer causality, in particular through structural modeling or quasi-experimental designs.

The paper is structured as follows. In Section II, we present a taxonomy of new traffic data sources and their potential for improving transportation studies. In Section III, we examine the economics of traffic externalities in the age of big data, reviewing recent empirical efforts to quantify welfare, productivity, and environmental costs. Section IV acknowledges the potential of big data analytics to improve our understanding of these externalities—through better measurement and accurate predictions—and, as a proof of concept, proposes a simple application that measures the causal link of stay-at-home orders on urban traffic congestion. Section V concludes with a discussion on policy opportunities and challenges.

II. Taxonomy of New Traffic Data Sources

It is useful to start by presenting a taxonomy of the different data sources used in big data analyses. We distinguish (i) direct physical sensing (in situ or remotely) from (ii) social media sources (so called human and social sensing) and (iii) urban sensing (through information generated by transportation operators). This is presented below.

A. In-situ and Remote Sensing

The primary sources of data for traffic studies traditionally included standard data collection efforts such as household traffic or commodity flow surveys, possibly complemented with first-generation in-situ and remote sensing systems such as inductive loops, overhead radar detectors, and static video cameras, among others. Recent developments in information and communication technology (ICT) have stimulated a proliferation of new traffic data sources. These include new generations of road-side static sensors such as LiDAR (Light Detection and Ranging), microwave radars, and acoustic sensors which measure speed, noise and traffic flow on the road network. The rapid penetration of mobile phone technology has also considerably lowered the costs of collecting travel behavior data. For example, GPS, GSM, and Bluetooth are increasingly used to generate real-time and reliable traffic volume data, travel speed and time, as well as origin-destination flows (see, e.g., Mitchell, 2014). It is also possible to measure mobility patterns from anonymized data extracted from call detail records (CDR).

Recently, the literature has also acknowledged the role of unmanned aerial vehicles (UAV), including drones and unmanned High-Altitude Pseudo-Satellites (HAPS) in providing valuable data for traffic analysis (see, e.g., Barmpounakis et al., 2016; Khan et al., 2017). Unlike static traffic monitoring systems and ground vehicles, UAVs can collect high resolution mobility data at fine grained spatial and temporal scales. UAVs can also perform data collection tasks similar to manned aerial vehicles while ensuring rela-

tively higher safety, rapidity, and cost-effectiveness (Puri, 2005). They can be useful in a wide range of monitoring and planning operations, including incident response, traffic surveillance on roadways and intersections, parking lot monitoring, origin-destination flow estimation, among other applications (Coifman et al., 2004; Puri, 2005). Like call detail records from mobile phones, drones are increasingly filling data gaps inherent to scarce or infrequently collected travel surveys.

B. Human and Social Sensing

In the digital era, virtual interactions on social media generate content or data (referred to as user generated content or UGC) on real-time trip experience, traffic conditions, and travel time unreliability. Attempts to characterize travel patterns or detect traffic anomalies, for instance, are relying on smartphone-compatible platforms such as Twitter, Foursquare, Instagram, and other social media. A good example of this comes from He et al. (2013) who examine the predictive power of travelers' social media activities in San Francisco and find a negative correlation between the intensity of tweets and traffic volume. More importantly, while most existing models have a short-term forecasting horizon, these authors show that tweet-based semantics can be exploited to significantly improve longer-term traffic predictions. In the same vein, self-reported positions on Foursquare and Instagram (Ribeiro et al., 2014), and traffic complaints on Twitter (Georgiou et al., 2015) have also been shown to be highly correlated with traffic congestion.

The sources of traffic obstruction can be analyzed using social sensing data describing commuters' reactions to real-time disruptive events. To do this, a social sensing software scrutinizes a given social media platform, looking for keywords such as "traffic", "congestion", "accident", and "roadwork", and extract useful information for real-time traffic analysis. For example, Pan et al. (2013) combine human mobility data from GPS tracers with social media data from Weibo (a Twitter-like social media in China) to characterize

real-time traffic anomalies that arise from accidents, roadworks, and other disruptive events. Similar approaches have been used to develop an intelligent interface platform in Dublin (Daly et al., 2013) and a smartphone application in the United Kingdom (see Hampson et al., 2017) that provide real-time information on traffic congestion sources. In the case of the United Kingdom, a mobile application known as Twitraffic provides real-time traffic information from users' tweets. On average, Twitraffic is believed to detect traffic incidents about 7.1 minutes before they are picked up by the UK government's highway agency data (Hampson et al., 2017).

C. Urban Sensing Technology

In addition to mobile and social sensors, credit cards, smart cards in public transit, retail scanners, and digital toll systems, are increasingly being used as "urban sensors" (Blazquez and Domenech, 2018). Data generated through the urban sensors can be used to measure broad patterns of urban mobility and to some extent traffic flows and congestion.

Transponders at toll plazas can be used to measure personal and commercial vehicle flows along a road segment (Kitchin, 2014), and eventually track variations in economic activities. For example, Askitas and Zimmermann (2013) explore data from an electronic toll collection system in Germany to generate a monthly Toll Index, which they use to nowcast business cycles. Smart cards in public transit, such as SmarTrip in the Washington metropolitan area and Oyster in London, can be used to track individual travel behavior, and generate useful information on economic and social outcomes. It is possible, for example, to characterize a cardholder's commuting trajectory while generating information on his/her place of residence and employment status by combining smart cards with credit card data (Wang and Moriarty, 2018). Smart cards also allow for granular analyses of transport behavior. As an example, Tao et al. (2014) explore a large smart card database to examine the spatial-temporal dynamics of Bus Rapid Transit (BRT) sys-

tems in Brisbane, Australia. Comparing BRT-based trips with trips taken in non-BRT urban public transit networks, the authors find that BRT-based trips exhibit significant spatial heterogeneity and are markedly different from ordinary bus travel.

III. Traffic Flows and Externalities

The use of big data in economic analyses of transportation covers a variety of topics that can broadly be categorized as (i) measures of health and environmental externalities associated with congestion, (ii) economic effect of congestion (in terms of employment and productivity) and (iii) welfare costs. We sequentially review these applications, starting with welfare costs.

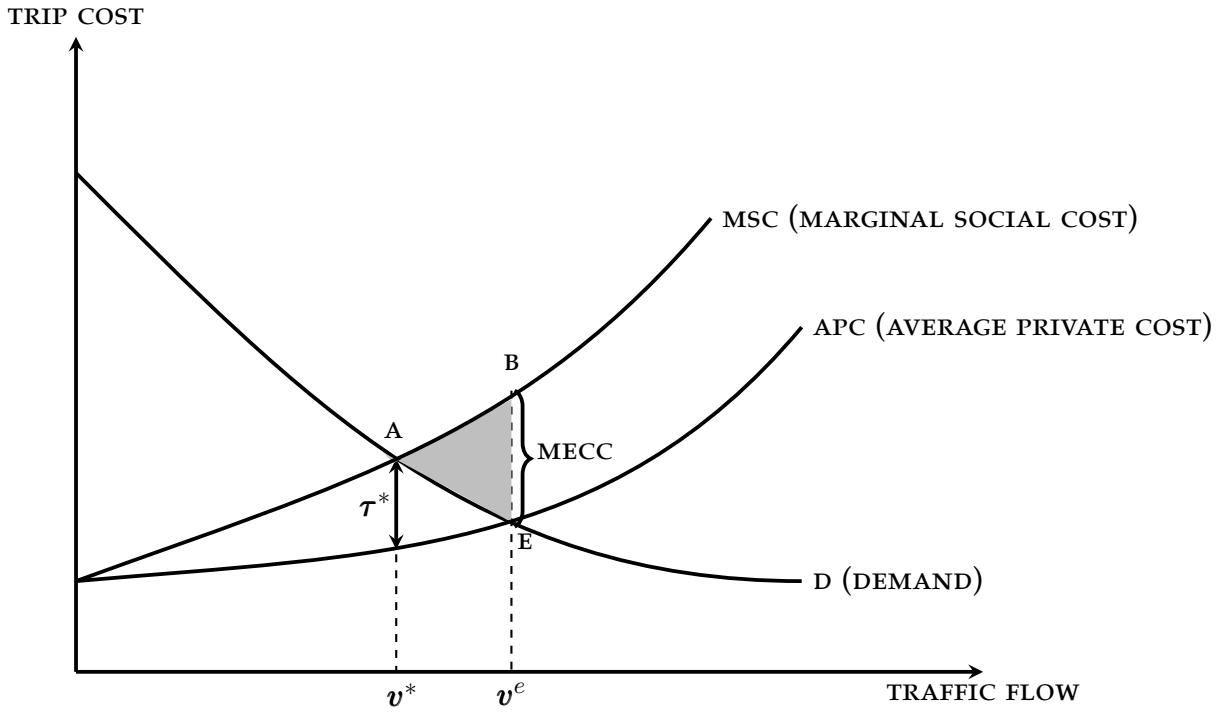
A. Welfare Costs

Traffic flows can entail several externalities (e.g., accidents, health effects, delays), which are often directly related to or exacerbated by congestion. Congestion in cities can have very large economic and social costs. In its 2015 Urban Mobility Scorecard (UMS), for instance, the Texas Transportation Institute analyzes congestion in 471 urban areas in the United States and finds that urban dwellers spent an extra 6.9 billion hours in traffic and wasted 3.1 billion gallons of fuel (Schrank et al., 2015). Converting these congestion-induced travel delays and fuel consumption into monetary values, the study concludes that the total cost of urban traffic congestion in the United States represented $US160 billion in 2014. In Cairo and Kuala Lumpur, two highly urbanized cities in the developing world, congestion costs estimated using similar approaches reach up to 3.6 and 2.2 percent of national GDP, respectively (Zachau et al., 2015). These figures mostly reflect the industry estimation standard in which the direct costs of congestion are computed as the aggregate travel delay—including time unreliability—relative to a free-flow benchmark. However, despite being widely referenced, economists are often skeptical

about the economic rationale behind these measurements (Litman, 2013). Duranton and Venables (2018) argue that the benchmark should not be the free-flow for three reasons: (i) Attaining the same observed equilibrium flow without congestion may be a physically impossible counterfactual; (ii) Free flow may not be compatible with the observed equilibrium demand for transport (which would increase given the absence of congestion); and (iii) Optimality may actually require some level of congestion. In fact, from an economic theory perspective, urban traffic congestion is better understood in terms of demand and supply (Walters, 1961; Mohring, 1976).

Figure 1 provides a graphical illustration of the standard model of congestion (which is generalizable to any type of traffic externality). The model focuses on the short run when traffic capacity is fixed. The inverse demand curve D expresses the aggregated flow or volume of travel demanded (x axis) as a function of trip cost in minutes per kilometer, i.e. the inverse of speed (y axis). It reflects users' willingness to pay for urban travel. The marginal and average cost functions on the graph express how the time costs of travel increases with traffic flow. Given congestion externalities, each marginal user increases overall congestion, so that the marginal social cost (MSC) of travel is greater than the average private cost (APC), as represented on Figure 1. The difference between the MSC and the APC—i.e., the vertical distance separating the two curves—is the marginal external congestion cost (MECC). It measures the reduced speed and related delays that each individual user entering the traffic stream imposes on all other travelers. The equilibrium volume of traffic, v^e, corresponds to the intersection of the curves D and APC, where the distance between the MSC and APC curves shows the equilibrium MECC (see Figure 1). However, the optimal traffic flow, v^*, equates the demand (curve D) and the marginal social cost (curve MSC). This optimum is achievable through Pigouvian taxation equivalent to the congestion charge τ^*. In the absence of congestion pricing, the shaded area ABE measures the deadweight loss of congestion. It is the true economic cost of traffic congestion.

Figure 1: The Basic Model of Traffic Congestion



Measuring the welfare loss associated with congestion is challenging given that standard travel surveys may not be available or have insufficient time and spatial coverage that makes the inverse demand curve or the cost functions difficult to recover (Small et al., 2007). Notable efforts include Walters (1961), Kraus et al. (1976), Keeler and Small (1977), and Dewees (1979) in the US and Canada. The advent of new data sources (in particular: Google Maps, traffic microwave sensors, and GPS) may thus offer promising potential for overcoming these challenges, including in developing countries where data leapfrogging could be occurring.

Akbar and Duranton (2017), for example, provide one of the first empirical attempts that leverage big data to measure the deadweight loss of congestion. Combining a travel survey with counterfactual trip information extracted from Google Maps, the authors successfully recover supply and demand for road travel in Bogotá. In their empirical setting, big data from Google Maps serves two main purposes. First, it provides information on distance and real-time trip duration that overcome extensive data require-

ments to estimate travel demand and supply. Second, counterfactual trip information extracted from Google Maps allows comparison of similar trips under different traffic scenarios, which helps attenuate endogeneity issues by breaking any correlation between the dependent variable and the error term.[2] The authors find a negligible welfare loss associated with traffic congestion, estimated to be less than 1 percent of the daily wage in Colombia. This result is consistent with mobility patterns in urban India, where the examination of traffic data collected via Google Maps (Akbar et al., 2018) or smartphone applications (Kreindler, 2018) also points to small welfare effects of congestion.

In a recent research, Yang et al. (2020) combine nearly 400 million real-time observations of speed and traffic volume data with a plausibly exogenous policy shock inducing a variation in traffic congestion (driving restrictions based on the last digit of the license plate number) to estimate the marginal external congestion cost in Beijing. Using traffic microwave sensors, they collected real-time traffic flows at two-minute intervals for the city's major roads in 2014. The authors find that the marginal external congestion cost can rise up to $0.30 per vehicle-km while the deadweight loss of road congestion represents about half percent of Beijing's GDP. The authors find that this welfare loss would increase about three fold if social and environmental externalities such as accidents and air pollution were included in the calculation.

B. Mobility, Employment and Productivity

There is an established literature that uses traditional data (mainly travel surveys) to analyze the reciprocal links between mobility and economic outcomes. For example, Bhat and Koppelman (1993) use the Dutch Mobility Panel to show that employment and household income have a significant influence on car ownership, trip generation,

[2] Potential identification challenges may arise from trip selection and the simultaneous determination of trip cost and flows. To overcome these issues, the authors combine a number of empirical strategies, including controlling for weather shocks, enlarging the study area to multiple road segments, and considering several counterfactual time costs for similar trips under various scenarios.

and mode choice. Looking at the reverse sense of causation (i.e., from traffic to economic outcomes), several papers point to externalities from cars (mainly congestion) as a significant bottleneck for economic activities. There is empirical evidence for the US that congestion hinders employment growth (Boarnet, 1997), lowers productivity (Fernald, 1999), shrinks output growth (Hymel, 2009) and reduces income (Jin and Rafferty, 2017).[3] The economic effects of congestion are found to be stronger in vehicle-intensive industries (e.g., manufacturing) relative to other industries, as documented for the US (Fernald, 1999) and the UK (Graham, 2007). These findings, however, have been challenged by a recent study, which reexamines US metropolitan area data over 30 years, and concludes that the impacts have been overestimated (Marshall and Dumbaugh, 2018).

By generating information on people, places, and activities at high spatial and temporal scales, big data has the potential to shed light on the causes and consequences of both mobility and congestion. Regarding the causes of congestion, a recent paper exploits vehicle counts from aerial imagery to document the impact of ride-hail services on congestion in New York City (Mangrum and Molnar, 2018). To our knowledge, the productivity and employment consequences of congestion have not yet been assessed using big data. As for mobility, several studies use big data to relate it to productivity and employment outcomes. For example, Toole et al. (2015) use CDR data to study the reciprocal links between employment status and individual mobility in two undisclosed European countries. In the first country, the authors focus on the closure of a large manufacturing firm and measure the reduction in the intra-urban mobility of laid-off workers inferred from changes in geo-referenced mobile phone activity. Conversely, the authors apply IA techniques in the second country to estimate aggregate unemployment rates from individual users' mobile phone behavior. Another study uses a gravity model to relate commuting flows estimated with CDR data to within-city variations in wages and productivity in Bangladesh and Sri Lanka (Kreindler and Miyauchi, 2017). It shows that

[3] For surveys of the economic impacts of transportation, see Redding and Turner (2015) and Berg et al. (2017) for the case of developing countries.

mobility is negatively affected by strikes in Bangladesh, which reduces output growth relative to a typical workday.

C. Health and Environmental Externalities

Transport involves various social and environmental externalities, including road accidents, air and noise pollution, and associated emotional stress and health issues. Car accidents and related injuries have been found to increase with traffic density (Bauernschuster et al., 2017), leading to higher automobile insurance premia (Edlin and Karaca-Mandic, 2006). The health effects of transport-induced pollution—mainly from nitrogen oxides (NO_x), carbon monoxide (CO), volatile organic compounds (VOC), particulate matter (PM), and unburned hydrocarbons—are well documented in the literature. These effects include cardiovascular and pulmonary problems (Bauernschuster et al., 2017), low birth weight, prematurity and increasing risks of infant mortality (Currie and Walker, 2011; Knittel et al., 2016). These health effects, in turn, may have economic implications. Empirical evidence from London (Duffy and McGoldrick, 1990) and Los Angeles (Evans and Carrère, 1991) suggests that exposure to highly congested traffic increases the level of psychological stress, which may negatively affect task performance (Stokols et al., 1978) and proofreading abilities (Schaeffer et al., 1988). In contrast to these findings, however, a recent study using a large sample of driving instances from the American Time Use Survey (ATUS) finds that congestion during rush hour commuting has little impact on happiness, sadness, stress, and fatigue.

With the rise of big data, the opportunity to better understand congestion-induced social and environmental externalities has drastically improved. Traffic emissions can now be tracked with more precision using urban sensors in combination with crowd-generated points of interest (Thakuriah et al., 2017). For example, Nyhan et al. (2016) analyze GPS data from a taxi fleet of over 15,000 vehicles in Singapore to estimate emissions from different sources of pollutants at high spatial resolution and temporal scale. Gately

et al. (2017) use a similar approach to estimate traffic emissions in the Boston metropolitan area. Their results suggest that only 10 percent of the road network accounts for 70 percent of traffic- related air pollution. Linking emissions data with privacy-preserving medical records can provide even better insights into the impacts of emissions on a variety of health outcomes (see, e.g., Kho et al., 2015). Reports or complaints about congestion or traffic-related incidents can also be tracked using social media data (Hasan et al., 2013). Combining traffic data from the California Department of Transportation with Twitter data, Georgiou et al. (2015) find a strong correlation between users' complaints and the severity of traffic congestion. Using Twitter data, Wang et al. (2015) estimate aggregate congestion in Chicago with about 80 percent accuracy.

IV. Combining Big Data and Economic Approaches for Prediction

Finally, an important trend in the big data literature is that it is not only used for past and real time measurement of phenomena, but increasingly used to make predictions. The predictive capability of big data analytics can be useful to construct credible counterfactual for causal inference using observational studies. In this section, we briefly review traffic flow prediction attempts that make use of big data, and detail how big data techniques can incorporate economic insights to evaluate the causal effect of driving restriction policies on traffic congestion.

A. Big Data, Machine Learning, and Econometrics

Transportation experts can rely on big data to generate accurate and timely information for traffic flow prediction, a crucial step for improving real-time traffic management and mitigating urban congestion (Lv et al., 2015; Chen et al., 2017). Traffic flow prediction usually begins with mining vast amounts of unstructured locational data extracted from physical and social sensing technologies. The data are then analyzed through

various standard statistical methods and machine learning techniques to detect traffic patterns and to build predictive models of those patterns. The performance of the underlying algorithms varies in many ways that reflect in part their ability to account for network-wide heterogeneity and spatiotemporal relations while maximizing in-sample and out-of-sample prediction (Einav and Levin, 2014a; Lv et al., 2015; Ma et al., 2017). Comparative studies between big data analytics and traditional statistical methods to predict traffic flows point to the superiority of the big data approach which can better account for non-linear relations (Lv et al., 2015; Ma et al., 2017).

Big data analytics and predictive modeling are also increasingly influencing urban management and congestion mitigation efforts (Hampson et al., 2017). The city of Tokyo, for instance, has partnered with a private firm to develop a smartphone compatible app, *Zenryoku Annai!*, that analyzes nearly 360 million observations every second to generate real-time information on the shortest and least-congested travel routes. A similar intelligent transport system (ITS) in Denmark, *Copenhagen Connecting*, was implemented to promote transport sustainability through real-time digital traffic control and weather adaptation options. Di Lorenzo et al. (2016) use data from Abidjan, Côte d'Ivoire to develop a CDR-based intelligent platform, *AllBoard*, which provides local authorities with real-time visualization and a public transit optimization system.

Unlike most machine learning specialists, who tend to focus primarily on predictive modeling from robust data correlations, economists are also concerned with establishing causality between variables. Formally, an econometric model seeking to identify the causal impact of traffic conditions C_i on a given outcome Y_i (e.g., health, employment, firm productivity, economic growth, pollution, etc.) for observation i (household, firm, road segment or network, city, metropolitan area, county, country, etc.) can be expressed as:

$$Y_i = \beta_0 + \beta_1 C_i + \mathbf{X_i'}\gamma + \varepsilon_i, \tag{1}$$

where $\mathbf{X_i}$ is a set of relevant exogenous variables whose effects are described by parameter γ, β_0 is the intercept, and ε_i is the idiosyncratic error term. The parameter of interest is β_1, which captures the impact of a change in congestion C_i on the outcome of interest Y_i.

The principal challenge in estimating equation 1 is the potential endogeneity of traffic conditions. Examples of such endogeneity issues may arise from the non-random supply of transport services or the spatial sorting of users of the transport system. For instance, although congestion may affect economic growth there can be a reverse causality whereby growth induces greater demand for transportation. Traditionally, the empirical economics literature has relied on instrumental variables approaches or natural or quasi-natural experiments to address these potential endogeneity issues (see, e.g., Hymel, 2009; Sweet, 2014). By facilitating access to new data sources, improving the measurement of relevant variables (including those that were previously unobservable), big data has the potential to attenuate some of the endogeneity biases through reduced measurement errors and better controls (Einav and Levin, 2014b; Glaeser et al., 2018), but does not solve all endogeneity issues.

For economists, the valued-added of big data also lies in the ability to explore a wider range of outcomes usually absent in traditional data sets. The often near-universal coverage of big data also opens up new avenues to examine the heterogeneous and distributional effects of transport conditions and policies.

B. Application: The Congestion Impact of the Covid-19 Lockdown

As an illustration, we examine traffic data constructed from publicly available information on Google Maps to estimate the congestion effect of mobility restriction policies during the Sars-Cov-2 pandemic.[4] We rely on the Bayesian Structural Time Series (BSTS)

[4] Mobility restriction policies are part of so-called non-pharmaceutical interventions (NPI) in response to the Covid-19 pandemic. They have been implemented at various geographical levels globally and are commonly referred to as *stay-at-home orders* or *lockdowns*.

model, a state-space model that can be used to construct counterfactuals from observational data in time series settings (Varian, 2014; Brodersen et al., 2015). Applying the BSTS to the traffic data, we compare actual congestion outcomes with an estimate of what *would have been* observed in the absence of a lockdown and infer the causal impact of the restriction policy on congestion. We focus on Bogotá, New Delhi, New York City, and Paris, four highly congested metropolitan areas that enforced stringent lockdown policies to curb the spread of Covid-19.

B.1. Measuring Traffic Congestion

Our data set comes from trafficindex.org, a website that uses the Google Maps traffic layers to compile time series observations on traffic congestion. Google Live Traffic Information, the underlying data source, reports high resolution traffic conditions from anonymized users of a smartphone-enabled application. The layers are based on a four-color scheme that characterizes the traffic status in real-time on a given road segment. More specifically, the color green indicates relatively free-flow traffic conditions while dark red denotes highly congested traffic. Orange and red, the two intermediate colors, correspond to medium and near-heavy congestion, respectively.

According to the website, the images of each road segment for both the entire city and the city center only were queried every 20 minutes, 24 hours per day, using the Google Maps Application Programming Interface.[5] Then, using the pixel-level information on of these images, the website computes a daily Traffic Congestion Index (TCI) at each level of aggregation (city center or entire city) as:

$$\textbf{TCI} = \sum_{k=0}^{3} k p_k, \tag{2}$$

[5] We had similar images queried for Paris from Google Maps and compared them to a sample of data sets from the website covering the same period. No major discrepancies between the two series were found.

where p_k is the percentage of total pixels in each location with color k, with k equal to 0 for green, 1 for orange, 2 for red, and 3 for dark red, respectively. To obtain the daily congestion index, we compute the index for every hour and take the average over 24 hours. By construction, the figure varies between 0 and 300, where 0 indicates a free-flow traffic condition and 300 a highly congested condition. We interpret the traffic congestion index (TCI) as a measure of aggregate delay from free-flow speed.[6]

We use and construct the daily congestion index over the period from March 20, 2019 to May 19, 2020 for several major cities with varying degrees of COVID-19 lockdowns. These include Paris and New York City, subject to strict lockdown measures starting on March 17 and March 20, 2020, respectively. We also include Bogotá and New Delhi, both of which started to enforce strict COVID-19 lockdown policies on March 25, 2020. We use the traffic conditions in Hong Kong SAR, China, a city that did not implement any COVID-19 lockdown measure, as our main control series for modeling the counterfactual scenarios for Bogotá, New Delhi, New York City, and Paris. Alternatively, in addition to using the TCI of Hong Kong SAR, we also use the TCI from cities such as Cairo, Sofia, Stockholm, and Tokyo that opted for relatively softer lockdown measures. Finally, we enrich our control traffic congestion index time series with city-level information on precipitation and temperature from Menne et al. (2012a,b). In what follows, we focus on core cities only. The methodology is, of course, replicable to measures of congestion at the greater metropolitan area level.

[6] Dasgupta et al. (2020) use the same underlying information from Google Maps to calculate the average speed within a city. Their approach, however, requires estimating a city-specific correspondence between speeds and colors and assuming that the correspondence is stable over time and over different locations in the city. It overlooks the fact that free flow may occur at low speeds in the presence of speed limitations, which are common in urban contexts. In contrast, the TCI focuses on the severity of delays while making it possible to aggregate the information over the whole network even when free-flow conditions might involve different speeds in different parts of the network.

B.2. Bayesian Structural Time Series

To estimate the causal impact of the Covid-19 lockdowns on traffic congestion in sample of selected cities, we use a time-varying-parameter (state-space) model known as Bayesian Structural Time Series (BSTS). As in Brodersen et al. (2015), we formally represent the BSTS model in the following block of equations:

$$\begin{aligned} y_t &= \mu_t + \tau_t + \beta^T \mathbf{x}_t + \varepsilon_t \\ \mu_t &= \mu_{t-1} + \delta_{t-1} + u_t \\ \delta_t &= \delta_{t-1} + v_t \\ \tau_t &= -\sum_{s=1}^{S-1} \tau_{t-s} + w_t, \end{aligned} \tag{3}$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma_t^2)$ and $\eta_t \sim \mathcal{N}(0, Q_t)$ are error terms independent of all other unknowns, with $\eta_t = (u_t, v_t, w_t)$ containing independent components of a Gaussian white noise.

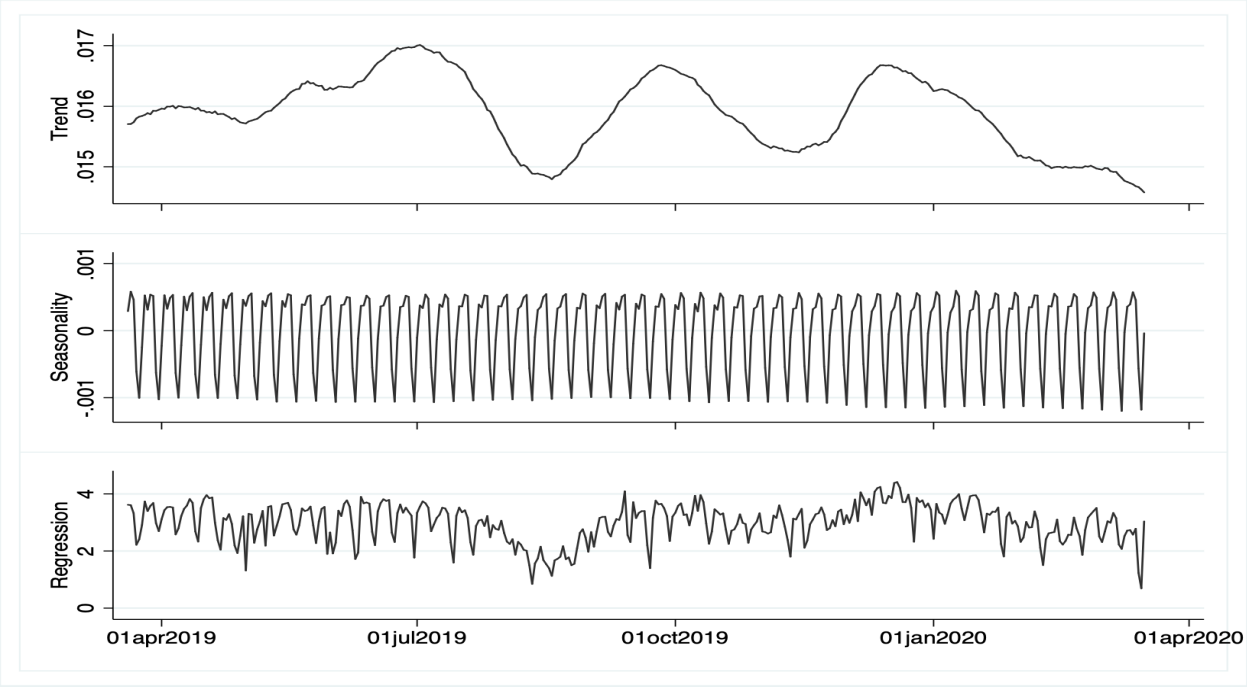
The model in equation 3 links the observed traffic congestion index time series y_t to three components: a local linear trend component μ_t, a seasonal component τ_t, and a regression component $\beta^T \mathbf{x}_t$, where $\mathbf{x}_t$ is a vector of TCI of cities in the control group as well as other controls (such as temperature and precipitation). The second equation of the model displays the current level of the local linear trend μ_t, which depends on its previous level μ_{t-1} and its expected increase between $t-1$ and t—hence the term δ_{t-1}. The latter is often thought of as the *slope* of the linear trend at time $t-1$ and may obey a random walk as specified in the third equation of the model (Scott and Varian, 2014). The local linear trend model is useful for short-term predictions as it is relatively flexible to account for variations in the local context (Brodersen et al., 2015). The fourth equation of the model depicts the joint contribution of the number of seasons S to the observed congestion index data y_t. We set $S = 7$ to capture the day-of-week effect in the daily traffic congestion index series. Thus, the *seasonal* component is incorporated as a set of

day-of-week dummy variables such that the total seasonal effect equals zero over a full cycle of 7 days.

For estimating the causal effect of the Covid-19 lockdowns in our sample of cities, the regression component is crucial. It allows us to construct counterfactual predictions of the daily traffic congestion index series from a synthetic control using traffic conditions from a city or a combination of cities that did not face Covid-19 lockdown measures (*untreated cities*). Unlike the local trend and seasonal sub-models, the observed traffic congestion index series from the pool of *untreated* cities allows us to capture the variance in traffic conditions in the treated cities. In this application, we primarily use Hong Kong SAR as a relevant untreated candidate city for two reasons other than data availability. First, Hong Kong SAR did not implement a Covid-19 lockdown over the entire period of the analysis. Second, and more importantly, it is implausible that traffic conditions in Hong Kong SAR are affected by the effects of lockdown-related changes in our sample of treated cities (Bogotá, New Delhi, New York City, and Paris), an important assumption for causal inference in the BSTS setting.

Using the natural logarithm of the traffic congestion index data from Paris before the lockdown—i.e. between March 20, 2019 and March 17, 2020—we show in figure 2 the contribution to the BSTS model of the trend, seasonal, and regression components, respectively. There is clearly a daily pattern in the congestion data characterized by significantly more variation in the seasonal component relative to both the trend and regression components. With substantially fewer variation than in the regression component, the local trend exhibits a few instances of major reduction in congestion, especially around summer and other national holidays (e.g., Armistice day), followed by relatively short-lived episodes of increased traffic congestion. Note that in addition to the traffic congestion index for Hong Kong SAR, the covariates in the regression component include precipitation and the natural logarithm of temperature. These two control time series are also unlikely to be themselves affected by the lockdown policy in the

Figure 2: Contribution of the Trend, Seasonal, and Regression Components (Paris)



short-term.

Another distinctive feature of the Bayesian structural time series model is the integration of a variable selection mechanism, known as *spike-and-slab* technique, in the regression sub-model. This is a two-step procedure that involves: (i) evaluating the probability that a given covariate is selected in the model (*spike*) and; (ii) shrinking the non-zero coefficients towards prior expectations (*slab*). In the case of many potential covariates, the spike-and-slab prior provides a powerful way of incorporating uncertainties about which variables to include based on their relative contribution to the prediction model. This is useful for reducing the complexity of the model in order to avoid overfitting, a recurrent problem in the big data literature.[7] However, it is worth acknowledging that overfitting is unlikely to be an issue in our application, given the limited number of covariates.

[7] For a technical discussion of the spike-and-slab approach, see the papers by Scott and Varian (2014) and Brodersen et al. (2015).

B.3. Causal Impact of the Covid-19 Lockdown on Congestion

After the World Health Organization (WHO) declared Covid-19 a global pandemic on March 11, 2020, many countries implemented restrictive policies of various degrees (stringent vs. lenient) and scope (national or subnational) to prevent further spread of the virus. These include, among others, social distancing, quarantine, isolation and community containment measures. As early as March 17, 2020, the French government banned all non-essential trips and required people leaving their homes for essential journeys to remain within a 1-km perimeter for up to 1 hour while carrying a signed and dated exception form. The initial 15-day stay-at-home order was later extended until May 5, 2020. The governor of New York, where the number of Covid-19 infections and deaths had drastically increased meanwhile followed suit and ordered that all non-essential businesses be shutdown on March 20, 2020. Both Colombia and India enforced similar countrywide stringent lockdown actions starting on March 25, 2020.

Using the above Bayesian structural time series model, we estimate the causal effect of the Covid-19 lockdown policies on traffic congestion in Bogotá, New Delhi, New York City, and Paris. All four metropolitan areas are consistently ranked among the most congested cities in the world. For example, Bogotá topped the Inrix rank of the world most congested cities in 2019, with drivers loosing on average the equivalent of 7 days and 23 hours a year (Reed, 2019).[8] Traffic congestion in New York City, the fourth most congested place in the United States according to Inrix, cost the city about $US11 billion in 2019. In Paris, a typical driver lost about 165 hours a year due to traffic congestion, making the city the 7th most congested in the world in 2019. Similarly, New Delhi was ranked the 8th most congested city of the world, according to the 2019 TomTom traffic congestion index. Together, while congestion remains a pressing policy challenge in these cities, existing policies are either politically difficult to adopt (e.g., congestion pricing) or have yet to deliver on their promises.

[8] Inrix ranks cities according to the number of hours spent in congestion (see www.inrix.com).

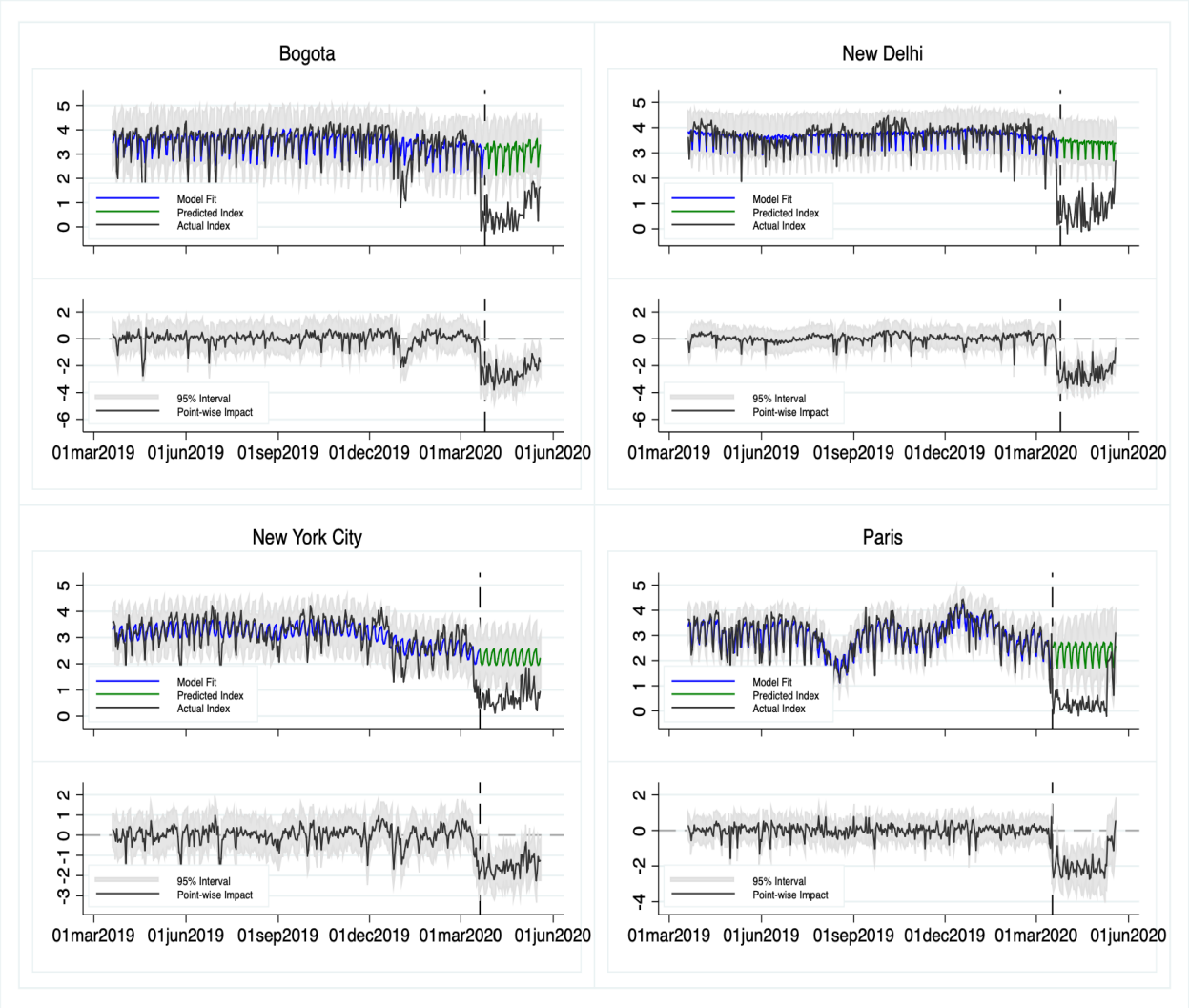
Result 1: Using Only Hong Kong SAR as Control City.

We begin by estimating the causal impact of a strict lockdown policy on traffic congestion in Bogotá, New Delhi, New York City, and Paris. We use traffic conditions in Hong Kong SAR as well as daily temperature and precipitation data for the relevant treated cities as well as for the control series. The control covariates of the regression sub-model in the BSTS are unlikely to be directly affected by the Covid-19 restriction policies in the treated cities, as Hong Kong SAR did not implement any lockdown measure and Covid-19 cannot plausibly explain short-term variations in the weather variables.

As shown in both Figure 3 and Panel A of Table 1, we find that Covid-19 lockdown policies drastically reduced daily traffic congestion—as measured by the logarithm of the traffic congestion index—in each of the four selected cities that implemented strict stay-at-home orders during the pandemic. Overall, the BSTS model fits very well the trajectory of the traffic congestion index during the pre-lockdown period in all cities. For example, the synthetic control series using Hong Kong SAR and the weather series did a remarkable performance of picking up nearly all episodes of spike and dip of the pre-lockdown traffic conditions, especially in New Delhi and Paris (Figure 3).

There is also an unambiguous divergence between observed traffic conditions and the counterfactual predictions following the implementation of the Covid-19 lockdown policies, with much less traffic congestion in the post-lockdown period. More specifically, as shown in Panel A of Table 1, the lockdown-induced reduction in the traffic congestion index varies between 68% and 81% in comparison to a counterfactual *business-as-usual* scenario in the treated cities. Bogotá and Paris exhibit the largest decline in relative congestion with 81% and 79.5%, respectively, followed by New Delhi (-77%) and New York City (-68%). These estimated causal effects are substantial and statistically significant, according to the 95% posterior confidence interval reported in brackets, which excludes 0 for all treated cities. Furthermore, the probability of obtaining this effect by chance is very small (the one-sided tail-area probability p-value is less than 0.01 for all cities),

Figure 3: The Congestion Impact of the Covid-19 Lockdown Policy



Notes: Causal effect of the Covid-19 lockdowns on traffic congestion in Bogotá, New Delhi, New York City, and Paris. For each city: (i) the upper panel shows the time series of the logarithm of the traffic congestion index; and (ii) the lower panel shows the daily incremental impact of the lockdown measure on congestion captured by the difference between the predicted counterfactual and the actual measure of congestion.

another evidence of the statistically significance of the estimated causal effect of travel restriction on congestion.

Result 2: Adding Other Control Cities to Hong-Kong SAR, China

As the global Covid-19 pandemic unfolded, other national and subnational government units opted for alternative restrictive measures without imposing strict stay-at-home orders. These alternative policy actions involved the implementation of large-scale

Table 1: The Causal Impact of Covid-19 Lockdown on Traffic Congestion

	Bogotá	New Delhi	New York	Paris
A. Result 1				
Actual	0.593	0.767	0.721	0.486
Prediction	3.118	3.337	2.235	2.367
	(0.120)	(0.117)	(0.206)	(0.399)
Relative effect	-80.99%	-77.00%	-67.76%	-79.48%
95% CI	[-89%, -74%]	[-84%, -70%]	[-85%, -50%]	[-114%, -47%]
Tail-area probability	0.001	0.001	0.003	0.001
B. Result 2				
Actual	0.593	0.767	0.721	0.486
Prediction	2.787	3.344	2.208	2.158
	(0.116)	(0.085)	(0.103)	(0.405)
Relative effect	-78.74%	-77.05%	-67.35%	-77.49%
95% CI	[-87%, -71%]	[-82%, -72%]	[-76%, -58%]	[-116%, -41%]
Tail-area probability	0.001	0.001	0.001	0.001

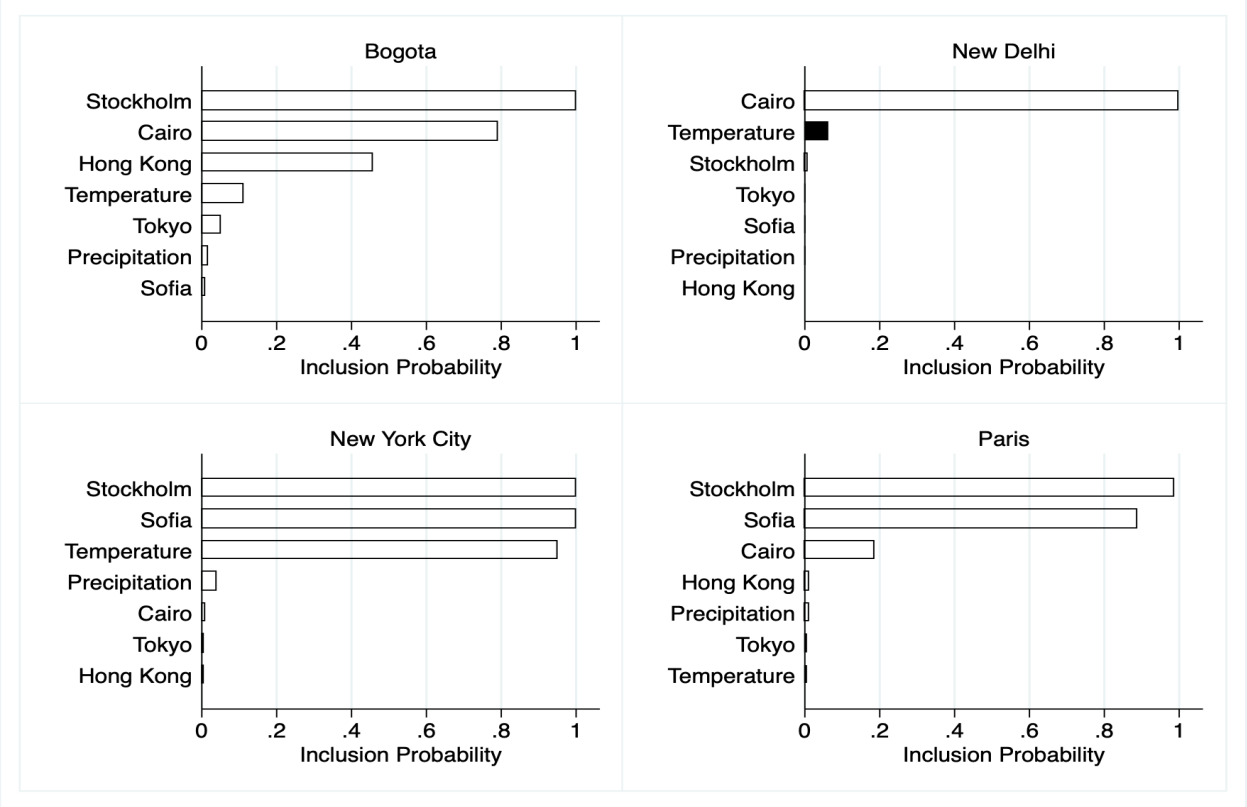
Notes: The table shows the effect of the Covid-19 lockdowns on traffic congestion in Bogotá, New Delhi, New York City, and Paris. For each city, we report the BSTS model using: (A) traffic conditions in Hong Kong and weather in treated cities as the synthetic control series; and (B) traffic condition in Hong Kong, Cairo, Sofia, Stockholm, and Tokyo as the synthetic control series.

social distancing, mandatory face mask-wearing measures, or state of emergency. Using a sample of cities for which we have the required data, we augment our initial list of synthetic control series—i.e. weather conditions in treated cities and congestion in Hong Kong SAR—with traffic congestion data from Cairo, Sofia, Stockholm, and Tokyo, all of which enforced some form of relatively flexible and less constraining lockdown policies. Like in the case of Hong Kong SAR, we do not expect traffic restrictions in the treated cities to directly influence traffic conditions in these cities. However, unlike Hong Kong SAR, people in these cities might have responded to the softer lockdown measures by changing their travel behavior in the midst of the Covid-19 outbreak. But this is unlikely to occur exactly the day when the strict lockdown measures where enforced in the treated cities.

We show in Figure 4 the posterior inclusion probability of the synthetic control series in the regression sub-model of the Bayesian Structural Time Series model. For Bogotá, the covariates with a probability of inclusion exceeding 0.05 are traffic conditions in Stockholm (with an inclusion probability of 1), Cairo (0.79), Hong Kong SAR (0.46),

Tokyo (0.05), and the temperature in the city (0.11). For New Delhi, traffic conditions in Cairo (1) and the city temperature (0.06) are the only predictors with an inclusion probability above 0.05. It is worth noting that temperature in New Delhi is negatively correlated with traffic congestion. As for traffic conditions in Stockholm and Sofia, they are very good predictors of traffic congestion in Paris and New York City, as they both exhibit inclusion probabilities over 0.85. Temperature is positively correlated to traffic conditions in New York City and exhibits an inclusion probability in the BSTS model of 0.95.

Figure 4: Posterior Inclusion Probabilities



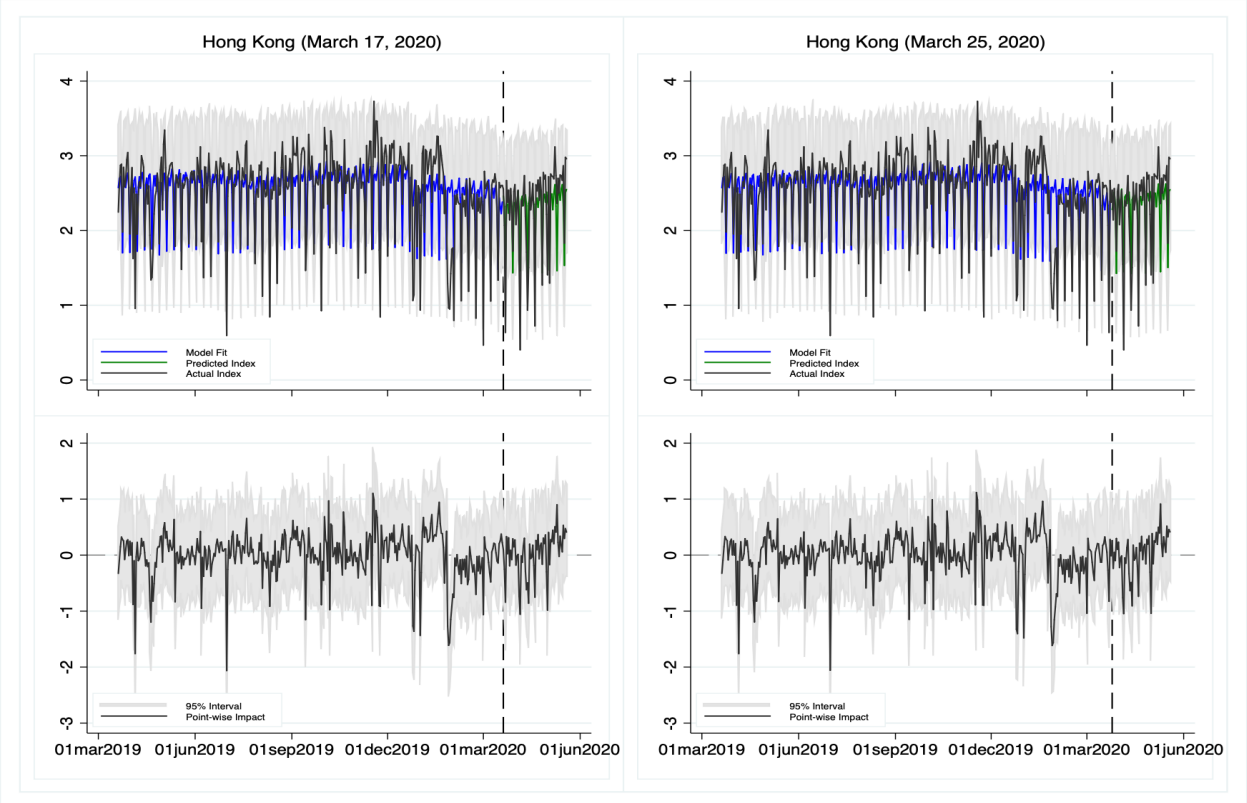
Notes: These graphs show the posterior inclusion probabilities for the predictors of the traffic congestion index for Bogotá, New Delhi, New York City, and Paris. White bars indicate positive coefficients while black bars show negative coefficients.

As shown in Panel B of Table 1, we find large differences between the observed traffic congestion and the model predictions of *what would have happened* without strict Covid-19 lockdown measures in the four cities in our study. In terms of relative effects, the lock-

down policies reduced traffic congestion by about 79% in Bogotá, 77% in New Delhi, 67% in New York City, and 77% in Paris. These relative changes in traffic congestion due to the lockdown measures are similar in magnitude to those from the previous analysis. Furthermore, these causal effects are large and statistically different from zero, with their 95% confidence intervals that almost perfectly overlap the 95% confidence intervals of the previous results. Put differently, the use of traffic conditions in cities with lenient lockdown measures as additional synthetic controls, did not significantly affect the previous analysis using only Hong Kong SAR as the control city. In sum, increasing the number of cities contributing to the synthetic control series did not affect the robustness of the initial finding that mobility restriction policies reduced traffic congestion.

Result 3: Placebo Analysis.

Figure 5: Placebo Test using Hong Kong SAR



Notes: Placebo test of a causal effect of the Covid-19 lockdown on traffic congestion in Hong Kong SAR. We consider two hypothetical treatment dates: (i) March 17, 2020 (left panel); and (ii) March 25, 2020 (right panel).

The above findings suggest that strict lockdown measures led to a significant reduction in traffic congestion. If this were truly the case only in the context of cities that enforced stringent lockdown policies, there should be no congestion effect in Hong Kong SAR, a city that avoided a lockdown. To examine this hypothesis, we use the cities of Cairo, Sofia, Stockholm, and Tokyo to construct a synthetic control series of traffic congestion that we compare to the observed traffic congestion index for Hong Kong SAR.[9] As shown in Figure 5, we find no significant effect of a strict lockdown policy on congestion in Hong Kong SAR. This finding is robust to the use of various hypothetical enforcement periods. More specifically, we find an overall 1.8% (1.04%) reduction in traffic congestion with a 95% confidence interval of [-8%, +11%] ([-7%, +9%]) when the hypothetical treatment date is set to March 17, 2020 (March 25, 2020). Regardless of the treatment date, these effects are quite small and statistically insignificant.

Together, we have documented a causal relationship between travel restriction policies during the Covid-19 crisis and traffic congestion in cities within the context of both developed and developing countries. Our findings are robust to the use of alternative cities with more or less lenient lockdown measures as synthetic control series. Moreover, as expected, we find no effect of lockdown on traffic congestion in Hong Kong SAR, a city which refrained from enforcing a stay-at-home order. These results suggest that policies that incentivize teleworking may effectively complement existing efforts to reduce traffic congestion and associated externalities. However, the overall economic impact would have to account for the reduction in transport activities and whether telecommuting can efficiently substitute for on-site work.

[9] Although, we also use temperature (in log) as an additional control covariate, we did not include a precipitation series due to data unavailability.

V. Discussion and Conclusion

Big data is increasingly generating attractive opportunities for multi-sectoral collaborations involving government agencies, the private sector, and research institutions. Collectively, these actors are working to identify, organize, and harness the potential of large-scale administrative or private sector data. For developing countries, this technological progress offers an opportunity to leapfrog from the scarcity of traditional household travel surveys—deemed to be costly and often inappropriately administered—to high dimensional and cost-effective big data resources. For instance, call detail records (CDR) are increasingly used as reliable alternatives to conventional travel surveys, either exclusively (Alexander et al., 2015; Çolak et al., 2015) or in combination with other data sources (Wang et al., 2012; Jiang et al., 2013; Iqbal et al., 2014; Widhalm et al., 2015).

The quest for better understanding the causes and consequences of mobility and congestion arising from traffic flows, or reflecting locational behaviors of individuals or firms, has generated promising prospects for policy design. The availability of large-scale data sets at high spatial and temporal resolution is useful to detect and predict mobility patterns and associated externalities. It also supports transport research and enables policy makers to assess the welfare impacts of their interventions, through various channels including the reduction of travel delays, accidents and pollution and the stimulation of output and labor productivity. Planners can capitalize on the predictive modeling techniques inherent to big data analytics to improve real-time traffic management and provide their constituents with accurate information on congestion, alternative routes, and multimodal transportation options. Big data may also help design the transition towards more energy-efficient transportation systems by providing real-time behavioral responses (via geocoded social media sensors) to congestion externalities. To illustrate this point, Wang and Moriarty (2018) present a smartphone-based personal travel assistant (PTA) which estimates the energy cost of an individual trip in Beijing, taking into account traffic conditions and modal choice.

Despite the potential benefits of using locational data with near-universal coverage to inform urban and transportation policies, there are challenges inherent to the big data revolution. One important challenge is the technical expertise required to extract millions of pieces of information from heterogeneous sources and process unstructured data sets into operational decision-making inputs. Another crucial challenge for generating and manipulating big data for transportation analyses relates to the preservation of privacy and confidentiality. Given their geo-referenced nature and the possibility to extract them without users' formal consent, urban and social sensing data as well as other user-generated content (UGC) can significantly compromise individual privacy. A combination of technical and regulatory approaches has been explored to address these issues without compromising the social and economic benefits. For example, Thakuriah et al. (2017) describe a number of technological solutions, including Privacy Enhancing Technologies (PET), synthetic data, and Trusted Third Party (TTP) mechanisms, that are relevant for privacy preservation. Many organizations and countries have already adapted their personal data protection framework and guidelines to the necessity to reap the benefits of big data while protecting privacy (International Transport Forum, 2015).

In most developing countries, the prospect of big data technology adoption in transportation remains slow to materialize. Possible reasons for the low take-up may stem from the lack of in-house data analytics expertise, low storage capacity for increasingly large data sets, as well as limited collaboration between state agencies, research centers, and the private sector. There are also political economy considerations arising from the coordination failures among the relevant actors and institutions, both in the data mining process and the design of the legal and regulatory environment. Despite these challenges, the penetration of mobile technology at a high pace across different demographic groups in developing countries may open new opportunities for sustainable transportation policies.

References

Akbar, Prottoy A, Victor Couture, Gilles Duranton, Ejaz Ghani, and Adam Storeygard. 2018. "Mobility and Congestion in Urban India." *World Bank Policy Research Working Paper*.

Akbar, Prottoy, and Gilles Duranton. 2017. "Measuring the Cost of Congestion in Highly Congested City: Bogotá." *Mimeo: University of Pennsylvania*.

Alexander, Lauren, Shan Jiang, Mikel Murga, and Marta C González. 2015. "Origin–Destination Trips by Purpose and Time of Day Inferred from Mobile Phone Data." *Transportation research part c: emerging technologies*, 58 240–250.

Askitas, Nikolaos, and Klaus F Zimmermann. 2013. "Nowcasting Business Cycles Using Toll Data." *Journal of Forecasting*, 32(4): 299–306.

Barmpounakis, Emmanouil N, Eleni I Vlahogianni, and John C Golias. 2016. "Extracting Kinematic Characteristics from Unmanned Aerial Vehicles."Technical report.

Bauernschuster, Stefan, Timo Hener, and Helmut Rainer. 2017. "When Labor Disputes Bring Cities to a Standstill: The Impact of Public Transit Strikes on Traffic, Accidents, Air Pollution, and Health." *American Economic Journal: Economic Policy*, 9(1): 1–37.

Berg, Claudia N, Uwe Deichmann, Yishen Liu, and Harris Selod. 2017. "Transport Policies and Development." *The Journal of Development Studies*, 53(4): 465–480.

Bhat, Chandra R, and Frank S Koppelman. 1993. "An Endogenous Switching Simultaneous Equation System of Employment, Income, and Car Ownership." *Transportation Research Part A: Policy and Practice*, 27(6): 447–459.

Blazquez, Desamparados, and Josep Domenech. 2018. "Big Data Sources and Methods for Social and Economic Analyses." *Technological Forecasting and Social Change*, 130 99–113.

Boarnet, Marlon G. 1997. "Infrastructure Services and the Productivity of Public Capital: The Case of Streets and Highways." *National tax journal* 39–57.

Brodersen, Kay H, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L Scott. 2015. "Inferring Causal Impact using Bayesian Structural Time-Series Models." *The Annals of Applied Statistics*, 9(1): 247–274.

Chen, Yuanfang, Mohsen Guizani, Yan Zhang, Lei Wang, Noel Crespi, and Gyu Myoung Lee. 2017. "When Traffic Flow Prediction Meets Wireless Big Data Analytics." *arXiv preprint arXiv:1709.08024*.

Coifman, Benjamin, Mark McCord, Rabi G Mishalani, and Keith Redmill. 2004. "Surface Transportation Surveillance from Unmanned Aerial Vehicles." In *Proc. of the 83rd Annual Meeting of the Transportation Research Board*.

Çolak, Serdar, Lauren P Alexander, Bernardo G Alvim, Shomik R Mehndiratta, and Marta C González. 2015. "Analyzing Cell Phone Location Data for Urban Travel: Current Methods, Limitations, and Opportunities." *Transportation research record: Journal of the transportation research board*(2526): 126–135.

Currie, Janet, and Reed Walker. 2011. "Traffic Congestion and Infant Health: Evidence from E-ZPass." *American Economic Journal: Applied Economics*, 3(1): 65–90.

Daly, Elizabeth M, Freddy Lecue, and Veli Bicer. 2013. "Westland Row Why So Slow? Fusing Social Media and Linked Data Sources for Understanding Real-Time Traffic Conditions." In *Proceedings of the 2013 international conference on Intelligent user interfaces*. 203–212, ACM.

Dasgupta, Susmita, Somik Lall, and David Wheeler. 2020. "Traffic, Air Pollution, and Distributional Impacts in Dar es Salaam: A Spatial Analysis with New Satellite Data."

Dewees, Donald N. 1979. "Estimating the Time Costs of Highway Congestion." *Econometrica: Journal of the Econometric Society* 1499–1512.

Di Lorenzo, Giusy, Marco Sbodio, Francesco Calabrese, Michele Berlingerio, Fabio Pinelli, and Rahul Nair. 2016. "Allaboard: Visual Exploration of Cellphone Mobility Data to Optimise Public Transport." *IEEE transactions on visualization and computer graphics*, 22(2): 1036–1050.

Duffy, Carol A, and Ann E McGoldrick. 1990. "Stress and the Bus Driver in the UK Transport Industry." *Work & Stress*, 4(1): 17–27.

Duranton, Gilles, and Anthony J. Venables. 2018. "Place-Based Policies for Development." *World Bank Policy Research Working Paper*.

Edlin, Aaron S, and Pinar Karaca-Mandic. 2006. "The Accident Externality from Driving." *Journal of Political Economy*, 114(5): 931–955.

Einav, Liran, and Jonathan Levin. 2014a. "The Data Revolution and Economic Analysis." *Innovation Policy and the Economy*, 14(1): 1–24.

Einav, Liran, and Jonathan Levin. 2014b. "Economics in the Age of Big Data." *Science*, 346(6210): , p. 1243089.

Evans, Gary W, and Sybil Carrère. 1991. "Traffic Congestion, Perceived Control, and Psychophysiological Stress among Urban Bus Drivers.." *Journal of Applied Psychology*, 76(5): , p. 658.

Fernald, John G. 1999. "Roads to prosperity? Assessing the Link Between Public Capital and Productivity." *American economic review*, 89(3): 619–638.

Gately, Conor K, Lucy R Hutyra, Scott Peterson, and Ian Sue Wing. 2017. "Urban Emissions Hotspots: Quantifying Vehicle Congestion and Air Pollution Using Mobile Phone GPS Data." *Environmental pollution*, 229 496–504.

Georgiou, Theodore, Amr El Abbadi, Xifeng Yan, and Jemin George. 2015. "Mining Complaints for Traffic-Jam Estimation: A Social Sensor Application." In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. 330–335, ACM.

Glaeser, Edward L, Scott Duke Kominers, Michael Luca, and Nikhil Naik. 2018. "Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life." *Economic Inquiry*, 56(1): 114–137.

Graham, Daniel J. 2007. "Variable Returns to Agglomeration and the Effect of Road Traffic Congestion." *Journal of Urban Economics*, 62(1): 103–120.

Hampson, Keith, CEO SBEnrc, Peter Newman, Charlie Hargroves, Bela Stantic, Kamal Weeratunga, MRWA Jannatun Haque, and NSW RMS. 2017. "Mining the Datasphere: Big Data, Technologies, and Transportation."Technical report, National Research Centre - Curtin University.

Hasan, Samiul, Xianyuan Zhan, and Satish V Ukkusuri. 2013. "Understanding Urban Human Activity and Mobility Patterns Using Large-Scale Location-Based Data from Online Social Media." In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing.*, p. 6, ACM.

He, Jingrui, Wei Shen, Phani Divakaruni, Laura Wynter, and Rick Lawrence. 2013. "Improving Traffic Prediction with Tweet Semantics.." In *IJCAI*. 1387–1393.

Hymel, Kent. 2009. "Does Traffic Congestion Reduce Employment Growth?" *Journal of Urban Economics*, 65(2): 127–135.

International Transport Forum. 2015. "Transport: Understanding and Assessing Options. Corporate Partnership Board Report."

Iqbal, Md Shahadat, Charisma F Choudhury, Pu Wang, and Marta C González. 2014. "Development of Origin–Destination Matrices Using Mobile Phone Call Data." *Transportation Research Part C: Emerging Technologies*, 40 63–74.

Jiang, Shan, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C González. 2013. "A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities." In *Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing.*, p. 2, ACM.

Jin, Jangik, and Peter Rafferty. 2017. "Does Congestion Negatively Affect Income Growth and Employment Growth? Empirical Evidence from Us Metropolitan Regions." *Transport Policy*, 55 1–8.

Keeler, Theodore E, and Kenneth A Small. 1977. "Optimal Peak-Load Pricing, Investment, and Service Levels on Urban Expressways." *Journal of Political Economy*, 85(1): 1–25.

Khan, Muhammad Arsalan, Wim Ectors, Tom Bellemans, Davy Janssens, and Geert Wets. 2017. "UAV-Based Traffic Analysis: A Universal Guiding Framework Based on Literature Survey." ELSEVIER SCIENCE BV.

Kho, Abel N, John P Cashy, Kathryn L Jackson, Adam R Pah, Satyender Goel, Jörn Boehnke, John Eric Humphries, Scott Duke Kominers, Bala N Hota, Shannon A Sims et al. 2015. "Design and Implementation of a Privacy Preserving Electronic Health Record Linkage Tool in Chicago." *Journal of the American Medical Informatics Association*, 22(5): 1072–1080.

Kitchin, Rob. 2014. "The Real-Time City? Big Data and Smart Urbanism." *GeoJournal*, 79(1): 1–14.

Knittel, Christopher R, Douglas L Miller, and Nicholas J Sanders. 2016. "Caution, Drivers! Children Present: Traffic, Pollution, and Infant Health." *Review of Economics and Statistics*, 98(2): 350–366.

Kraus, Marvin, Herbert Mohring, and Thomas Pinfold. 1976. "The Welfare Costs of Nonoptimum Pricing and Investment Policies for Freeway Transportation." *The American Economic Review*, 66(4): 532–547.

Kreindler, Gabriel. 2018. "The Welfare Effect of Road Congestion Pricing: Experimental Evidence and Equilibrium Implications." *Mimeo: Massachusetts Institute of Technology*.

Kreindler, Gabriel, and Yuhei Miyauchi. 2017. "Billions of Calls Away from Home: Measuring Commuting and Productivity inside Cities with Cell Phone Records." *Mimeo: Massachusetts Institute of Technology*.

Litman, Todd. 2013. "Congestion Costing Critique: Critical Evaluation of the 'Urban Mobility Report'."

Lv, Yisheng, Yanjie Duan, Wenwen Kang, Zhengxi Li, Fei-Yue Wang et al. 2015. "Traffic Flow Prediction with Big Data: A Deep Learning Approach." *IEEE Trans. Intelligent Transportation Systems*, 16(2): 865–873.

Ma, Xiaolei, Zhuang Dai, Zhengbing He, Jihui Ma, Yong Wang, and Yunpeng Wang. 2017. "Learning Traffic As Images: A Deep Convolutional Neural Network for Large-Scale Transportation Network Speed Prediction." *Sensors*, 17(4): , p. 818.

Mangrum, Daniel, and Alejandro Molnar. 2018. "The Marginal Congestion of a Taxi in New York City." *Processed, Vanderbilt University*.

Marshall, Wesley E, and Eric Dumbaugh. 2018. "Revisiting the Relationship Between Traffic Congestion and the Economy: A Longitudinal Examination of US Metropolitan Areas." *Transportation* 1–40.

Menne, Matthew J, Imke Durre, Bryant Korzeniewski, Shelley McNeal, Kristy Thomas, Xungang Yin, Steven Anthony, Ron Ray, Russell S Vose, Byron E Gleason et al. 2012a. "Global Historical Climatology Network-Daily (GHCN-Daily), Version 3." *NOAA National Climatic Data Center*, 10, p. V5D21VHZ.

Menne, Matthew J, Imke Durre, Russell S Vose, Byron E Gleason, and Tamara G Houston. 2012b. "An Overview of the Global Historical Climatology Network-Daily Database." *Journal of atmospheric and oceanic technology*, 29(7): 897–910.

Mitchell, D. 2014. "New Traffic Data Sources–An Overview." *Bureau of Infrastructure, Transport and Regional Economics, Canberra, ACT, Australia*.

Mohring, Herbert. 1976. *Transportation Economics*.: Ballinger.

Nyhan, Marguerite, Stanislav Sobolevsky, Chaogui Kang, Prudence Robinson, Andrea Corti, Michael Szell, David Streets, Zifeng Lu, Rex Britter, Steven RH Barrett et al. 2016. "Predicting Vehicular Emissions in High Spatial Resolution Using Pervasively Measured Transportation Data and Microscopic Emissions Model." *Atmospheric Environment*, 140 352–363.

Pan, Bei, Yu Zheng, David Wilkie, and Cyrus Shahabi. 2013. "Crowd Sensing of Traffic Anomalies Based on Human Mobility and Social Media." In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 344–353, ACM.

Puri, Anuj. 2005. "A Survey of Unmanned Aerial Vehicles (UAV) for Traffic Surveillance." *Department of computer science and engineering, University of South Florida* 1–29.

Redding, Stephen J, and Matthew A Turner. 2015. "Transportation costs and the spatial organization of economic activity." In *Handbook of regional and urban economics*. 5: Elsevier, 1339–1398.

Reed, Trevor. 2019. "INRIX Global Traffic Scorecard."

Ribeiro, Anna Izabel João Tostes, Thiago Henrique Silva, Fátima Duarte-Figueiredo, and Antonio AF Loureiro. 2014. "Studying Traffic Conditions by Analyzing Foursquare and Instagram Data." In *Proceedings of the 11th ACM symposium on Performance evaluation of wireless ad hoc, sensor, & ubiquitous networks*. 17–24, ACM.

Schaeffer, Monica H, Stacey W Street, Jerome E Singer, and Andrew Baum. 1988. "Effects of Control on the Stress Reactions of Commuters 1." *Journal of Applied Social Psychology*, 18(11): 944–957.

Schrank, David, Bill Eisele, Tim Lomax, and Jim Bak. 2015. "2015 Urban Mobility Scorecard."

Scott, Steven L, and Hal R Varian. 2014. "Predicting the Present with Bayesian Structural Time Series." *International Journal of Mathematical Modelling and Numerical Optimisation*, 5(1-2): 4–23.

Small, Kenneth A, Erik T Verhoef, and Robin Lindsey. 2007. *The Economics of Urban Transportation*.: Routledge.

Stokols, Daniel, Raymond W Novaco, Jeannette Stokols, and Joan Campbell. 1978. "Traffic Congestion, Type A Behavior, and Stress.." *Journal of Applied Psychology*, 63(4): , p. 467.

Sweet, Matthias N. 2014. "Do Firms Flee Traffic Congestion?" *Journal of Transport Geography*, 35 40–49.

Tao, Sui, Jonathan Corcoran, Iderlina Mateo-Babiano, and David Rohde. 2014. "Exploring Bus Rapid Transit Passenger Travel Behaviour Using Big Data." *Applied geography*, 53 90–104.

Thakuriah, Piyushimita Vonu, Nebiyou Y Tilahun, and Moira Zellner. 2017. "Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery." In *Seeing cities through big data*.: Springer, 11–45.

Toole, Jameson L, Yu-Ru Lin, Erich Muehlegger, Daniel Shoag, Marta C González, and David Lazer. 2015. "Tracking Employment Shocks Using Mobile Phone Data." *Journal of The Royal Society Interface*, 12(107): , p. 20150185.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives*, 28(2): 3–28.

Walters, Alan A. 1961. "The Theory and Measurement of Private and Social Cost of Highway Congestion." *Econometrica: Journal of the Econometric Society* 676–699.

Wang, Pu, Timothy Hunter, Alexandre M Bayen, Katja Schechtner, and Marta C González. 2012. "Understanding Road Usage Patterns in Urban Areas." *Scientific reports*, 2, p. 1001.

Wang, Senzhang, Lifang He, Leon Stenneth, Philip S Yu, and Zhoujun Li. 2015. "Citywide Traffic Congestion Estimation with Social Media." In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems.*, p. 34, ACM.

Wang, Stephen Jia, and Patrick Moriarty. 2018. *Big Data for Urban Sustainability*.: Springer.

Widhalm, Peter, Yingxiang Yang, Michael Ulm, Shounak Athavale, and Marta C González. 2015. "Discovering Urban Activity Patterns in Cell Phone Data." *Transportation*, 42(4): 597–623.

Yang, Jun, Avralt-Od Purevjav, and Shanjun Li. 2020. "The Marginal Cost of Traffic Congestion and Road Pricing: Evidence from a Natural Experiment in Beijing." *American Economic Journal: Economic Policy*, 12(1): 418–53.

Zachau, Ulrich, Sudhir Shetty, Mathew A Verghis, and Frederico Gil Sander. 2015. *Malaysia Economic Monitor June 2015 Transforming Urban Transport*.: The World Bank.