Online appendix:

Mobility and congestion in urban India

Prottoy A. Akbar[†]
Aalto University

Victor Couture[§]
University of British Columbia

Gilles Duranton[‡]
University of Pennsylvania

Adam Storeygard[¶]
Tufts University

November 2022

Abstract: This appendix provides further methodological details about our data construction and additional results for our paper "Mobility and congestion in urban India".

Key words: urban transportation, roads, traffic congestion, travel speed determinants, cities
JEL classification: R41, O18

[†]Aalto University and Helsinki Graduate School of Economics (email: prottoy.akbar@aalto.fi); https://www.prottoyamanakbar.com/.

[§]Vancouver School of Economics, University of British Columbia (email: victor.couture@ubc.ca); https://economics.ubc.ca/faculty-and-staff/victor-couture/.

[‡]Wharton School, University of Pennsylvania (email: duranton@wharton.upenn.edu); https://real-estate.wharton.upenn.edu/profile/21470/.

[¶]Department of Economics, Tufts University (email: Adam.Storeygard@tufts.edu); https://sites.google.com/site/adamstoreygard/.

Appendix A. Main data sources

A. City Definitions

Our city universe and definitions come from two sources. City point locations and populations are from the World Urbanization Prospects, 2018 revision (WUP; United Nations (2019)), which contains all 181 cities in India and 139 in the US with a (projected) population of at least 300,000 in 2018. City boundaries are based on two datasets from the Global Human Settlements Layer (GHSL) version 2016A reporting conditions circa 2014-15: Settlement Model (SMOD; Pesaresi and Freire, 2016), which defines each 1-kilometer grid cell as urban or not, and the 38-meter Built-up grid (BUILT; Corbane *et al.*, 2018). We combine these datasets in several steps for both India and the US:

1. We split "twin" cities that WUP treats as one, but we believe are better conceptualized as two for our purposes (San Francisco-Oakland, Dallas-Fort Worth, and St. Petersburg-Tampa, US). We allocate the total population of the pair to each member based on alternative population sources.

2. We alter the location of some cities as follows. We first compare each city's coordinates to those reported for the same city (based on a search for name and country) by Google Maps (GM). For any city where the discrepancy is greater than 5 kilometers, we manually search to determine which is more appropriate. For two Indian cities we chose GM: Hubli-Dharwad and Kakinada. If neither is appropriate, for example because the GM search returns a smaller place of the same name, we pick another central point in GM: Darbhanga (train station), Korba (central intersection) and Tirunelveli (central temple). We further use GM for all US cities.

3. We assign the nearest set of contiguous SMOD cells (SMOD polygon hereafter) to each WUP point.

4. We drop (a) two cities that would have had too small a population (well under 300,000) if we carried out an analogous split in step 1 above (Bridgeport-Stamford and Poughkeepsie-Newburgh, US), (b) five cities more than 5 kilometers from an SMOD polygon after the spatial join in step 3 (Asheville, Augusta-Richmond County, Jackson (Mississippi), Concord, and Myrtle Beach, US), and (c) Thanjavur, India, whose SMOD polygon is implausibly small (one grid cell).

5. We move WUP points that are outside of but less than 5 kilometers from the nearest SMOD polygon into a more appropriate location within the polygon (Thoothukkudi (Tuticorin), India and Little Rock and Victorville-Hesperia-Apple Valley, US). This does not change the city sample but creates a more plausible city center for trip sampling.

6. We merge WUP points that fall within the same SMOD polygon in one case where we deem it appropriate because they are essentially subcenters of a larger metro area, summing their populations and using the point location of the largest city in the SMOD polygon (The Woodlands absorbed into Houston, US).

7. In the remaining cases where multiple WUP points fall within the same SMOD polygon, we split the polygon into smaller polygons, one surrounding each point. Where relevant, we split the polygon at an international border (Detroit, San Diego, El Paso and Laredo, US) or a large water body (Oakland/San Francisco, US). Where no such clear division is available, we split the SMOD polygon so that the resulting area assigned to each city is proportional to its population, with an elasticity of 0.57, based on Ahlfeldt and Pietrostefani (2019). When there are two cities in the polygon, we use the following algorithm. Let the two cities A and B have center points C_A and C_B, and populations P_A and P_B. We define their boundary as the set of points Z such that $dist(Z, C_A)/dist(Z, C_B) = (P_A/P_B)^{0.57/2}$. That defines a quadratic form. Without loss of generality, define a new Euclidean coordinate system (x, y) such that C_A is the origin and C_B is at $(0, b)$. Then the boundary is defined by: $x^2 + y^2 = [x^2 + (b - y)^2](P_A/P_B)^{0.57}$. Using this algorithm, the following city pairs are split: Raniganj/Durgapur; Durg-Bhilainagar/Raipur; Bhiwandi/Mumbai; Kanhangad/Kannur; Ottappalam/Kozhikode (Calicut); and Kayamkulam/Kollam, India; Round Lake Beach-McHenry-Grayslake/Chicago; and Trenton/Philadelphia, US. In three remaining cases, three cities fall within the same polygon: Mission Viejo/ Riverside-San Bernardino/Los Angeles Denton-Lewisville/Fort Worth/Dallas, and San Jose/Oakland/San Francisco, US. In these cases, we define boundaries pairwise, and then apply them in descending order of city size. The algorithm produces a few small exclaves: areas assigned to a city center that are only connected to it via another city's polygon. In these cases we manually re-assign the exclave to the city to which it is adjacent.

8. Select all BUILT pixel centroids within each resulting city polygon, create a 500-meter buffer around them, and add any areas wholly surrounded by this buffer polygon, to define the final city polygon.

Panel A of figure A.1 below shows our city boundary, and the built-up areas within it, for a medium-sized city, Jamnagar in Gujarat, which we use for illustrative purpose throughout this appendix.

B. Trip Definitions

This section describes how we determine the within-city trips to query on GM. We define a *trip* as an ordered pair of points (origin and destination) within the same city. A *trip instance* is a trip taken at a specific time on a specific day. A *location*/point refers to a pair of longitude-latitude coordinates identifying the centroid of a roughly 40-meter GHSL pixel. We require that trip endpoints are at least one kilometer apart in haversine length, for three reasons. First, the rounding of travel times and lengths introduce potentially non-classical measurement error in our computations of travel speed. Second, GM does not always return a driving time under traffic conditions for very short trips. Even when it does, the returned travel times are sometimes very inconsistent and the routes involve unnecessary detours. Third, walking is an easy alternative to driving for short trips, and sources of error such as the unobserved time cost of finding parking, etc. will be a more significant component of the trip.

Our target sample for city c is 15 $\sqrt{Pop_c}$ trips, where Pop_c is the projected 2018 population of city c from United Nations (2019), and 21 trip instances per trip, to ensure variation across times of day. That is approximately 158,000 trip instances for our smallest cities, 223,000 instances for a median-sized city, and 1,522,000 instances for the largest city (Delhi).[1]

We define four types of trips: radial (2/9 of all trips), circumferential (1/9), gravity (1/3), and amenity (1/3). For each trip between an origin and destination, we also sample the reverse trip.

[1]By comparison, in the 2017 US National Household Travel Survey (NHTS), the 180th, 100th, 50th, 10th and 1st most sampled US metro areas have about 200, 900, 2,500, 12,000, and 49,000 trips, respectively (Duranton, Nagpal, and Turner, 2021).

Radial trips

Radial trips are defined in a polar coordinate system with respect to a city center. They have one end at a randomly located point within 1.5 kilometers of the city center as defined above. Distance from the center is drawn from a truncated normal distribution with mean 0, standard deviation 0.75 kilometer and support $[0, 1.5]$ kilometers. For each 'central' endpoint, we determine the other endpoint using one of two methods with equal probability:

1. *Absolute distances* of $AbsDist \in \{2, 5, 10, 15\}$ kilometers (equally weighted) are drawn. For each of these four distances, we pick a random point within the built-up area of the city that is between $(AbsDist - 0.2)$ and $(AbsDist + 0.2)$ kilometers from the first endpoint.[2] See panel B of figure A.1 for illustration with the city of Jamnagar. Lighter shades of red distinguish longer trips.

2. *Distance percentiles* relative to the largest possible distance for any trip between built-up areas of the city are drawn from a uniform distribution from the 1st to 99th percentile (excluding distances less than 500 meters). See panel C of figure A.1 for illustration with the city of Jamnagar.

If a city has no valid trips for a given absolute distance +/-0.2 kilometer, the trips assigned to that distance are reallocated to the distance percentiles sample.[3] Similarly, if there are not enough unique 40 m pixels that are $AbsDist$ +/-0.2 kilometer from the center destination to meet a given absolute distance's quota, the remainder of the quota is filled with randomly drawn distance percentiles instead.

Circumferential trips

Like radial trips, circumferential trips are also defined in a polar coordinate system with respect to a city center. Circumferential trips originate at a random origin at least 2 kilometers away from the city center. The analogous destination is at the same distance

[2]In the absence of a built-up point between $(AbsDist - k, AbsDist + k)$, we increment k by 0.1 and search again. We stop when either all picked central endpoints have been paired with another random endpoint or until $k = 1$ (at which point, we pick a new central endpoint).

[3]Only 67 (or roughly a third) of the cities have a maximum distance to centroid of 15 kilometers or more. 110 (or 61%) of the cities have a maximum distance of 10 kilometers or more. 169 (or 94%) of the cities have a maximum distance of 5 kilometers or more, and all cities have a maximum distance greater than 2 kilometers (with the smallest maximum distance being 3.6 kilometers).

(+/-0.4 kilometer) from the centroid, 30 (+/-3) degrees clockwise or counter-clockwise from the origin. See panel D of figure A.1 for illustration with the city of Jamnagar.

Gravity trips

Gravity trips are designed to match the length profile of trips sampled in the US NHTS and the Bogotá Travel Survey (Duranton *et al.*, 2021, Akbar and Duranton, 2018). We identified each location-pair using the following algorithm:

1. Consider a randomly picked initial point ($GravityPoint$) and a length ($GravityLength$ kilometers) drawn from a truncated Pareto distribution with shape parameter 1 and with support between 1 kilometer and 250 kilometers (corresponding to a mean of roughly 5.52 kilometers).

2. Choose a point randomly from among all points at a straight-line length between $(GravityLength - 0.2)$ kilometer and $(GravityLength + 0.2)$ kilometer from the point $GravityPoint$. If there are no such points, start over from (1) with a new pair of $(GravityPoint, GravityLength)$.

See panel E of figure A.1 for illustration with the city of Jamnagar. Lighter shades of red distinguish longer trips.

Amenity trips

Amenity trips join a random endpoint with another endpoint corresponding to one of 12 amenity types (e.g. schools, meals, religion). First, we systematically search Google Places for all establishments of each type in a city to identify the most popular business categories associated with them. Then, we use these categories as search keywords to identify the most "prominent" trip destination returned by GM near each trip origin.

The weighting of trips across these amenity types is based on a mapping of amenity types to trip destination purposes from the 2017 US NHTS. The NHTS has 20 categories of trip destination purpose, a subset of which we aggregate up to 12 amenity types (trip share in parentheses, conditional on being in relevant subset): Buy goods (27.1%), work (20.2%), buy meals (14.7%), recreational activities (6.4%), exercise (6.4%), buy services (4.7%), schools (4.5%), religious or community activities (3.8%), healthcare including adult and child care (4.0%), change transport (1.9%), other general errands (5.6%), and other (0.8%).

The set of keywords we use to search for amenities on GM were picked from among the most popular business categories (assigned by Google) among the establishments within 50 kilometers of each city's center.[4] In particular, to identify the keywords for each city:

1. For each amenity type, we exclude business categories that are in the 1% tail of rarest business categories (across all cities) unless doing so leaves us with less than 30 categories – in which case, the size of the tail we remove gets reduced until we have 30 labels. We call this the truncated global category list.

2. We drop any city and amenity type pair with no categories from the global category list. We re-apportion the share of trips to be defined for the city-type pair evenly across all other amenity types in the city.

3. Starting from the global category list for each amenity type, we remove the 1% tail of rarest business categories in the city-type pair. If that leaves us with less than 10 categories, we reduce the size of the tail to remove until we get back to 10. We call this the truncated city category list.

The number of amenities to search for using each keyword in the truncated city category list is proportional to the number of times the keyword appears as a business category in the city.

On each query, GM returns an ordered list of up to 20 establishments, one of which we choose as our amenity endpoint. We want to make sure that the locations we are choosing match our expectations of the trip purpose category and do not require trips that are either too far beyond the extent of the city or too short for reliable travel times. So, we restrict amenity endpoints to establishments:

- with at least one business category from our list of truncated global categories of the corresponding amenity type.

- that are less than 50 kilometers and more than 1 kilometer away in (haversine) distance from where the query is centered.

[4]Choosing different search keywords for each city makes sure that we are defining trips to establishments in places where they exist. For example, churches are very common in some Indian cities but rare in others. Assuming their frequency also signals their popularity as a trip destination, we define more religion trips to churches in cities where they are more common.

We choose as amenity endpoint the establishment that is listed first among the remaining ones. When queries by a particular keyword repeatedly fail to return nearby establishments with a relevant global category, we re-apportion the trip share of that keyword proportionally among the remaining keywords for the amenity type.

The set and ordering of establishments returned by a query depend on the zoom level on the GM interface. GM relies on a combination of proximity and "prominence" to identify the 20 most relevant establishments to return.[5] At high zoom levels, GM puts more weight on establishments close to the center of the search than more prominent establishments further away. We fix the resolution at "14z" for all amenity trips. For some sense of how high a zoom level this is, note that on a typical 19-inch monitor in the Chrome browser, the map covers an area of approximately 8-12 square kilometers.

Panels F, G, and H of figure A.1 illustrate for the city of Jamnagar our selection of school, religion, and healthcare trips, respectively.

C. Querying trips on Google Maps

Our target sample was 2,735,442 trips across all cities and strategies. Between June 5 and November 13 of 2019, we queried 57,175,086 instances of 2,733,330 trips. Of them, we ignore 0.12% of instances, including: (a) 11,220 instances (and 12 trips) that GM did not return any route for, (b) 53,874 instances (and 2,129 trips) with implausible routes where the travel distance was shorter than the straight-line distance, and (c) 6,811 instances (and 220 trips) where the routes were too indirect (travel distance was at least 50 kilometers greater than the straight-line distance). We ended up with 2,730,969 queried trips, or 99.8% of our target. Across cities, the mean is 99.9% with a coefficient of variation of 0.5%.[6] The median (as well as the mean) trip was queried 21 times (with a standard deviation of 4.9) and 90% of the trips were queried at least 15 times.

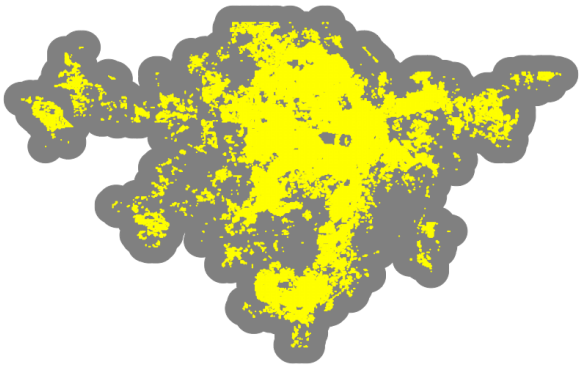
We wanted the distribution of trip departure/query times to roughly resemble the distribution of departure times on a typical weekday.[7] However, we also wanted enough trip queries from each time period of the day for the fixed effects to be precisely estimated, so we

[5] An establishment's "prominence" is determined "by its ranking in GM's index, global popularity, and other factors"

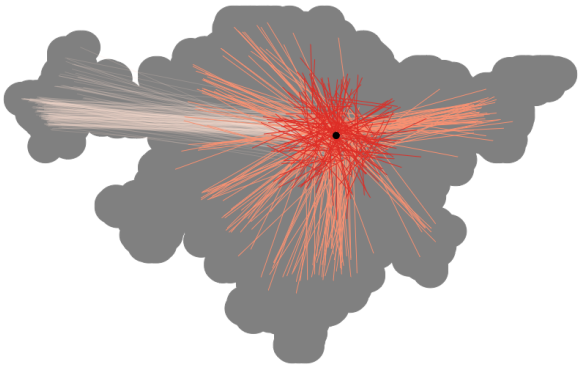
[6] 44% of the trips were queried in June and 1% were queried in November, with the remainder evenly split across the four months in between.

[7] We rely on the 2017 NHTS in the US.

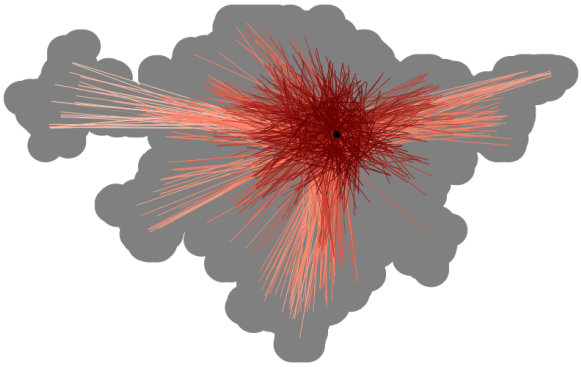
Figure A.1: Illustrations for the city of Jamnagar



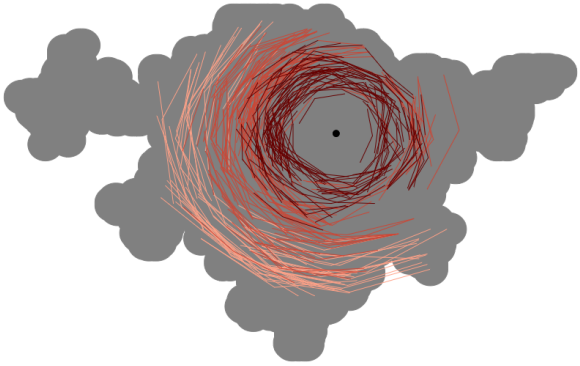
Panel A: Built-up and overall area



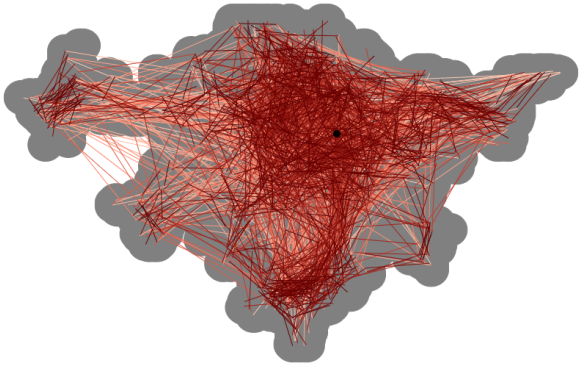
Panel B: Radial trips of absolute lengths
2 km, 5 km, 10 km, and 15 km from the center



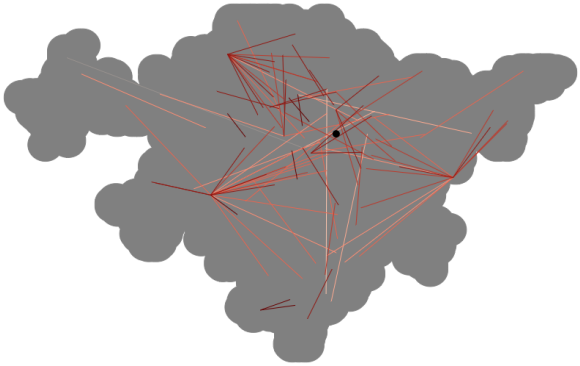
Panel C: Radial trips over uniformly
picked distance percentiles



Panel D: Circumferential trips around the center

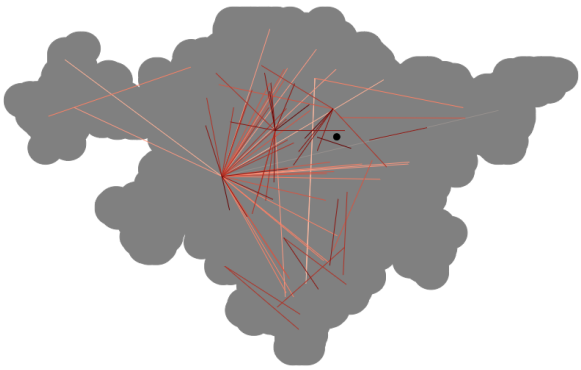


Panel E: Gravity trips

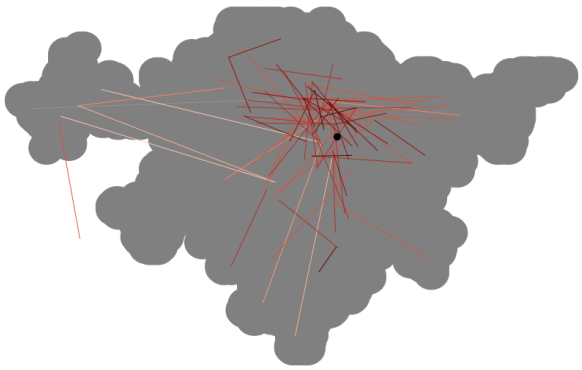


Panel F: School trips

Figure A.1 (continued): Illustrations for the city of Jamnagar



Panel G: Religion trips



Panel H: Healthcare trips

over-sampled sparse overnight periods to make sure that the quietest hours of the day get no fewer than one-fifth of the number of queries at the busiest hour of the day. In particular, at any hour of the day, we tried to query the following share of the day's total queries: 9 PM – 6 AM: 1.6%, 6 AM: 4.1%, 7 AM: 6.8%, 8 AM: 6.2%, 9 & 10 AM: 5.0%, 11 AM: 5.7%, 12 & 1 PM: 6.0%, 2 PM: 6.6%, 3 PM: 7.5%, 4 PM: 8.0%, 5 PM: 7.4%, 6 PM: 5.2%, 7 PM: 3.5%, 8 PM: 2.5%.

To be able to perform large numbers of queries on GM over short periods of time, we distributed the job across multiple processors that would query simultaneously. We automated the process such that at any time of a day, we activated as many processors as needed to satisfy the intended number of instances per trip as well as the planned hourly distribution of queries.

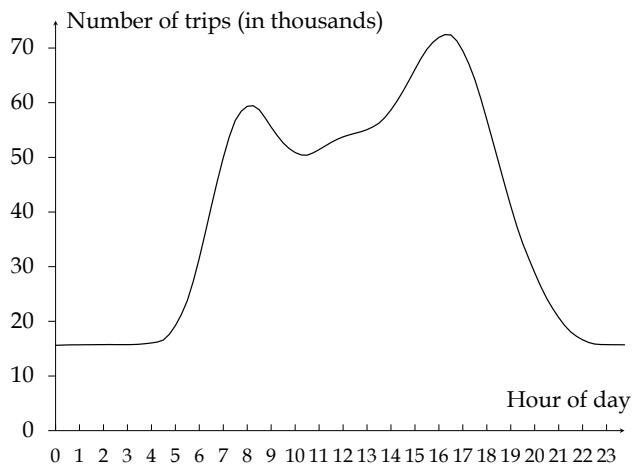
Panel A of figure A.2 shows the realized distribution of query times across hours of the day, and it very closely matches the planned distribution. The median trip was queried at 8 different 15-minute time slots of the day and 90% of the trips were queried at 4 or more time slots. The query times for each trip were spread out throughout the day: 88% of the trips were queried at least once between 6 AM and 6 PM and once between 6 PM and 6 AM.

We wanted to have an even spread of days and times across cities and trip types/strategies. So the order in which the trips were queried was randomized. Panel B of figure A.2 shows the stable realized proportion of trip types across hours of the day.

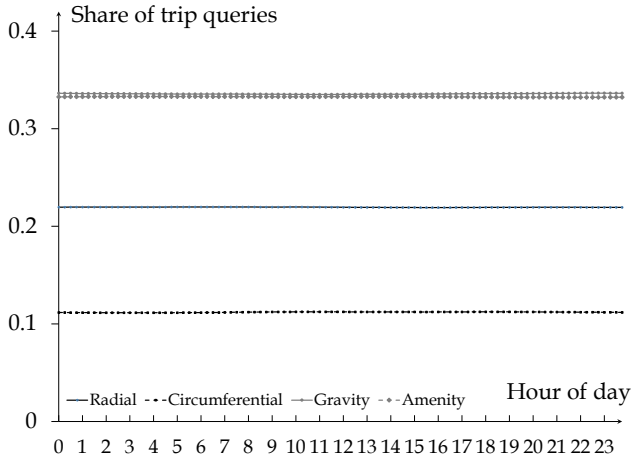
Walking and transit trips

We do not expect walking times for a given trip to vary by either the day or the hour of day. However, walking speeds do vary based on slope and the density of the network of streets

Figure A.2: Queries by time of the day



Panel A: Number of trip instances by time of day



Panel B: Proportion of types by time of day

and pedestrian paths. So, unlike for driving times, we aimed to query each trip only once for walking times. We also try to query each trip via walking at the same departure time as one of the trip's driving instances, while making sure the overall distribution of departure times via walking would resemble the distribution of departure times via driving (as in Figure A.2). We executed the walking queries between October 9 and December 15 of 2019 and, due to a technical problem, ended up querying more than we planned (2 instances per trip on average). GM returned at least one walking route for all but 89 (out of 2,735,442) trips.

GM does not generally track public transit in real time, but instead relies on fixed schedules that are periodically updated by transit authorities using the open-source General Transit Feed Specification. Thus, for any given trip, we do not expect any meaningful variation across weekdays in our travel times by transit. However, transit availability may differ significantly between day and night. So, we aimed to query two instances of each trip via transit: one during the day (7 AM - 7 PM), one at night (7 PM - 7 AM), both during weekdays (Monday-Friday). Conditional on day or night status, we draw departure times for each transit instance from the same distribution that we impose on the driving instances. Wherever possible, we also try to match the exact departure times of two of the trip's driving

instances.[8] We executed the transit queries between October 14 and December 21 of 2019, and due to a technical hiccup, ended up querying more than we planned (3 instances per trip on average).

There are several important caveats to the transit data. First, GM returned no routes on 17% of the queries and for 25% of the trips. Second, we do not expect the schedules to include informal transit providers, which own the large majority of India's bus fleet.[9] Third, some returned routes are implausible. Specifically, we exclude routes that (1) require walking all the way (69% of instances). Of instances that return a transit route, we also exclude those that (2) require waiting over an hour to start the trip (8% of day time instances and 39% of night time instances), or (3) are slower than their walking counterpart (38% of instances), which happens when GM uses inter-city rail, presumably because it is the only nearby transit alternative, to create highly convoluted itineraries. Following these exclusions, only 18% of our trips offer viable transit alternatives on at least one instance, and they are highly concentrated in the largest cities. In 159 of our 180 cities, less than 10% of trips are viable by transit. We cannot distinguish whether the absence of a viable transit route is due to limitations in the city's transit network or limitations in GM's coverage of the transit network. With that in mind, we report the 15 cities with at least half our trips covered by GM in Table A.1.

Appendix B. Other Data Sources

Open Street Map: Road network and water bodies data

Our measures of road network characteristics and water bodies come from OpenStreetMap (OSM; `https://openstreetmap.org`), a collaborative worldwide mapping project. We used OSMnx (Boeing, 2017) to download the OpenStreetMap network within each city (with boundaries as defined above) on two occasions for separate purposes. In May 2020, we downloaded the entire network of driving roads as a shapefile, along with information on road class. We use this network for the purpose of mapping our GM trip routes to OSM

[8] Because some driving trips are never queried at night, matching the departure times requires us to sometimes query both transit instances of a trip during the day. Overall, 57% of our transit trip instances took place during the day.

[9] See `https://data.gov.in/catalog/number-buses-owned-public-and-private-sectors-india`, last accessed, 6 September 2018.

Table A.1: Ranking of cities by transit network coverage

Rank	City	State	Coverage
1	Bangalore	Karnataka	0.90
2	Pune (Poona)	Maharashtra	0.88
3	Chennai (Madras)	Tamil Nadu	0.86
4	Ahmadabad	Gujarat	0.86
5	Mumbai (Bombay)	Maharashtra	0.78
6	Rajkot	Gujarat	0.76
7	Coimbatore	Tamil Nadu	0.76
8	Surat	Gujarat	0.76
9	Hyderabad	Telangana	0.72
10	Mysore	Karnataka	0.72
11	Delhi	Delhi	0.67
12	Vadodara	Gujarat	0.64
13	Jaipur	Rajasthan	0.61
14	Lucknow	Uttar Pradesh	0.56
15	Bhiwandi	Maharashtra	0.54

Notes: Coverage refers to the share of trips with at least one instance with a viable transit route returned by GM.

road classes and counting intersections along the route. In December 2020, we downloaded summary statistics on the total road network in each city from OSMnx (intersections count and angle distribution, total road length by type, grid-like road network, road elevation, grade and bearing distribution, and river and coastline length). We use these measures to construct city-level covariates.

Google Places: Establishment data

We collected data on the location and characteristics of Google Places establishments by searching trip destinations (e.g. banks) on GM. For each search, we specify a keyword (e.g. "bank"), a pair of coordinates to center the search on, and a resolution zoom level. GM returns up to 20 establishments on each search. Our objective is to collect every establishment on Google Places that is a potential trip destination in each of our cities in India. Couture (2016) shows that a similar methodology returns almost all restaurants in the United States.

Keywords We divided our search keywords into 14 trip purpose categories, based on definitions created by the NHTS: work, school, exercise, transport, childcare, adultcare, goods, services, meals, errands, healthcare, religion, other. Within these 14 categories, we identified

Table B.1: Keywords to search within each trip purpose category

Category	Keywords (in order)
work	'office'
school	'school'
exercise	'gym', 'sport complex'
transport	'transit station', 'bus station', 'train station', 'subway station', 'airport', 'taxi stand'
childcare	'day care center', 'child care', 'nursery school', 'preschool'
adultcare	'adult day care center', 'assisted living facility'
goods	'supermarket', 'store', 'bakery', 'book store', 'car dealer', 'clothing store', 'convenience store', 'department store', 'electronics store', 'florist', 'furniture store', 'gas station', 'hardware store', 'home goods store', 'jewelry store', 'liquor store', 'pet store', 'shoe store', 'shopping mall'
services	'bank', 'auto repair shop', 'hair salon', 'beauty salon', 'dry cleaner', 'pet care', 'accounting', 'ATM', 'bicycle store', 'car rental', 'car repair', 'car wash', 'electrician', 'funeral home', 'hair care', 'laundry', 'lawyer', 'insurance agency', 'locksmith', 'movie rental', 'moving company', 'painter', 'pharmacy', 'plumber', 'real estate agency', 'roofing contractor', 'travel agency', 'veterinary care'
meals	'Restaurant', 'cafe', 'meal delivery', 'meal takeaway'
errands	'police station', 'post office', 'library', 'embassy', 'city hall', 'courthouse', 'fire station', 'local government office', 'police'
recreation	'bar', 'theater', 'movie theater', 'museum', 'art gallery', 'park', 'amusement park', 'aquarium', 'bowling alley', 'campground', 'casino', 'night club', 'stadium', 'zoo'
healthcare	'hospital', 'doctor', 'dental clinic', 'medical clinic', 'medical center', 'therapy', 'dentist', 'physiotherapist', 'spa'
religion	'church', 'mosque', 'temple', 'community center', 'synagogue'
other	'cemetery', 'lodging', 'parking', 'RV park', 'storage'

109 keywords from Google Places' list of recognized 'Place Types'. Table B.1 shows all the keywords within each category.

Zoom level GM's establishment search algorithm returns establishment based on both proximity and prominence. In general, a high zoom level means that GM puts more weight on proximity than on prominence, which is desirable given our objective of collecting all establishments, not just prominent ones. So we set our zoom level high at "17z". For some sense of how high a zoom level this is, note that on a typical 19-inch monitor in the Chrome browser with the side panel minimized, this implies a coverage area of approximately one square kilometer, such that individual building outlines appear, even for buildings less than 5 meters wide, even in small cities.

Number of queries We seek to collect all establishments by making a sufficient number of queries at random locations within cities. For each category and city:

1. We pick the first keyword
2. We make X queries in random locations throughout the city. We set X to be proportional to $(\text{area})^{1/3}$ to make more queries in larger cities with more ground to cover.
3. We count the share Y of queries that return new establishments i.e. establishments that have not shown up on previous iterations.
4. If $Y \geq 10\%$, we return to step 2 and repeat
5. If $Y < 10\%$, we move to the next keyword and repeat from step 2. We stop when we have searched all keywords.

Location of queries For each query, we pick location coordinates randomly from within the built-up extent of the city. But since GM only returns up to 20 establishments on each query, we sample more in areas with high expected establishment density. In particular, we assign each pixel within the city a number of $D = u(10 + d - p)$ where d is the (normalized) distance to the city center, p is the (normalized) population density, and u is a draw from a uniform distribution between 0 and 1.[10] Then, pixels are drawn sequentially for querying in increasing order of D.

Weather data

Hourly weather data (rain, thunderstorm, temperature, humidity, wind speed, dewpoint, and atmospheric pressure) were obtained from Meteostat (`https://dev.meteostat.net/`, last accessed, 9 April 2021), which compiles its data from other organizations like the National Oceanic and Atmospheric Administration (NOAA), Deutscher Wetterdienst and Environment Canada. For each city, we identify up to three nearest weather stations within 40 kilometers of the city center and query the Meteostat API for their weather data at every hour of every day for which we collect travel times from GM. For cities with more than one nearby weather station, we use weighted averages of the weather variables where the weights are the inverse haversine distances between the station and the city center.

[10] Population density measures are based on data at the 1-kilometer grid level from GPW (CIESIN 2016).

We recovered weather data for 162 cities during the trips collection period. The median city-day has 8 weather readings and more than 90% of city-days have at least 2. On an average day, 24 (typically large) cities report weather every hour of the day. The number of weather readings per day for a given city varies little across days. The remaining 18 cities are missing data because Meteostat does not have a weather station within 40 kilometers of the city center. Over the five months when we collected weather data, it rained 6.5% of the time and there were thunderstorms 0.4% of the time.

Public holidays

According to the website `https://www.officeholidays.com/countries/india/2019`, there were 51 holidays falling on a weekday during our main data collection period between June and November 2019. Using descriptions provided by the data source, we retained 37 more important ones corresponding to religious festivals like Diwali (festival of lights celebrated in many parts of India) and major civic celebrations like Independence Day. Following a search of Google News we also identified 15 strikes in the largest Indian cities, including six that we deemed more important.

For the six national holidays during our data collection period, we estimated the same specifications as in column 7 of table 2, enriching it with an indicator variable corresponding to each of these public holidays. For five out of six of these public holidays we estimate a positive effect on speed, ranging from 0.5% to about 3%. 0.5% corresponds to a 'Saturday effect' as estimated in column 1 of table 2 while 3% is closer to a Sunday effect (4.5%) in the same specification. Interestingly, the only insignificant effect, a precisely estimated zero, is for Anant Chaturdashi, when only civil servants have the day off.

The remaining 31 public holidays are not celebrated everywhere. Separately for each state, we thus duplicate again the specification of column 7 of table 2 adding an indicator variable for each of 25 statewide celebrations. We have as many as eight religious celebrations and civic holidays in Odisha and as little as one in Punjab. Across 21 separate regressions, we estimate 81 coefficients for state public holidays. Of these 53 are positive and significant, 20 are negative and significant, and 8 are insignificant. Positive and significant coefficients range from slightly above zero like a Saturday to 5 or 6%, slightly above the faster travel speed enjoyed on Sundays. Higher coefficient (in the 10-15% faster range) are estimated only for the Kut festival in Manipur and the major festivals of Diwali and Durga Purja in Assam. The negative coefficients we estimate are mostly in the -0.5% to -3.5% slower

than a standard week day. Out of 20 negative coefficients, 11 pertain to the last day or two of the Ganesh Chaturthi festival. The closure of this major Hindu celebration honoring the deity Ganesh is in many parts of India associated with large processions. Another three negative coefficients are associated with the Ayudha Purja celebration, including in Maharashtra where people decorate their vehicles and exhibit them. A further three are associated with state election days.

Turning to strikes, we estimate 2% slower traffic on 22 October 2019 during the national strike for banks. We also estimate indicators for days when large strikes took place in the four main cities of India: Delhi, Mumbai, Kolkata, and Bangalore. Out of 15 strike indicators we estimate in four separate regressions, eight are negative and significant and five are positive and significant. Interestingly, the largest positive coefficient we estimate is for Delhi during the 'public transport driver' strike on 19 November 2019. Road travel was about 4.5% faster than a normal weekday following the absence of taxis and rickshaws, while bus and subway drivers were not striking. On the other hand, the youth demonstrations following the global climate strike led to slower traffic in Mumbai and Kolkata on 20 September 2019.

2011 Census of India

We obtained boundaries of towns and villages (fourth level administrative units) from the World Bank's South Asia Spatial Database (Li, Rama, Galdo, and Pinto, 2015). Other information is from the Census of India website. Specifically, population and number of households by town and village is from the Primary Census Abstract Data Tables (`https://censusindia.gov.in/pca/pcadata/pca.html`). Share of population possessing a car or motorcycle by town and village is from Percentage of Households to Total Households by Amenities and Assets (`https://censusindia.gov.in/2011census/HLO/HL_PCA/Houselisting-housing-HLPCA.html`). Pucca and kutcha road length and count of streetlight poles, by town only, are from the District Census Hand Book, Town Amenities files (`https://censusindia.gov.in/2011census/dchb/DCHB.html`). Average commute distance for urban non-agricultural workers by mode by district, are from Table B-28, 'Other Workers' By Distance From Residence To Place Of Work And Mode Of Travel To Place Of Work - 2011. (`https://censusindia.gov.in/2011census/B-series/B_28.html`).

National Sample Survey

Daily labor earnings by district and sector are from the Employment and Unemployment Survey of the National Sample Survey (NSS-EUE) 2011–12. District-level aggregates, in current US dollars per day, are from the World Bank's South Asia Spatial Database (Li *et al.*, 2015).

Other sources

Source for 63% as the share of the urban population with smartphones in India in 2017: `https://www.statista.com/statistics/894084/india-urban-smartphone-penetration/` (accessed 24 February 2021).

Source for mobile network coverage: Collins Bartholomew Mobile Coverage Explorer `https://guides.library.upenn.edu/MobCovExp` (accessed 8 July 2020)

Source for public holidays in India: `https://www.officeholidays.com/countries/india/2019.php` (accessed 24 February 2021).

Night lights data from from DMSP satellite F-18 2013, based on Baugh *et al.* (2010), are available at `https://eogdata.mines.edu/products/dmsp/`. Image and data processing by NOAA's National Geophysical Data Center, based on data collected by US Air Force Weather Agency.

Within-city population density is from WorldPop (Bondarenko *et al.*, 2020) and Gridded Population of the World (GPW), version 4 (CIESIN 2016).

Appendix C. Variable Construction

A. Trip-level variables

Many variables in Table 2 require dividing a trip into a set of segments. We define these segments based on GM directions. For instance, the first segment of a trip begins at the trip origin, and ends at the location where GM provide the first new direction (e.g. "turn left"). Every location where GM provides a new direction corresponds the end of the previous trip segment and the beginning of a new segment. Together, the set of trip segments characterize the trip route.

Distance to city center The city center is defined in Appendix A. We define a trip's distance to the city center as its average distance to the city center, integrated along a straight line

path from the trip's origin to the trip's destination. In about 0.01% of cases where the integral cannot be calculated due to rounding in the log term, we use an average of origin and destination distance to center.

Upward and downward gradient GM only returns the change in elevation on a trip segment when requesting walking time (and biking time, which is not available in India). To ensure that this "walking" query takes the same route as the "driving" trip in our sample, we input the end of each segment on that route as a throughpoint in our GM walking request.[11] For each of these walking requests, GM returns segment distance in meters and change in elevation in meters. To compute the upward gradient measure, we only keep segments with a positive change in elevation. For these segments, we divide the sum of the change in elevation by the sum of distance.[12] We compute the downward gradient measure in a similar way.

Share of trip on each road class Each edge in the Open Street Maps network receives a tag which characterizes its road class. We follow the OSM road classification and divide roads into six classes:

1. Motorways: Expressways consisting of restricted access dual carriage ways with 2 or more lanes in each direction with limited access via interchange. We also include the less frequent OSM type "trunks" in the motorways category. Trunks are national highways connecting major cities.[13]
2. Primary Roads: State highways linking major population centers.
3. Secondary Roads: District roads linking smaller population centers, or major through routes within city limits.
4. Tertiary Roads: Other roads linking towns and villages, important routes linking different localities within a city.
5. Residential Roads: Roads used for local traffic, not major through routes.

[11] GM accepts a maximum of eight throughpoints in addition to trip origin and destination, so we implement the same throughpoint selection algorithm described in Appendix D to replicate Intents trips. The algorithm described in section Appendix D uses smartphone "pings" as candidate throughpoints. Here we instead use the end of each trip segment as candidate throughpoints.

[12] Some changes in elevation are implausibly large. We winsorize change in elevation at 75 meters per segment, corresponding to the 90^{th} percentile of the distribution of maximum segment change in elevation across trips.

[13] These definitions would be roughly similar in countries other than India, and for instance motorways are equivalent to freeways in the United States.

6. Other: Includes roads tagged as "Unclassified" (least important roads in a network) or "Road" (roads whose class is unidentified by OSM).[14]

We assign a road class to each GM trip segment as follow. First, we create a buffer zone with a 2 meter radius around the two endpoints of each segment. Then, if the buffers at both endpoints of the segment intersect with an OSM edge of the same road class, we assign the segment to that road class.[15] If the two endpoints of a segment intersect with two edges of different road classes, we set that segment to missing.

We then take any unassigned segments and repeat the above steps after increasing the buffer size from 2 to 5 meters. We then iterate again with a 10, 20, and 30 meter buffer. We classify a segment as having "missing" road class if the algorithm above does not successfully assign it a road class.

Below we compute the length, in kilometers, of the different classes of roads in each city. We find that motorways account for 3.0% of the roadway in Indian cities, primary roads account for 2.9%, secondary roads account for 4.0% and tertiary roads account for 8.3%. Hence the four most important road classes account for less than 20% of the roadway. However these road classes account for nearly 70% of travel in our simulated trips. More precisely, we can compute the share of a given trip on a given road class as the distance-weighted share of its segments that have been assigned a given road class. We end up with 22% of trip distance on motorways, 18% on primary roads, 14% on secondary roads, 14% on tertiary roads, 10% on residential roads, 5% on other roads, and 19% missing.

Number of Intersections We compute the number of intersections for each trip as the number of OSM nodes within 20 meters of the trip route. Recall that the trip route is a set of segments from GM, where each segment is a straight line connecting its two endpoints. An OSM node is where two edges intersect, so the intersection of a single carriage with a dual carriage road would have two nodes.

Number of Right Turns We compute the number of right turns on a trip using the same GM directions that we used to define segments. Specifically, we consider a step direction to be a turn if it contains the strings "TURN", "SHARP RIGHT", "SHARP LEFT", "TAKE

[14]We ignore service roads (designed to access a venue) as well as all special road tags (e.g. pedestrian roads, cycleways, etc). However, we include all the "link" tags in their corresponding road class, for instance we classify a "motorway_link" tag (e.g. a highway ramp) as a motorway.

[15]If only one point (beginning or end) intersects with an OSM edge of a given road class, and the other point does not intersect anything, we assign that road class to the segment.

THE RAMP ON THE RIGHT", or "TAKE THE RAMP ON THE LEFT". We consider a step direction to be a right turn if it is a turn and it additionally contains the word "RIGHT". Right turns are against traffic in India as driving is on the left.

Number of Google Maps establishment We compute the number of Google establishments for each trip as the number of establishments in our Google Places sample located within 20 meters of any segment along the trip route.

B. City-level variables

Area Area within the city boundary as defined in the main text, in square kilometers.

Elevation variance Specifically, this is the variation of elevation across nodes in the OSM road network. Node elevations are queried using the GM API accessed via OSMNX.

Road length Each edge in the OSM network receives a tag which characterizes its road class. The four road classes that we include in our measures of city road length (motorways, primary, secondary, tertiary) are, as described above, the most important classes of roads. For each class we measure total road length in kilometers.[16]

We note that certain cities have incomplete street networks on OSM. For instance, some cities have a large share of missing OSM roads in our mapping of GM trip route to the OSM road network described above. Even in these cities however, visual comparison of the OSM road network with the GM road network (or satellite images) suggests that motorways, primary, and secondary roads are never missing, while tertiary roads, which also enter our "major roads" measure, are only infrequently missing. The coefficient on "major roads" in Table I.1 is still significant when we exclude tertiary roads, but smaller by 30 to 50 percent.

Grid-like road network OSMNX calculates the compass bearing ("bearing" for short) from each directed edge's origin node to its destination node. The bearing captures the orientation of the edge relative to north. We use a city's distribution of edge bearings to characterize how 'grid-like' its road network is in two separate ways: 'orientation' which captures the share of edges conforming to the network's main grid orientation, and 'Gini' which captures the dispersion in the distribution of edge bearings.

[16] In the OSM network, both carriage ways of a motorway count as separate edges (i.e. once in each direction). We experimented with counting dual carriage ways only once when measuring length, and also with measuring lane-kilometers, instead of just edge kilometers. These adjustments generate measures of length by road class that are very highly correlated with our preferred measure.

Table C.1: Summary statistics for Indian cities

	Mean	St. dev.	Min.	Max.
Population ('000, Census/our delineation)	1,220	2,848	63	22,683
Log population (Census/our delineation)	6.41	0.97	4.14	10.0
Population ('000; UN, 2018)	1,486	3,108	306	28,514
Population growth 1990-2018 (%)	135	113	22	944
Area (km^2)	246	469	19	5,068
Log area	4.97	0.91	2.96	8.53
Total roads length (km)	1,789	3591	25	33,730
Major roads (km)	326	670	8	6,136
Share households with car/jeep/van (%)	9.25	5.29	0.75	31.8
Share households with scooter/motorcycle/moped (%)	38.8	12.7	3.30	71.9
Mean daily earnings ($)	5.00	1.93	2.20	15.1
Elevation variance	232	781	1.53	9411
Water length (km)	99.8	188	0	1526
Lights per km	35.9	69.9	0.048	712
Network	0.151	0.054	0.064	0.397

Notes: Cross-city averages not weighted by population. 180 cities. Major roads include motorways, primary roads, secondary roads, and tertiary roads. Earnings are for industrial workers. For population, the two sources differ both because of the target year and because they are based on different boundaries. See below for the computation of the network variable.

Orientation. A grid is a set of roads intersecting at perpendicular angles. In a grid all bearings are either perpendicular or parallel to each other. The orientation grid metric measures the proportion of edges in a city's road network that conform to the dominant grid orientation in that they are perpendicular or parallel to the modal edge bearing.

Let g index each edge in the road network of city c, and let x_{cg} be the edge bearing rounded to the nearest degree, and x_c^{modal} be the modal edge bearing modulo 90 of city c. For example, if a city's grid were oriented N-E-S-W, then x_c^{modal} would equal 0. Let $\delta_{gc,x_c^{modal},\nu}$ be an indicator for whether edge g in city c conforms to grid orientation x_c^{modal} within a bandwidth error of ν:

$$\delta_{ic,x_c^{modal},\nu} = \begin{cases} 1 & \text{if } (x_g - x_{0ct}) \mod 90 <= \nu \\ 1 & \text{if } (x_g - x_{0ct}) \mod 90 >= (90 - \nu) \\ 0 & \text{else.} \end{cases} \tag{C1}$$

We then compute our grid-like measure as:

$$Orientation_c = \frac{\sum_{g \in I_c} \delta_{gc,x_c^{modal},\nu}}{Q_c}, \tag{C2}$$

Figure C.1: Most and least grid-like city road network using orientation grid metric

Panel A: Chandigarh - Grid Score = 0.4

Panel B: Shillong - Grid Score = 0.06

Maps of the OSM road networks, for Chandigarh, the most grid-like city, and Shillong, the least-grid-like city. We measure how grid-like a city is based on the share of edges in its road network conforming to the grid's main orientation, i.e., whose compass bearings are within 2 degrees of the modulo 90 modal bearing in the network.

where I_c is the set of all edges in city c, and Q_c is the number of edges in I_c.

In the paper, we report results using a narrow error bandwidth of $\nu = 2°$. We experimented with a wider bandwidth of 5°. We also experimented with allowing for more than one dominant grid orientation, because for instance larger cities can have smaller sub-grids whose orientation differs from that of the main grid.[17] These variations produce highly correlated rankings of cities, and we therefore prefer the simplest version above. Visual inspection suggests that our methodology performs well at ranking road networks by how grid-like they are. Figure C.1 shows the most and least grid-like cities according to the orientation metric, side-by-side.[18]

Gini. We modify the definition of the Gini index for income inequality to measure the normalized dispersion of edge bearings. For each city c, we define 360 different possible bearings, indexed by k, and ranked by their frequency such that $k = 1$ is the least frequent bearing and $k = 360$ is the most frequent bearing. In a perfectly gridded city, the four most frequent bearings, spaced 90 degrees apart, would account for 100% of edge bearings. Therefore, we can interpret high values of the following Gini index as corresponding to

[17]We also experimented with weighting edges by length, but such measures appear to overestimate how grid-like small cities with few very long roads are.

[18]We noted above that smaller road classes are missing in some cities, which could bias our grid measure if certain types of road are more grid-like. Our results are robust to limiting the sample to the subset of cities for which we have a more complete road network. It would be possible to compute measures of how grid-like the road network is separately for different types of road defined above, instead of only for the total road network. However, visual inspection suggests that these measures do not perform well at capturing overall how grid-like cities are. For instance motorways are often curved and outside of the main grid.

cities with a more grid-like network:

$$Gini_c = \frac{Q_c \times 360 - 2\sum_{k=1}^{360}\sum_{l=1}^{k}\theta_{cl}}{Q_c \times 360}, \tag{C3}$$

where θ_{cl} is the number of edges in city c with bearing l.

The assumption of 360 possible distinct bearings is arbitrary, and we also computed Gini indices after rounding up each bearing to the nearest even degree (i.e. by assuming 180 possible bearings.) We also experimented with defining modulo 90 bearings (instead of modulo 360 as above).[19] These variations produce Gini indices that are highly correlated with the index defined above that we use in the paper.

Water length We calculate the length of river centerlines, coastlines, and lakeshores within each city from OSM.

Population Data at the town and village (fourth administrative) level are from the 2011 Census. City values are area-weighted sums of the town and village values. In other words, we first sum populations from all towns/villages falling entirely within the city boundary. For towns/villages that cross the city boundary, we add the a share of the town's population equal to the share of the town/village's area falling within the city.

Share of car and motorcycle owners Data on the share of households that possess a "Car/Jeep/Van" and the share that own a "Scooter/Motor Cycle/Moped" at the town and village (fourth administrative) level are from the 2011 Census. City-level values are weighted averages of these town/village values, using each town/village's share of households in the city, and pro-rating towns/villages that fall partially within the city as described in the calculation of population (above).

Mean daily earnings Cities that fall within a single district are assigned the district's value. Cities that cross a district boundary are assigned a population-weighted average analogous to the car share described above.

Street lights "Electricity-Road Lighting Connection (Numbers)", a count of street lights, is available at the town level in the 2011 Census. We sum counts for all towns within the city. For towns straddling the city boundary, we add a fraction of the count equal to the share of the town's land area falling in the city.

[19]For some smaller cities with sparser road networks, the number of distinct edge bearings is less than 360. In these cases, we adjust the calculation to consider only the total set of bearings present in that city, which may be less than 360.

Share of paved roads Length of paved (Pucca) and dirt (Kutcha) roads at the town level are from the 2011 Census. We aggregate to the city level analogously to streetlights. Share paved is length paved divided by total (paved+dirt) length.

Potholes We obtain data on potholes from an app designed by Intents Mobi (`https://intents.mobi/`). In Appendix D, we provide more details on how Intents recruits drivers and collects geolocation data (we also use an app from Intents to compare GM trips with actual trips, and we describe this exercise at length). Intents' app identifies a candidate pothole as an event of sudden acceleration in both horizontal and vertical movement. At least three such events in the same location identify a pothole. The dataset that we obtain from Intents contains the location of all potholes, and pings that geolocate drivers and allow us to measure their distance traveled in kilometers. Our final dataset has 23,512 potholes and 1.03 million kilometers of travel, collected from September 9 2019 to December 18 2019 in 131 cities.

In each city, we compute potholes per kilometers as the average number of potholes encountered per kilometers driven in a city.[20] Over our entire sample, the average number of potholes per kilometers driven is 0.023. Some cities have very few kilometers driven, so we removed all cities where the standard error of our measure of potholes per kilometers is above a cut-off of 0.05.[21] Using that metric, 87 cities remain in our sample. Our results were not sensitive to varying this cut-off.

Night lights Night lights is the sum of all pixels' digital number for 2013 within city boundaries.

Spatial Gini of population The spatial Gini of population for city c is defined as

$$PopGini_c = \frac{\sum_{i=1}^{N_c}\sum_{j=1}^{N_c}|P_{ic} - P_{jc}|}{2N_c^2\overline{P_c}}, \tag{C4}$$

where the population of pixel i in city c is P_{ic}, N_c is the number of pixels in city c, and $\overline{P_c}$ is the mean value of all pixels in c. Population of each 100-meter pixel is from WorldPop.

[20]Note that our measure is per kilometers driven in the Intents data, not per overall kilometers of roads available in a city. So a pothole on a frequently used road contributes more to our measure than a pothole on a rarely used road.

[21]To compute this standard error, we assume that encountering a pothole on any given kilometer is a binomial event that happens with a constant probability $p = 0.023$ (computed across all cities). We can then approximate the standard error of p in a given city as $p(1-p) \times$ (km driven).

Appendix D. Using Intents app data to evaluate Google Maps accuracy

Intents trip data To evaluate the accuracy of GM speed data, we collect actual trip data in Indian cities using a smartphone app. Our app was designed specifically for our project by Intents Mobi (`https://intents.mobi/`), a mobile app developer with experience creating geolocation apps for the Indian market.

We collected data from September 9th to December 18th 2019. Until October 2019 our app was available in 20 cities representative of the size and regional distribution of cities within the 11 Hindi-speaking majority states (the app instructions are in English and Hindi). App usage was geofenced (restricted) to within 50 kilometers of the center of large cities and 25 kilometers of the center of small and medium size cities. In late October 2019 however, at Intents' request we extended app availability to all of India, because recruitment was being hindered by low ratings our app was receiving from prospective users unable to use it in their city. App users are paid per kilometer driven in a motor vehicle. On average we paid 1 INR per kilometer on daytime trips and 2 INR (1.5 to about 3 cents) per kilometer at night (11 PM to 5 AM), but Intents varied rates across cities, time of day, and day of week to achieve targets of 15,000 kilometers traveled per city, and 10% of night trips from 11 PM to 5 AM. In effect this meant higher rates in smaller cities at night.[22] We did not collect any information about these drivers, but Intents recruited drivers through online advertisements targeted at frequent drivers, like Uber or Ola drivers.

Drivers press a button when they start and finish a trip. For the entire duration of the trip, every second the app records a "ping" consisting of the phone's current position and time. We collect these pings as received by Intents' servers, as well as individual trip ID numbers constructed by Intents.[23] In total, we received 65,881,813 million pings in 249,938 trips. After eliminating trips not in one of our cities (13% of trips), trips with a straight-line distance from origin to destination of 0 (11%), trips shorter than 0.2 kilometer (18%), trips longer than 100 kilometers (0.2%), trips slower than 2.5 km/h (8.9%), trips faster than 100 km/h (1.0%), and weekend trips (24%), our clean weekday sample of trips is 90,894 trips

[22] Our instructions were simple and made to fit on a single screen, written in Hindi and English. Users had to be car drivers, click on a button at the beginning and end of each trip, and sign off when not in the car. Rates and limits were clear, with 6 hours maximum paid per day, and a maximum of 50 kilometers of the higher night rate.

[23] An app user may forget to begin and end trips within the app. In that case, Intents ends the trip if the vehicle has not changed position for more than 10 minutes (the trip ends at the time of first ping in that final location). Also, sequences of pings sometimes fail to be recorded on the provider's server. In our analysis sample, the maximum gap between pings is less than one minute for more than 75% of our trips.

Table D.1: Cities with most Intents trips

City	# Intents trips	Population rank
Delhi	35,525	1
Udaipur	7,289	103
Jaipur	6,660	10
Lucknow	6,094	11
Chandigarh	4,920	55
Nellore	4,530	74
Indore	3,920	16
Bangalore	2,823	4
Chennai (Madras)	2,804	5
Bhopal	2,309	22

Note: Number of Intents trips in cleaned sample for the 10 cities with the most Intents trips. Population rank is among all cities in our sample based on 2018 UN WUP population.

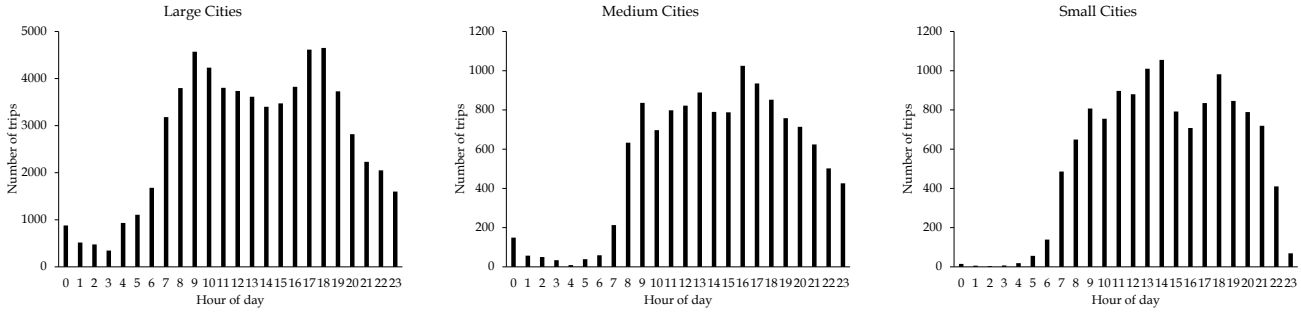
in 89 cities. Table D.1 shows the 10 cities with the most Intents trips in the Intents sample, with city name, number of Intents trips, and city population rank.[24] Overall, we have one city (Delhi) with a large sample of 35,500 trips, and 13 cities with smaller samples between 1,000 and 7,500 trips. These samples are too small for city-specific analysis, so we divide our sample into three different city-size bins, large cities with population rank 1 to 20 (65,260 trips), medium size cities with rank 21 to 60 (12,699 trips), and small cities with rank over 60 (12,935 trips.)

Figure D.1 shows the number of trips at each hour of the day in large, medium and small cities. Night time trips in small cities are exceedingly rare, despite the higher pay rate that we offered to incentivize them. In small cities, we have only 17 trips taken during the three hours from 1 AM to 4 AM, representing about 0.1 percent of trips in these cities. In the analysis that follows, this prevents us from measuring average speed at night in small cities with precision.

Google Maps replication of Intents trips To assess the accuracy of GM travel speed, we attempted to replicate each Intents trip on GM as faithfully as possible. For each Intents trip,

[24]Only five of these top ten cities (Delhi, Udaipur, Lucknow, Chandigarh, and Bhopal) are in the 20 that we had initially selected for data collection, with an additional city (Indore) being in the in a list of 10 alternative cities that we provided Intents with in case data collection was difficult in our preferred cities. Overall, we achieved only modest success at restricting our sample to a chosen set of cities.

Figure D.1: Number of Intents trips by hour of day



Total number of Intents trips per hour of the day by city size bins, in cleaned sample. Large cities have population rank from 1 to 20, medium cities from 21 to 60, and small cities from 61 and up.

we query a matching trip on GM that takes the same route, and starts within 5 minutes of the Intents start time.

To characterize the exact route taken by each Intents trip, we divide each trip at "throughpoints". The throughpoints are locations ("pings") where we observe the driver along the route between origin and destination. We can input these throughpoints into GM, to ensure that GM returns travel time for a route that is as similar as possible to that of the matching Intents trip. However, GM accepts a maximum of 8 throughpoints, so to decide which Intents pings to use as throughpoints, we implement the following algorithm:

1. We define $N = min(2D - 1, 8)$, where D is trip length in kilometers, as the maximum number of throughpoints on a trip. As an example, if a trip is 2 kilometers, $N = 3$, or three throughpoints dividing the trip into four segments of 0.5 kilometer on average.

2. The next step is to decide which pings to define as throughpoints. We aim to have throughpoints at roughly equal distance from one another along the route of the trip. If there are fewer than N pings, then each of them is a throughpoint. Otherwise, we divide the route of each trip into $N + 1$ subtrips of equal length and define the N endpoints of these subtrips as candidate throughpoints. For each candidate throughpoint, we select the nearest ping along the route as the actual throughpoint.[25]

3. We then eliminate throughpoints that are redundant, in the sense of falling near a straight line between the previous and following throughpoint. Specifically, we

[25]In rare cases where ping frequency is low and the route from one ping to its closest neighboring ping in one directions includes an entire subtrip, then we designate this isolated ping as a throughpoint. Also, this approach can generate fewer than N throughpoints, usually when pings are bunched at one end of the trip or the other. In that case, no information is lost by having fewer than N throughpoints.

eliminate any throughpoint x located between two throughpoint y and z where the straightline distance yz is within 10% of the sum of distances yx and xz. This is because GM often responds to such redundant throughpoints inappropriately, for example with an extra loop around the block.

Sometimes, our GM query fails, in the sense that it doesn't return a trip with a route similar to that of the Intents trip.[26] This happens when we have very few Intents pings with which to establish a route, or when these pings form a pattern that is too convoluted. In the analysis that follows, we eliminate the 33 percent of trips for which there is a more than 20 percent difference between the Intents distance and its GM equivalent.

Comparing Intents and Google Maps speeds We now compare trip speed across three different samples: Intents trips, the GM replication of Intents trips, and the GM trips simulated using the radial, circumferential, gravity, and amenity methodology described in the paper. For all computations that follow, we weight the simulated trips to ensure that every city has the same total weight in both the Intents and the simulated trip samples.

Panel A of figure D.2 plots average trip speed at each half hour of the day in large, medium, and small cities, for all three samples. In all panels, we show two standard error confidence bands around the estimates from the Intents trip sample.[27] We find that the average speed of Intents trips is nearly identical to that of their GM replication within every city size bin and hour of the day. We conclude that on average, GM and Intents speed are similar when computed from similar trip samples. However, average speed appears lower in our overall GM sample of simulated trips, especially at night.

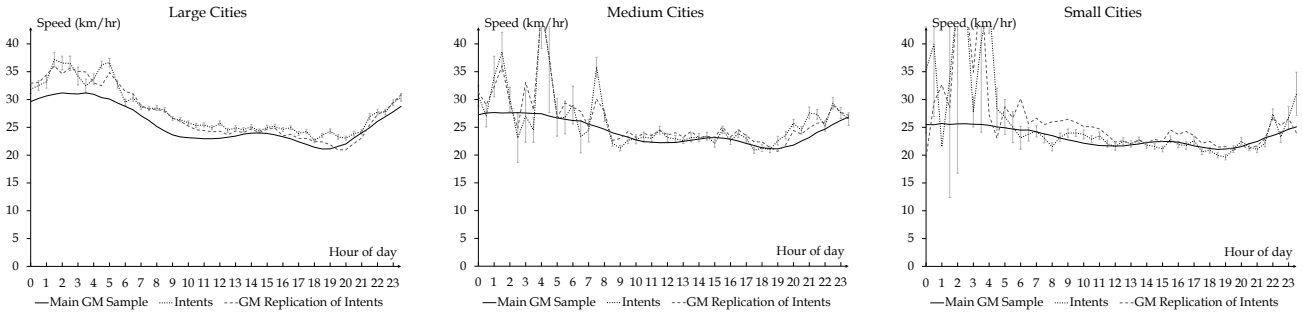
This speed difference is not surprising, because many factors that affect driving speed systematically differ between our Intents and simulated trip samples. In section 4, we showed that trip length, and distance to the city center, are important determinants of trips speed. So in panel B of Figure D.2, we show estimated time effects from a regression of log travel speed on log trip length, log distance to city center, and time effects.[28] Overall, the time effects for all three samples align remarkably well. The better fit relative to panel A is due to the trip length control. In the Intents sample, night time trips are almost twice

[26]Our query also fails if GM rejects a throughpoint, which happens for 2 percent of trips.

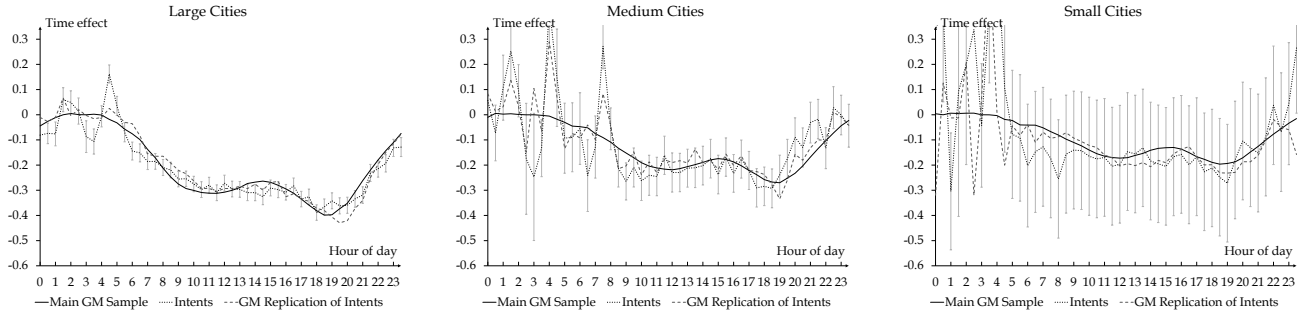
[27]Confidence bands for the main GM simulated sample (not shown) are very small, and confidence bands for the GM replication of Intents trip (not shown) are generally similar to those shown for Intents trips.

[28]We normalize the time effects in the GM simulated sample to equal zero at 3 AM. We then normalize the time effects within each sample so that all three curves take the same value at 5 PM.

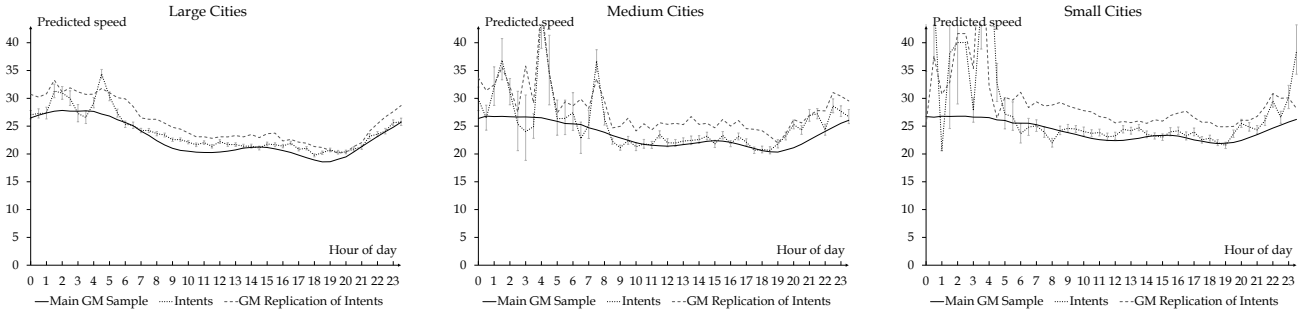
Figure D.2: Trip sample speed comparisons



Panel A: Average trip speed



Panel B: Estimated time effects



Panel C: Fitted speed

The figure compares speed at each half hour of the day, for our main GM simulated trip sample (solid line), the Intents trip sample (dotted line), and the GM replication of Intents trips (dashed line). Panel A shows average speed within each half hour. Panel B shows estimated time effects in a regression of log travel speed on log trip length, log distance to city center, and time effects. Panel C shows fitted speed values from the same Panel B regression. All confidence bands shown are two standard errors. Confidence bands for the main GM simulated sample (not shown) are very small, and confidence bands for the GM replication of Intents trip (not shown) are generally similar to those shown for Intents trips. All averages and regressions for the main GM simulated sample are weighted to match the distribution of trips across cities in the Intents sample.

as long as day time trips, so they appear faster in the absence of a trip length control. In our simulated sample, trip length is constant across time of day by construction. This demonstrates that with appropriate trip-level controls, our simulated GM sample replicates congestion patterns from actual trips across cities of different sizes. For instance, averaging the time effects in Panel B suggests that as reported in the main text, the speed difference

between night time (11 PM to 6 AM) and peak time (10 AM to 2 PM and 4:30 PM to 9 PM) trips in large cities is 29.4% in our main simulated GM sample, 28.0% in the Intents data, and 31.7% in the GM replication of Intent trips. In medium sized cities, the night time vs peak time differences in these three samples is 20.5%, 21.9%, and 21.9%.

Finally, in panel C of figure D.2 we compare the average speed in our three samples, computed as a fitted value from the same regression used to obtain the time effects. We compute all fitted values at the average trip length and average distance to the city center from the full simulated GM sample. We find that average fitted speed is more similar across samples than mean speed, but speeds from Intents trips and their GM replication still appear somewhat faster than our simulated trips. We do not find this remaining difference concerning given how special the Intents trip sample is.[29] The Intents sample provides actual trip speed, but we are not confident that Intents' trip ID accurately defines discrete trips that individuals take. As documented above, the raw Intents sample had a significant share of extremely short trips, and of trips whose route is so convoluted as to not be replicable by GM. Our cleaning procedures reduce the impact of such problems, but they cannot entirely resolve them. Another issue with panel C is that the three fitted speed curves shift up or down relative to one another depending on the trip length used to compute the fitted value, because the estimated coefficient on trip length in the speed regression varies across samples. In particular, the estimated impact of trip length on speed is smallest in the Intents sample, likely due to measurement error in trip length.

Overall, this exercise shows that GM can replicate actual speed even in smaller Indian cities. In addition, congestion patterns from a sample of actual trips resemble those from our sample of simulated GM trips, after appropriately controlling for trip characteristics.

Appendix E. Comparison with Uber Movement data

We use data from Uber Movement (`https://movement.uber.com`) based on actual trips taken by Uber riders for five Indian cities, Mumbai, Delhi, Hyderabad, Bangalore, and Kolkata. Uber provides data by multi-hour period by day, or alternatively, by individual hour by calendar quarter. Given the importance of within-day variation in speed, we use the latter for the last three quarters of 2019 when we collected our main data. Uber Movement divides each city into small zones and provides a travel time between zone pairs for each hour of

[29] One explanation for faster Intents trips is that Uber and Ola drivers limit their driving to faster roads.

the day. For example, in Bangalore, there are 198 zones, averaging between three and four square kilometers each, and Uber Movement reports travel times for 93.7% of all $198 \times 197 \times 24 = 936,342$ zone pair-hours.[30]

If these were average speeds of actual trips between these zone pairs, we could compare them directly to our GM trips. However, they differ from real trips in a critical way. Uber Movement computes zone-pair travel times using all trips that pass through a pair of zones. Because these zones are fairly small, we expect a typical Uber trip to generate information for many pairs. For example, a real trip that passes through ten zones will provide travel time information for 10×(10-1)/2 = 45 different pairs. We call these pairs trip segments. These segments will contain a different distribution of zones than the actual trip. For example, the beginning and end zones represent 20% of the zones in an actual ten-zone trip. But in the 45 trip segments, they represent only 9%. This is a problem because we expect that the beginning and end of a trip are on average slower than their middle, for the same reason that long trips are faster than short trips—drivers must use slow local roads at the beginning and end of many trips.

A second complication of comparing our GM data with the Uber Movement data is that there is no way of knowing the actual road distance traveled between two zones in an Uber Movement trip segment. Therefore, we measure the haversine distance between the centroids of the zones of the endpoints of either each trip instance (GM) or each trip segment (Uber Movement) and compute the corresponding effective speed.

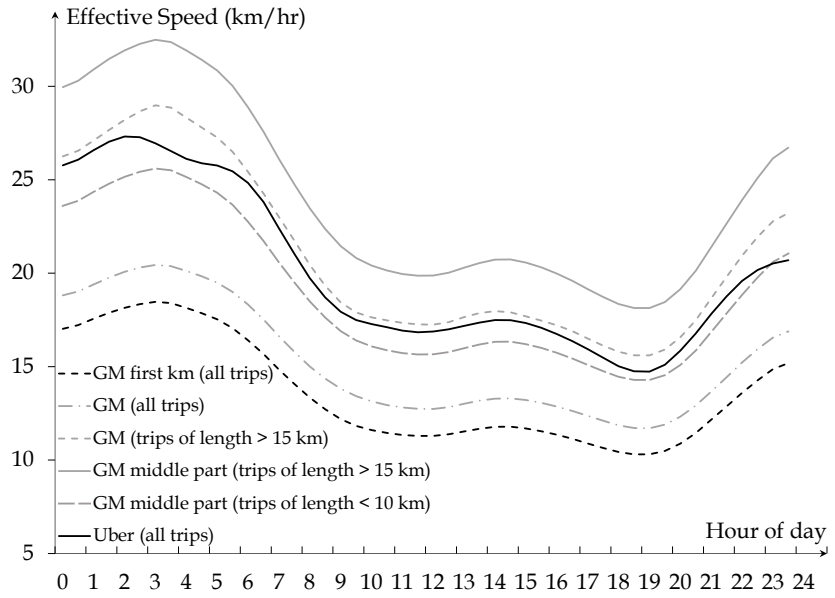
To make the samples comparable, we restrict attention to GM trip instances matched to a corresponding Uber Movement observation by zone of origin, zone of departure, and hour of departure. Uber Movement's effective definition of each city is generally modestly smaller than ours. In Bangalore, 91.5% of our 1,048,682 GM trip instances have both their origin and their destination in a zone reported by Uber Movement. Uber Movement provides matching travel times data (origin, destination, and hour) for 93.7% of these, or 899,100. Conversely, only 13.1% of zone-pair-hours reported by Uber Movement have a corresponding GM trip in our Bangalore sample.[31]

Consistent with our conjecture above, the effective travel speeds computed from Uber Movement travel times are much faster than the corresponding speeds computed from

[30]In Bangalore, these zones correspond to official city wards. We exclude the travel time information that Uber Movement reports for each zone to itself, which is provided only for the last hour of the day.

[31]Using instead all Uber Movement observations only makes minor differences. On the other hand, GM trip instances within the Uber-defined city limits is further from the center on average.

Figure E.1: Comparing Uber Movement with different parts of Google Maps trips



Pooled data for five cities (Mumbai, Delhi, Hyderabad, Bangalore, and Kolkata). Series computed as described in the text.

simulated trip instances obtained from GM. We plot these speeds for all hours of the day in figure E.1. For instance, at 6 AM, we observe an average effective speed of nearly 25 kilometers per hour for Uber Movement instead of about 18 kilometers per hour for GM. Note that we pool data for all five cities together but the patterns are the same for each city taken individually. Note also that we do not condition out trip characteristics as we do in the rest of the paper because some characteristics like trip type are missing from the Uber data while others like trip length cannot be directly compared across both groups.

Since Uber Movement oversamples the fast middle parts of trips as noted above, it is more appropriate to compare speeds computed from Uber Movement data with speeds computed from longer trips or from the middle part of GM trips instead of the entire trips. To compute speed for the middle part of trips, we proceed as follows. GM provides length and duration data for each step of a trip instance (that is, each portion of a trip between changes of direction) in meters and seconds. Unfortunately, step travel speed is not updated in real-time. It appears to reflect an average travel speed for that step.

We create step-specific speeds as follows. First, we compute the speed of segment n that

is part of trip i, Z_{in}. Segment n could be the first kilometer, or the middle (defined as all but the first two and last two kilometers). Because step travel speed is not available in real time, this speed is time invariant. We then compute the average speed of segment n relative to the average speed of an entire trip, $\frac{Z_n}{Z}$. Next, we compute speed during hour τ, S_τ, by averaging across all observed (real-time) speeds of whole trip instances during this hour. Finally, we compute speed for segment n at hour τ as $S_{n\tau} = \frac{Z_n}{Z} \times S_\tau$.

The resulting speed profiles by time of the day are plotted in figure E.1. While the first kilometer of trips is sizeably slower than their average speed, the middle part of trips is much faster.[32] Throughout the day, we find that the speeds we infer from Uber Movement travel times are marginally higher than the middle part of trips shorter than 10 kilometers. The speeds we infer from Uber Movement are also marginally lower than the speeds we estimate for the trip instances longer than 15 kilometers (and significantly lower than the middle part of these long trips, which are most likely to use fast roads). These findings are consistent with how Uber Movement measures travel times, which overweights the middle part of trips.

Appendix F. Alternative speed indices

In this appendix, we derive a number of alternative speed indices. Then, we describe how these alternative indices correlate with our benchmark index, the city-fixed effect in equation (2), using the specification from column (4) of Table 2.

A. Laspeyres and Paasche Indices

Travel conditions may vary across cities in ways that may not be well captured by our benchmark equation (2). For instance, Figure 2 suggests that peak hours are relatively slower and last longer in more congested cities. To capture this, we estimate a more flexible version of equation (2) where we allow both the intercept and the vector of coefficients to vary across cities:

$$\log S_i = \alpha_{c(i)} X_i' + s_{c(i)} + \epsilon_i \,. \tag{F1}$$

[32] Although we do not plot them here, we find that the last kilometer of trips is about as slow as the first. The second and second to last kilometers are also slower than the trip average but faster than the first and last kilometers.

Equation (F1) includes many coefficients for each city. Comparing for instance the time of day effect for trips between 9:30 and 10 PM across 180 cities will not be insightful. Rather than keep all these coefficients separate, we aggregate them into index measures of speed for each city.

More specifically, we proceed as follows. We first estimate equation (F1) for each city separately. Each of these 180 regressions can be used to generate a predicted speed for all trips in the data, telling us how fast trip i would be if it were taken in city c: $\hat{S}_{ci} = \exp\left(\hat{\alpha}_c X_i' + \hat{\phi}_c^2/2\right)$, where $\hat{\phi}_c$ is the estimate of the standard deviation of ϵ_i in city c. We also predict speeds from an analogous 'national' regression using all trip instances by imposing common coefficients regardless of the city of travel: $\hat{S}_i = \exp\left(\hat{\alpha} X_i' + \hat{\phi}^2/2\right)$.

Then, we compute a predicted duration for each trip i if it were to take place in city c ($\hat{T}_{ci} = D_i/\hat{S}_{ci}$) or 'nationally' ($\hat{T}_i = D_i/\hat{S}_i$). Finally we can compute a relative speed index for each city:

$$L_c = \frac{\sum_i \hat{T}_i}{\sum_i \hat{T}_{ci}}\,. \tag{F2}$$

The index L_c represents the time it would take to conduct all trip instances in the data at the estimated speed for city c relative to the predicted time it would take to conduct these trips at the average estimated 'national' speed. L_c is a unitless scalar, but we can multiply it by $\sum_i D_i / \sum_i \hat{T}_i$, the average national speed, to transform it into a predicted speed for city i.

We note that the index L_c defined in equation (F2) resembles a Laspeyres price index in the sense that we compare the speed of trips across Indian cities for the same national bundle of trip instances. Like a standard Laspeyres index, L_c may be sensitive to sampling error or to out-of-sample predictions.

Alternatively, we can compute the predicted time it takes to undertake all city c trips in city c relative to the predicted time it takes to undertake all city c trips from a national regression. That is, we can compute:

$$P_c = \frac{\sum_{i \in c} \hat{T}_i}{\sum_{i \in c} \hat{T}_{ci}}\,. \tag{F3}$$

This alternative speed index is analogous to a Paasche price index. Because we compare city trips at predicted city speed to city trips at predicted national speed, this Paasche index will be less sensitive to the problems of out-of-sample predictions that may afflict the Laspeyres index above. It is also straightforward to compute the corresponding Fisher index: $F_c = \sqrt{L_c \times P_c}$.

B. Logit indices

We can also compute a broad class of speed indices derived from economic theory. We first specify a simple utility maximization framework that resembles the logit model of travel demand of Ben-Akiva and Lerman (1985). The travel decision is a discrete choice over a set of trip destinations. Cheaper (shorter) trips receive more weight, with the strength of that relationship governed by the elasticity of substitution σ. We define the utility from visiting the destination of trip i in city c as:

$$u_{ci} = \log(b_{ci}) + (1-\sigma)\log(t_{ci}) + \epsilon_{ci}, \tag{F4}$$

where $t_{ci} = \xi T_{ci}$ is the time cost of a trip to destination i in city c that takes T_{ci} units of time at value of time ξ per unit, and ϵ_{ci}, the random component of utility, has a Type I extreme value distribution.[33] Note that the value of time ξ does not vary across cities, as we compute speed indices for a representative traveler. The parameter $\sigma > 1$ is an elasticity of substitution across destinations, and b_{ci} is a trip-specific quality parameter capturing all factors other than time costs making some destinations more desirable than others.[34]

As shown by Anderson *et al.* (1992, pp. 60–61), the expected utility of a traveler in city c is equal to the expected value of u_{ci}'s maximum across the N_c travel destinations available in city c:

$$\mathbf{E}\left(\max_{i\in N_c}\{u_{ci}\}\right) = \log\left(\sum_{i=1}^{N_c}\exp\left[\log(b_{ci}) + (1-\sigma)\log(t_{ci})\right]\right) = \log\left(\sum_{i=1}^{N_c} b_{ci}t_{ci}^{1-\sigma}\right). \tag{F5}$$

Now consider two cities, c and c'. Define a relative price index $G_{c,c'}$ as the factor by which travel costs in city c would have to change in order to equalize expected utility in the two cities:

$$\log\left(\sum_{i=1}^{N_c} b_{ci}(G_{c,c'}t_{ci})^{1-\sigma}\right) = \log\left(\sum_{i=1}^{N_c} b_{c'i}t_{c'i}^{1-\sigma}\right), \tag{F6}$$

so that:

$$G_{c,c'} = \left(\frac{\sum_i^{N_{c'}} b_{c'i}t_{c'i}^{1-\sigma}}{\sum_i^{N_c} b_{ci}t_{ci}^{1-\sigma}}\right)^{1/(1-\sigma)} = \left(\frac{\sum_i^{N_{c'}} b_{c'i}T_{c'i}^{1-\sigma}}{\sum_i^{N_c} b_{ci}T_{ci}^{1-\sigma}}\right)^{1/(1-\sigma)}, \tag{F7}$$

[33] Ben-Akiva and Lerman (1985) show how to derive a travel accessibility index from a logit model of travel demand. Anderson, de Palma, and Thisse (1992) show the correspondence between the logit and CES models.

[34] In Table G.2, we present an index computed at $\sigma = 0$. Technically, values of $\sigma < 1$ are inconsistent with utility maximization. In practice, the index at $\sigma = 0$ simply weights all trips equally and intuitively corresponds to a perfect complement case.

where the second equality uses $t_{ci} = \xi T_{ci}$. The relative price index $G_{c,c'}$ is best characterized as a relative travel accessibility index. It is low when comparing a city c with many destinations to another city c' with few destinations (gains from variety), and when travel in c is short-distance and fast and travel is c' is long-distance and slow.

We now develop a simple non-parametric procedure to isolate a pure speed index determined only by speed differences across cities. Instead of comparing two cities c and c', we compute this speed index in each city c relative to a nationally representative city. To do this, we replace the denominator of $G_{c,c'}$ with a 'national index' that has exactly the same distribution of trip length as in city c, and the same number of trips:

$$G_c = \left(\frac{\sum_{i \in c} b_{ci} T_{ci}^{1-\sigma}}{\sum_{i \in c} b_{ci} \overline{T}_{ci}^{1-\sigma}} \right)^{1/(\sigma-1)} . \tag{F8}$$

Note that we inverted the index to ensure that G_c increases with faster speed (the index derived above is a price index increasing with time costs.) We compute $\overline{T}_{ci}$ as the average travel time of all trips in the national sample with length within 1% of that of trip i in city c.[35] Note that both the city-level numerator and the national-level denominator of G_c have the same number of trips, and the same distribution of trip lengths. The index in each city is therefore free of gains from variety and gains from closer proximity to travel destinations, and determined only by speed differences relative to a national sample.

Instead of tackling the difficult problem of estimating the parameters of G_c, we compute G_c for a wide range of values of σ and b_{ci}. The quality of a destination is unobserved, but if we restrict our sample to amenity trips, the assumption that $b_{ci} = 1$ is a reasonable starting point, because we sampled amenity trips to match the trip shares in the US NHTS. We also compute variants of G_c using random draws of $b_{ci} \in U[1, 100]$ for each trip, thus randomly allowing certain destinations to be more desirable and to carry a higher weight in the index.

We also experiment with richer nesting structures, in which trips to similar destination types (e.g. work, buy goods, recreational activities, etc) are more substitutable. To do so, we divide trips into M groups and compute the following nested CES/logit speed index (Sheu, 2014):

$$G_c^{nest} = \frac{\left(\sum_{m=1}^{M} G_{mc}^{1-\mu} \right)^{\frac{1}{1-\mu}}}{\left(\sum_{m=1}^{M} \overline{G}_{mc}^{1-\mu} \right)^{\frac{1}{1-\mu}}} , \tag{F9}$$

[35] We drop any trip with fewer than 10 corresponding trips within 1% of its length in the national sample (less than 0.01% of trips).

and

$$G_{mc} = \left(\sum_{i=1}^{N_{mc}} b_{ci} T_{ci}^{1-\sigma} \right)^{\frac{1}{1-\sigma}}, \quad \overline{G}_{mc} = \left(\sum_{i=1}^{N_{mc}} b_{ci} \overline{T}_{ci}^{1-\sigma} \right)^{\frac{1}{1-\sigma}}, \tag{F10}$$

where $\mu > 1$ is the elasticity of substitution across groups, $\sigma \geq \mu$ is the elasticity of substitution within groups, and N_{mc} is the number of trips in group m in city c. We define groups as the amenity types recorded in Appendix A. In this case, the nested index G_c^{nest} puts less weight on destination types that are relatively slower in city c; travelers substitute away from them because they are costlier. We compute these indices using exactly the same non-parametric methodology described above and compute G_c^{nest} for various values of μ and σ and distributions of b_{ci}. We also experiment with alternative nesting structures defined by time (e.g. off peak, peak, high peak) or area (e.g. rings).

C. Logit Index with Departure Time Decision

One nested-logit model of special interest allows travelers to choose a departure hour. We describe this model and its results in detail here. Our objective is to develop a model that matches the distribution of trip departure times at each hour of the day, which is a key moment of trip level data. To do so, we calibrate the 'quality' of each hour in a given city to match the share of travelers departing at that hour, given travel cost. The intuition is straightforward: if a large share of trips depart at peak hours despite high peak travel costs (i.e. low speed), then peak hour trips must be of higher quality and receive higher weight in our speed index.

So we set up a nested-logit model where travelers first choose a travel destination $i \in N_c$ (outer nest), and then a departure hour $h \in \{0, 1, ..., 23\}$ (inner nest). N_c is the total number of trips in our sample in city c. We can write the speed index from this model as in equation (F9) and (F10), with appropriate subscripts, as follow:

$$G_c^{nest} = \left(\frac{\sum_{i=1}^{N_c} G_{ci}^{1-\mu}}{\sum_{i=1}^{N_c} \overline{G}_{ci}^{1-\mu}} \right)^{\frac{1}{\mu-1}}, \tag{F11}$$

with:

$$G_{ci} = \left(\sum_{h=0}^{23} b_{ch} T_{chi}^{1-\sigma} \right)^{\frac{1}{1-\sigma}}, \quad \overline{G}_{ci} = \left(\sum_{h=0}^{23} b_{ch} \overline{T}_{chi}^{1-\sigma} \right)^{\frac{1}{1-\sigma}}, \tag{F12}$$

where $\overline{T}_{chi}$ is the national average duration of a trip of the same length and departing at the same hour as trip T_{chi}, $\mu > 1$ is the elasticity of substitution across trip destinations (across

nests), and $\sigma \geq \mu$ is the elasticity of substitution across departure hours (within nest). We again inverted the nested-logit price index to ensure that G_c^{nest} increases with faster speed.

First, we compute the parameters b_{ch} that capture the quality of a trip taken at hour h in city c. Within each nest i in city c, the utility from choosing a departure hour h is given by equation (F4), rewritten with appropriate subscripts as: $u_{chi} = \log(b_{ch}) + (1-\sigma)\log(t_{chi}) + \epsilon_{chi}$, which delivers hourly travel shares $\log\left(\frac{s_{chi}}{s_{c0i}}\right) = \log(b_{ch}) - (\sigma-1)\log\left(\frac{t_{chi}}{t_{c0i}}\right)$. As before, $t_{chi} = \xi T_{chi}$ is the time cost of a trip to destination i in city c at hour h that takes T_{chi} units of time at value of time ξ per unit. We then isolate the b_{ch} as follows:

$$b_{ch} = \frac{s_{ch}}{s_{c0}}\left(\frac{t_{ch}}{t_{c0}}\right)^{(\sigma-1)} = \frac{s_{ch}}{s_{c0}}\left(\frac{T_{ch}}{T_{c0}}\right)^{(\sigma-1)}. \tag{F13}$$

Note that b_{c0} is normalized to 1, and that we dropped the i subscript because we calibrate b_{ch} for a representative trip in each city. The second equality uses the assumption that travel cost t is proportional to travel time T. s_{ch} is the share of trips departing at hour h in city c. We get these hourly departure shares from the Intents data described in Appendix D. For the 14 cities with at least 1,000 Intents trips, we compute a city-specific b_{ch}. For all other cities, we pool data into three city size bins: large cities with population ranks from 1 to 20, medium sized cities ranked 21 to 60, and small cities ranked 60+. To compute T_{ch}, the travel time of a representative trip at hour h in city c, we estimate the following regression:

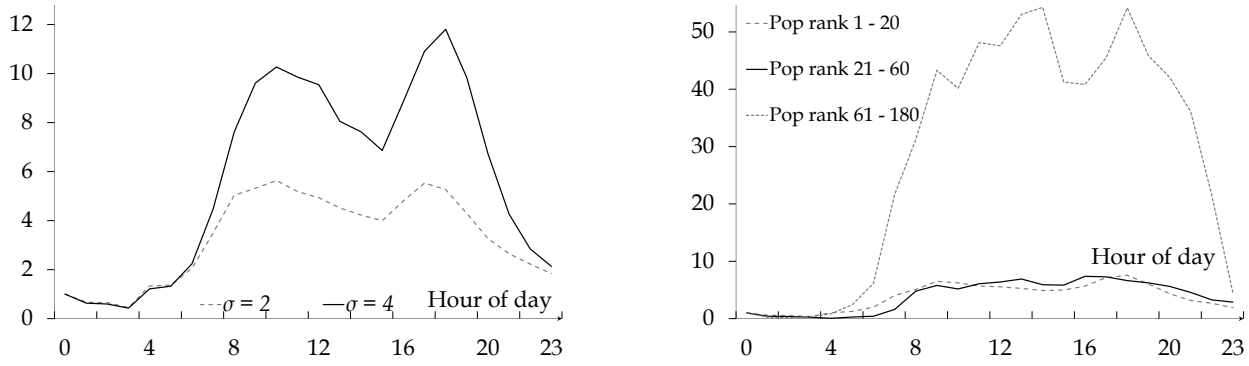
$$\log(T_{chi}) = \beta \log(distance_{chi}) + \sum_{h=1}^{23} hour_h + \epsilon_{chi}, \tag{F14}$$

where $hour_h$ is a dummy equal to one when a trip departs at hour h. We estimate equation (F14) from a sample of all trips in city (or city size bin) c in our GM trip sample. T_{ch} is the fitted value of that regression at the average trip length in city (or city size bin) c and at $hour_h = 1$.

Note that for any elasticity of substitution across hours of day σ and city-specific representative trip time T_{ch}, we can find quality parameters b_{ch} to exactly match the city-specific hourly departure shares s_{ch}. We lack an estimate of σ, so we compute b_{ch} for a wide range of values of σ.[36] Panel A of figure F.1 shows b_{ch} in Delhi for $\sigma = 2$ and $\sigma = 4$ at each hour of the day. As expected, a larger σ means that departure hours are more substitutable, so during peak hours the model calls for higher b_{ch} to explain the combination of high trip

[36] σ has two counterbalancing effects on our speed index. The direct effect of a large σ is to put less weight on the most expensive hours of the day at peak time. The indirect effect of a large σ is to call for high b_{ch} at peak hours, which puts *more* weight on the most expensive hours of the day at peak time.

Figure F.1: Hourly quality parameter b_{ch}



Panel A: b_{ch} for different σ in Delhi

Panel B: b_{ch} by city population bins ($\sigma = 2$)

Panel A shows the hourly quality parameter b_{ch} calibrated from equation (F13), computed for Delhi with $\sigma = 2$ (dotted line) and $\sigma = 4$ (solid line). Panel B shows b_{ch} computed for large cities (size rank 1 to 20, dashed line), medium cities (size rank 21 to 60, solid line), and small cities (size rank 61 to 180, dotted line).

shares and high travel costs. Panel B of figure F.1 shows b_{ch} for the large, medium, and small city size bins. We find that even though day time trips are much higher quality than night time trips everywhere, the difference between day and night time quality is much higher in smaller cities. This reflects the observed choice of travelers in smaller cities, who take almost no trips at night, perhaps because they have fewer safe and entertaining night time trip destinations available to them. As a result, peak time trips get relatively more weight in smaller cities.

Next, we compute the inner-nest index G_{ci} for each trip i. To compute T_{chi}, the time that trip i would take at each hour h in city c, we estimate the regression in equation (F14) separately for each city. T_{chi} is then a fitted value from that regression for a trip of the same distance as trip i that departs at hour h. We compute $\overline{T}_{chi}$ in exactly the same way, but as a fitted value from a national regression.

Finally, we compute the speed index G_c^{nest} in each city for values of σ and μ between 2 and 6. Relative to our benchmark index, which weights every trip equally, G_c^{nest} reweights peak hour trips through three different channels. First, the assumption that departure hours are substitutable puts less weight on peak time trips, which are most expensive (i.e. slowest). Second, and working in the opposite direction, the calibrated b_{ch} put *more* weight on peak time trips, which are of higher quality. The third channel comes from having different b_{ch} in different cities. As shown in Figure F.1, peak time trips get more weight in smaller cities. Overall, this reweighting of peak time trips makes G_c^{nest} somewhat less correlated

with congestion than our benchmark index.[37]

We find a rank correlation of 0.92 between our benchmark speed index and G_c^{nest} for reasonable values of σ and μ ($\sigma = \mu = 2$), and 0.79 for implausibly large values of σ and μ ($\sigma = \mu = 6$). These correlations are even higher if we restrict our sample to the 16 cities with city-specific b_{ch}. This strong correlation between the various speed indices that we produce is consistent with a key result of our paper that slow cities are generally slower at all times.[38]

Appendix G. Additional results, trip regressions

A. Additional specifications

Table G.1 reports a number of variants of our benchmark specification in table 2 column 4. Column 1 uses log effective speed (haversine length divided by time) instead of actual speed as dependent variable. The speed-effective trip length elasticity is slightly lower than the speed-actual trip length, because using effective length to define speeds reduces the distances of longer trips more than those of shorter trips, flattening the relationship with between log trip speed and log trip length. Column 2 uses speed under "typical" traffic conditions at the time of day of the request (also provided by GM) as dependent variable, and columns 3 and 4 limit to peak (10 AM to 2 PM and 4:30 PM to 9 PM) and high peak (6 to 8 PM) trips, respectively. Results are very similar to the baseline specification in table 2 column 4.

Column 5 limits attention to radial trips at peak hours going towards the center in the morning and back towards the periphery in the evening, mimicking archetypal commuting patterns. Because the trips are radial, beginning or ending near the city center, trip length and average distance to city center are mechanically highly correlated. As such, the distance to center relationship is overwhelmed by the trip length relationship and the coefficient on distance to the center is insignificant.

[37] Across a sample of all cities, our benchmark speed index has a correlation of -0.60 with the congestion factor, versus -0.45 for G_c^{nest}, computed at $\sigma = 4$ and $\mu = 2$.

[38] In fact, modeling the departure time choice generates relatively minor reweighting of trips relative to our benchmark index. This is because travel demand—and therefore the b_{ch} shown in Figure F.1—varies little throughout the day. In the Intents national sample, the hourly departure share between 8 AM and 9 PM only varies from a minimum of 0.05 from 8 PM to 9 PM to a maximum of 0.07 from 6 PM to 7 PM. All three city size bins show similar patterns. This time period account for about 80% of trips. One caveat is that the Intents users sample may not be representative of the general population.

Table G.1: Determinants of log trip speed, variants

	(1) Effective length	(2) Typical traffic	(3) Peak	(4) High peak	(5) 'Com-mutes'	(6) Congest. weighted	(7) Night	(8) Day
log trip length	0.14^a	0.19^a	0.19^a	0.18^a	0.25^a	0.21^a	0.19^a	0.19^a
	(0.0076)	(0.0052)	(0.0043)	(0.0042)	(0.011)	(0.0063)	(0.0045)	(0.0042)
log distance to center	0.10^a	0.083^a	0.098^a	0.100^a	0.0040	0.091^a	0.074^a	0.088^a
	(0.0055)	(0.0050)	(0.0057)	(0.0061)	(0.0094)	(0.0064)	(0.0053)	(0.0051)
Observations	41,991,655	41,991,655	18,735,401	3,532,661	2,058,065	41,991,655	9,154,230	32,837,425
R^2	0.33	0.59	0.56	0.56	0.70	0.64	0.62	0.56

Notes: OLS regressions including city, weekday, time of day (for each 30 minute period), and trip type indicators and weather controls for 180 cities as in table 2. The dependent variable is log effective speed in column 1, log speed under "typical" traffic conditions in column 2, log speed at peak hours (departure time between 10 AM and 2 PM or between 4:30 PM and 9:00 PM) in column 3, log speed at high peak hours (between 6 PM and 8:00 PM) in column 4, log speed for radial trips going towards the center during the 'morning' peak and away from the center during the evening peak in column 5, log speed computed after weighting trips using a congestion weight given by (Trip duration/trip duration in absence of traffic)$^{1/\lambda}$ with $\lambda = 0.3$, in column 6, log speed at night (sunset to sunrise) in column 7, and log speed during the day (sunrise to sunset) in column 8. Robust standard errors in parentheses. Intents weights used in all columns except column 6. a, b, c: significant at 1%, 5%, 10%.

Column 6 weights trips by their measured congestion factor, as in table G.2, Panel G, with elasticity of trip speed with respect to the density of vehicles, $\lambda = 0.3$. The speed-trip length elasticity increases slightly, consistent with longer trips more effectively avoiding congestion.

Columns 7 and 8 restrict attention to night (sunset to sunrise) and daytime (sunrise to sunset) trips, respectively. Results are very similar, with the only difference being a stronger distance to center effect for daytime trips consistent with greater daytime congestion.

B. Correlations of our baseline speed index with alternative indices

We report cross-city correlations between alternative speed indices and our benchmark index (the city fixed effects estimated from the specification reported in column 4 of table 2) in table G.2. We also report the standard deviation, maximum and minimum of each speed index. Standard deviations vary very little, except for the mean speed indices, which are constructed on a different (unlogged) scale.

Panel A compares our benchmark speed index to the analogous indices estimated in the other columns of table 2 that include various trip level controls. All these correlations are

Table G.2: Pairwise Spearman rank correlations with our benchmark speed index

Index	Corr.	Std. Dev.	Min	Max
Panel A: Columns from table 2				
(1)	0.942	0.119	-0.359	0.282
(2)	0.944	0.120	-0.362	0.285
(3)	0.998	0.114	-0.327	0.260
(5)	0.984	0.115	-0.353	0.230
(6)	0.918	0.122	-0.348	0.218
(7)	0.924	0.103	-0.319	0.184
Panel B: Trip subsamples				
Radial	0.928	0.134	-0.445	0.342
Circumferential	0.913	0.118	-0.319	0.328
Gravity	0.981	0.115	-0.329	0.238
Amenities	0.984	0.115	-0.335	0.280
0.5% sample	0.997	0.117	-0.337	0.273
0.1% sample	0.985	0.119	-0.347	0.275
Panel C: Mean speeds				
Simple mean	0.680	3.81	14.2	34.4
Mean unweighted by length	0.729	3.32	14.2	32.0
Mean of "typical" traffic speed	0.672	3.74	13.9	34.0
Mean of uncongested speed	0.570	4.10	15.9	38.0
Mean effective speed	0.677	2.82	8.76	26.9
Panel D: Variants of our main index				
Effective speed	0.892	0.132	-0.457	0.315
"Typical" traffic	0.988	0.114	-0.296	0.271
No traffic	0.956	0.097	-0.341	0.234
Fastest trip instance	0.982	0.106	-0.348	0.253
Off peak	0.985	0.109	-0.311	0.238
Peak	0.990	0.124	-0.365	0.288
High peak	0.970	0.130	-0.385	0.282
Peak radial	0.918	0.144	-0.412	0.361
Day (dawn to dusk)	0.994	0.119	-0.334	0.269
Night (dusk to dawn)	0.972	0.113	-0.349	0.268
Night (6 PM to 6 AM)	0.990	0.112	-0.352	0.253
Panel E: Full indices				
Laspeyres	0.898	0.140	0.671	1.416
Paasche	0.977	0.117	0.706	1.339
Fisher	0.953	0.126	0.688	1.345
Logit/CES ($\sigma = 0$)	0.955	0.120	0.657	1.317
Logit/CES ($\sigma = 2$)	0.913	0.125	0.739	1.431
Logit/CES ($\sigma = 4$)	0.822	0.152	0.775	1.688
Nested logit ($\sigma = 2$, $\mu = 2$)	0.918	0.124	0.719	1.422
Nested logit ($\sigma = 4$, $\mu = 4$)	0.846	0.128	0.698	1.419
Panel F: Distance to center				
Trips within 5 km of center	0.965	0.127	-0.329	0.386
Trips within 3 km of center	0.910	0.131	-0.337	0.396
Weight by inverse dist. to center	0.944	0.121	-0.333	0.339
Panel G: Weight by powered congestion factor				
$\lambda = 0.2$	0.936	0.146	-0.457	0.300
$\lambda = 0.3$	0.956	0.133	-0.415	0.290

Notes: 180 cities. For each alternative index, the first column reports the Spearman rank correlation with our benchmark index (table 2 column 4). Other columns report standard summary statistics.

above 0.91.

Panel B compares our benchmark index to the analogous indices estimated using the same specification but considering different types of trips separately. The correlations are again high. The lowest at 0.91 is with perhaps our most artificial type of trips, circumferential trips, and the highest is with perhaps our most realistic, amenity trips. Also in panel B, we show that after randomly dropping a large fraction of our trips (99.9%) within each city, we still obtain an index with a correlation of almost 0.99 with our benchmark index. This shows that city-level speed can be credibly measured using samples much smaller than those we collected.

Next, panel C compares our benchmark index to various measures of mean speed. The correlations are much lower than in the previous two panels. For instance, the correlation between our benchmark speed index and mean speed computed as total travel length divided by total travel time (equation 1) is only 0.68. This is because trip length has a large explanatory power on trip speed, and average trip length varies systematically across cities. As a result, mean speeds are sensitive to sampling strategies, unlike our preferred speed indices that control for trip length.

Panel D reports correlations between our benchmark speed index and speed indices computed from the estimations reported in table G.1, and related variants. Notably, the correlation of our benchmark speed index with an index that measures speed using effective (haversine), rather than traveled, trip length is 0.892. The 20 slowest cities using our benchmark speed index are all among the 25 slowest cities by effective speed. We can thus rule out the possibility that slow cities are more efficient at transporting travelers farther for the same number of straight line kilometers traveled. Slow cities are just slow.

The final part of panel D reports correlations between our benchmark index and speed indices computed in the same manner as our benchmark but from observations taken at specific hours of the day. For instance, an index computed using only trips at high peak time has a correlation of 0.970 with our benchmark index, and one calculated using only off peak trips has a correlation of 0.985 with our benchmark.[39]

Panel E reports correlations between our benchmark index and more sophisticated Laspeyres, Paasche, Fisher, and logit/CES indices. The Laspeyres index computed from equation (F2) using the same specification as for our benchmark index, but allowing all re-

[39]We did not compute speed indices from models with limited scheduling flexibility, as in Kreindler (2018). These models would essentially put higher weights on peak time trips, so our approach here is simply to show that speed indices based on *only* peak time trips are highly correlated with those based on all trips.

gression coefficients to vary across cities, has a correlation of 0.90 with our benchmark index. The Paasche and Fisher indices show even higher correlations. The logit/CES index from equation (F8), estimated at $\sigma = 0$ (the perfect complement case for which all trips receive equal weight) has a 0.95 correlation with our benchmark index. The correlation remains high at 0.82 even for an extreme value of $\sigma = 4$, which gives a two-kilometer trip about 400 times the weight of a longer 15-kilometer trip.[40] These correlations remain similarly high across a wide range of random destination quality draws b_{ci} (not shown). The nested-logit index from equation (F9) also delivers high correlations with our benchmark index even for large σ and μ equal to 4. We already described the high correlation between our benchmark index and a nested-logit index from a model with departure time choice. Other reasonable nesting patterns invariably generate high correlations with our benchmark index, suggesting that our results are robust to modeling travel demand with richer substitution patterns. This finding further confirms that our benchmark index provides a robust characterization of travel cost differences across cities, because slow cities tend to be slow at all times, for all types of trip destinations, and across the city.

Panel F considers indices computed from a sample restricted to trips that are close to the city center, or computed using higher weights for trips close to the center. These indices all have a correlation above 0.91 with our benchmark index.

In panel G we take another approach to weighting each trip by how likely it is to be taken, using the implicit density of vehicles along the route as a proxy. To do so, we assume that (i) the speed of a trip instance is reduced from the maximum for that trip solely by vehicle congestion, (ii) the elasticity of trip speed with respect to the density of vehicles, λ, is constant, and (iii) the density of vehicles is constant along the route. Under these assumptions, we can weight each trip i by its length, D_i, times the implicit density of vehicles, $(T_i/T_i^{nt})^{1/\lambda}$. While these assumptions are unlikely to be strictly true, they manage to capture the fact that more vehicles slow down traffic and thus slower trip instances should receive a higher weight given that they represent more travelers. We use $\lambda = 0.2$, which is a standard value in the traffic modelling literature (Small and Verhoef, 2007). We also use a higher value $\lambda = 0.3$ that reduces the weight put on slow trips, since slower speeds in India may not be caused only by more vehicle traffic. With both values, the indices are highly correlated with our benchmark index.

[40] Atkin, Faber, and Gonzalez-Navarro (2018) estimate an elasticity of substitution across retail stores slightly smaller than 4 for poor Mexican households. This is almost certainly an upper bound: the index considered here covers a much broader set of destinations that are unlikely to be as substitutable as retail stores.

Appendix H. Additional results, variance decomposition

Table H.1 reports variants of table 4 for restricted subsamples. Broadly, congestion plays a larger role near the city center, for peak radial trips, and (mechanically) congestion-weighted, and that is especially true for the largest cities. However, the role of uncongested speed remains substantial, and only falls substantially below the role of the congestion factor when restricting to high peak in city centers of the largest cities. The role of congestion varies little across trip types, but is generally slightly larger for radial and circumferential trips than amenity or gravity trips. Throughout all of these samples, the covariance between uncongested speed and the congestion factor remains negative and fairly small.

Appendix I. Additional results, city regressions

Table I.1 adds three additional regressors to the specifications presented in table 5. Results are presented in the main text.

Table I.2 reports several variants of table 5 column 3 predicting city speed indices. Column 1 reports standardized coefficients. Consistent with the main text discussion of explanatory power, population has by far the largest standardized coefficient in magnitude, followed by area and roads. In column 2, replacing Census-based city populations with estimates from United Nations (2019) reduces the coefficient on population by more than a third, consistent with more measurement error.

Removing tertiary, and both tertiary and secondary roads, respectively, from the definition of major roads in columns 3 and 4 reduces the coefficient on the roads measure by a factor of two and three, respectively. They remain precisely estimated. Such roads appear to be important. Column 5 adds a measure of paved roads from the Census. The coefficient is small and insignificant, likely because paved roads are poorly measured. Column 6 introduces the Intents potholes, measured as the inverse hyperbolic sine of potholes per kilometer to allow us to include cities with zero recorded potholes. We restrict the sample to 87 cities for which Intents has enough driving data to limit the standard error of the measure to at most 0.05. The coefficient is small, with a large standard error. We interpret this large standard error as meaning that we do not have enough systematic data to make a careful comparison across cities in this dimension. Column 7 replaces our preferred grid-like

Table H.1: Variance decompositions of our baseline speed index, variants

Sample	Cities	All trips			High peak trips		
		Uncongested speed	Congestion factor	Covariance	Uncongested speed	Congestion factor	Covariance
Panel A: Distance to city center less than 5 km							
All	180	0.665	0.161	-0.087	0.530	0.288	-0.091
Smallest 50%	90	0.810	0.063	-0.064	0.824	0.111	-0.032
Largest 50%	90	0.566	0.236	-0.099	0.408	0.362	-0.115
Largest 25%	45	0.463	0.351	-0.093	0.327	0.463	-0.105
Largest 10%	18	0.407	0.440	-0.076	0.290	0.524	-0.093
Panel B: Distance to city center less than 3 km							
All	180	0.643	0.169	-0.094	0.520	0.292	-0.094
Smallest 50%	90	0.746	0.067	-0.094	0.739	0.112	-0.075
Largest 50%	90	0.586	0.233	-0.090	0.441	0.358	-0.100
Largest 25%	45	0.488	0.348	-0.082	0.360	0.462	-0.089
Largest 10%	18	0.457	0.462	-0.041	0.336	0.560	-0.052
Panel C: By trip type							
Radial	180	0.745	0.153	-0.051	0.623	0.285	-0.046
Circumferential	180	0.737	0.129	-0.067	0.590	0.287	-0.062
Gravity	180	0.715	0.107	-0.089	0.575	0.228	-0.099
Amenities	180	0.676	0.126	-0.099	0.541	0.265	-0.097
Panel D:		Peak radial trips ('commutes')			Congestion weighted trips		
All	180	0.642	0.246	-0.056	0.554	0.231	-0.107
Smallest 50%	90	0.880	0.106	-0.007	0.684	0.121	-0.097
Largest 50%	90	0.504	0.316	-0.090	0.481	0.325	-0.097
Largest 25%	45	0.393	0.408	-0.100	0.371	0.397	-0.116
Largest 10%	18	0.366	0.570	-0.032	0.327	0.478	-0.098

Note: High peak hours are 6 - 8 PM and peak hours are 10 AM to 2 PM and 4:30 PM to 9 PM For the congestion weight we use $\lambda = 0.3$.

network measure with an alternative: the Gini coefficient of the distribution of compass bearings of all edges in the OSM street network. It is also associated with faster travel, with a similar magnitude (in both cases a one standard deviation higher value is associated with about 15% faster travel).

Column 8 replaces our earnings measure with an alternative income measure, night lights. It again suggests a hill-shaped relationship between income and speed. Finally, column 9 adds an index of city shape: the first principal component of the six main measures of city shape (perimeter, spin, dispersion, range, proximity, and cohesion) for 2010 from

Table I.1: Further correlates of city indices

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent variable	Speed index			Uncongested speed			Congestion factor		
log population	-0.18^a	-0.18^a	-0.18^a	-0.15^a	-0.15^a	-0.14^a	0.036^a	0.033^a	0.040^a
	(0.016)	(0.015)	(0.016)	(0.012)	(0.013)	(0.013)	(0.0071)	(0.0062)	(0.0067)
log area	0.090^a	0.10^a	0.095^a	0.060^a	0.073^a	0.067^a	-0.030^a	-0.029^a	-0.028^a
	(0.020)	(0.019)	(0.019)	(0.019)	(0.018)	(0.019)	(0.0083)	(0.0071)	(0.0078)
Elevation variance	-0.036^a	-0.034^a	-0.036^a	-0.027^a	-0.029^a	-0.028^a	0.0090^a	0.0050^b	0.0084^b
	(0.0047)	(0.0047)	(0.0046)	(0.0028)	(0.0039)	(0.0028)	(0.0033)	(0.0022)	(0.0033)
Water length	-0.12^a	-0.089^c	-0.13^a	-0.065^c	-0.054	-0.080^b	0.060^a	0.035	0.054^b
	(0.041)	(0.045)	(0.040)	(0.036)	(0.039)	(0.035)	(0.021)	(0.023)	(0.022)
log major roads	0.071^a	0.054^a	0.069^a	0.076^a	0.060^a	0.073^a	0.0049	0.0062	0.0037
	(0.016)	(0.015)	(0.016)	(0.017)	(0.016)	(0.017)	(0.0053)	(0.0047)	(0.0051)
log street lights	0.011^b	0.0067	0.010^b	0.010^b	0.0075	0.0091^b	-0.00094	0.00082	-0.0014
	(0.0050)	(0.0047)	(0.0050)	(0.0045)	(0.0046)	(0.0043)	(0.0018)	(0.0019)	(0.0018)
Network	0.30^a	0.25^a	0.30^a	0.26^a	0.20^b	0.28^a	-0.034	-0.055^c	-0.025
	(0.086)	(0.085)	(0.088)	(0.076)	(0.080)	(0.074)	(0.033)	(0.031)	(0.033)
Earnings	0.026^a	0.030^a	0.028^a	0.012	0.016^b	0.014^c	-0.014^a	-0.014^a	-0.013^a
	(0.0089)	(0.0084)	(0.0089)	(0.0081)	(0.0076)	(0.0079)	(0.0038)	(0.0038)	(0.0039)
Earnings^2	-0.0020^a	-0.0022^a	-0.0021^a	-0.00065	-0.00096^c	-0.00080^c	0.0013^a	0.0012^a	0.0013^a
	(0.00053)	(0.00053)	(0.00053)	(0.00048)	(0.00049)	(0.00047)	(0.00023)	(0.00023)	(0.00024)
Pop. growth 90-18	0.026			0.042^b			0.016^b		
	(0.018)			(0.016)			(0.0072)		
Share w. car		-0.065			0.18			0.24^a	
		(0.15)			(0.14)			(0.058)	
Share w. motorcycle		0.17^a			0.065			-0.11^a	
		(0.061)			(0.060)			(0.022)	
Spatial Gini pop.			-0.075			-0.15^b			-0.078^a
			(0.073)			(0.064)			(0.023)
R^2	0.64	0.66	0.64	0.61	0.61	0.61	0.52	0.58	0.54
Observations	180	180	180	180	180	180	180	180	180

Notes: OLS regressions with a constant in all columns. Robust standard errors in parentheses. a, b, c: significant at 1%, 5%, 10%. This table duplicates table 5 in the main text. We additionally include log population growth between 1990 and 2018 from UN data, the shares of households with access to a motorbike and to a car from the 2011 census, and a Gini index for the spatial distribution of population. For the latter, we compute a Gini index for the distribution of population across 90-meter pixels using the gridded population estimates from WorldPop.

Harari (2020). More sprawl (less compactness) is associated with slower travel. We note that relative to Harari (2020), we are considering a static context, and controlling directly for water bodies and elevation variance that may affect a city's shape.

Table I.3 reports several variants of table 5 column 3 with alternative speed indices as dependent variable. Column 1 uses effective (haversine) distance between trip origin and destination to calculate trip speed, thereby penalizing more circuitous trips. Column 2 uses GM's estimate of 'typical' travel time to calculate trip speed. Columns 3 and 4 restrict to trips

Table I.2: Correlates of city indices, alternative and additional explanatory variables

Variant	(1) Reference spec. Beta coef.	(2) UN 2018 population	(3) Top 3 road classes	(4) Top 2 road classes	(5) Paved roads	(6) Potholes	(7) Altern. network measure	(8) Altern. income (nightlights)	(9) City shape Harari index
log population	-1.54^a	-0.11^a	-0.17^a	-0.17^a	-0.19^a	-0.18^a	-0.18^a	-0.16^a	-0.16^a
		(0.023)	(0.015)	(0.016)	(0.016)	(0.025)	(0.016)	(0.018)	(0.016)
log area	0.75^a	0.016	0.12^a	0.13^a	0.094^a	0.061^b	0.095^a	0.10^a	0.092^a
		(0.021)	(0.018)	(0.017)	(0.020)	(0.025)	(0.020)	(0.025)	(0.018)
Elevation variance	-0.24^a	-0.024^a	-0.032^a	-0.034^a	-0.036^a	-0.046^b	-0.036^a	-0.029^a	-0.035^a
		(0.0067)	(0.0038)	(0.0040)	(0.0047)	(0.020)	(0.0049)	(0.0046)	(0.0043)
Water length	-0.22^a	-0.14^a	-0.15^a	-0.16^a	-0.13^a	-0.092^c	-0.14^a	-0.066	-0.10^b
		(0.049)	(0.041)	(0.042)	(0.041)	(0.047)	(0.042)	(0.051)	(0.043)
log major roads	0.67^a	0.069^a	0.036^a	0.022^b	0.069^a	0.077^a	0.071^a	0.073^a	0.063^a
		(0.021)	(0.012)	(0.0090)	(0.016)	(0.017)	(0.016)	(0.015)	(0.013)
log street lights	0.15^b	0.0048	0.012^b	0.012^b	0.0091^c	0.013^c	0.0093^c	0.016^a	0.012^b
		(0.0066)	(0.0053)	(0.0053)	(0.0047)	(0.0067)	(0.0051)	(0.0050)	(0.0053)
Network	0.14^a	0.16	0.34^a	0.36^a	0.28^a	0.20^c	0.16^b	0.35^a	0.35^a
		(0.12)	(0.096)	(0.10)	(0.086)	(0.11)	(0.067)	(0.10)	(0.099)
Earnings	0.47^a	0.046^a	0.031^a	0.033^a	0.029^a	0.025^b	0.030^a	0.23^c	0.038^a
		(0.013)	(0.0088)	(0.0089)	(0.0086)	(0.012)	(0.0085)	(0.12)	(0.0089)
Earnings^2	-0.47^a	-0.0033^a	-0.0023^a	-0.0023^a	-0.0021^a	-0.0019^a	-0.0022^a	-0.015^b	-0.0025^a
		(0.00076)	(0.00055)	(0.00055)	(0.00051)	(0.00067)	(0.00051)	(0.0065)	(0.00055)
log paved roads					0.0056				
					(0.0081)				
asinh potholes per km						0.0069			
						(0.029)			
Harari index									-0.0092^b
									(0.0036)
R-squared	0.64	0.42	0.62	0.60	0.64	0.67	0.63	0.64	0.67
Observations	180	180	180	180	180	87	180	179	155

Notes: OLS regressions with a constant in all columns. Robust standard errors in parentheses. a, b, c: significant at 1%, 5%, 10%. This table reports variants of main text table 5 column 3. Column 1 reports standardized coefficients (beta coefficients). Column 2 uses city population from the UN in 2018. Columns 3 and 4 measure road length without tertiary roads and without secondary and tertiary roads, respectively. We use the inverse hyperbolic sine instead of a log transformation to avoid excluding a small number of zeroes. Columns 5 adds a measure of (log) paved roads computed from the 2011 census. Columns 6 adds a measure of (log) potholes per kilometer computed using Intents data. Column 7 considers an alternative network index, the Gini coefficient for the distribution of edge compass bearings in the road network, which also measures how grid-like the road network of a city is. Columns 8 uses log total night lights and its square to measure income and its square. Column 9 includes an index of city shape: the first principal component of the six main measures of city shape (perimeter, spin, dispersion, range, proximity, and cohesion) for 2010 from Harari (2020).

at peak (10 AM to 2 PM and 4:30 PM to 9 PM) and high peak (6 PM to 8 PM) hours, respectively. Column 5 restricts to radial inward trips during the morning peak and radial outward trips during the evening peak, mimicking typical monocentric city commute patterns. Column

Table I.3: Correlates of city indices, alternative indices

Dependent variable	(1) Effective speed	(2) Usual traffic	(3) Peak hour	(4) High peak hour	(5) Commute speed	(6) Congestion weight	(7) Night speed	(8) Day speed	(9) Full controls
log population	-0.17^a	-0.19^a	-0.20^a	-0.21^a	-0.21^a	-0.19^a	-0.18^a	-0.19^a	-0.13^a
	(0.020)	(0.015)	(0.017)	(0.017)	(0.020)	(0.020)	(0.016)	(0.016)	(0.015)
log area	0.068^b	0.098^a	0.11^a	0.11^a	0.14^a	0.064^a	0.12^a	0.090^a	0.046^b
	(0.026)	(0.019)	(0.020)	(0.020)	(0.027)	(0.024)	(0.018)	(0.020)	(0.018)
Elevation variance	-0.064^a	-0.033^a	-0.037^a	-0.033^a	-0.039^a	-0.047^a	-0.036^a	-0.042^a	-0.027^a
	(0.0065)	(0.0033)	(0.0055)	(0.0071)	(0.0063)	(0.0076)	(0.0046)	(0.0044)	(0.0058)
Water length	-0.19^a	-0.090^b	-0.13^a	-0.12^a	-0.12^b	-0.15^a	-0.13^a	-0.13^a	-0.16^a
	(0.042)	(0.042)	(0.042)	(0.042)	(0.058)	(0.045)	(0.040)	(0.040)	(0.033)
log major roads	0.077^a	0.065^a	0.062^a	0.053^a	0.061^b	0.086^a	0.051^a	0.075^a	0.086^a
	(0.018)	(0.016)	(0.017)	(0.017)	(0.024)	(0.020)	(0.015)	(0.017)	(0.015)
log street lights	0.0089	0.011^b	0.010^c	0.0088^c	0.0043	0.012^c	0.0088^c	0.011^b	0.011^b
	(0.0057)	(0.0052)	(0.0053)	(0.0052)	(0.0066)	(0.0061)	(0.0048)	(0.0052)	(0.0052)
Network	0.016	0.27^a	0.31^a	0.31^a	0.45^a	0.35^a	0.33^a	0.27^a	0.23^b
	(0.12)	(0.085)	(0.089)	(0.091)	(0.11)	(0.11)	(0.083)	(0.090)	(0.098)
Earnings	0.031^a	0.028^a	0.035^a	0.043^a	0.046^a	0.027^b	0.034^a	0.024^a	0.021^b
	(0.011)	(0.0086)	(0.0092)	(0.0094)	(0.012)	(0.011)	(0.0085)	(0.0092)	(0.0087)
Earnings^2	-0.0023^a	-0.0019^a	-0.0028^a	-0.0034^a	-0.0037^a	-0.0024^a	-0.0025^a	-0.0018^a	-0.0016^a
	(0.00069)	(0.00052)	(0.00055)	(0.00055)	(0.00072)	(0.00065)	(0.00052)	(0.00058)	(0.00053)
R^2	0.54	0.64	0.64	0.65	0.57	0.60	0.63	0.64	0.57
Observations	180	180	180	180	180	180	180	180	180

Notes: OLS regressions with a constant in all columns. Robust standard errors in parentheses. a, b, c: significant at 1%, 5%, 10%. This table duplicates table 5 in the main text. The dependent variable is computed in the same manner as in column 4 of table 2 in columns 1-8 using effective speed in column 1, GM's usual traffic speed in column 2, peak hour speed (10 AM to 2 PM and 4:30 PM to 9 PM) in column 3, speed at high peak hours (6 PM to 8 PM) in column 4, speed for trips mimicking traditional commutes (radial trips towards the center during the morning peak and away during the evening peak) in column 5, speed for which trips are weighted by a congestion factor (each trip's weight is given by the ratio of duration over duration in absence of traffic elevated to the power 0.3) in column 6, night speed (sunset to sunrise) in column 7, and day speed (sunrise to sunset) in column 8. Column 9 uses the speed fixed effects estimated in the full specification of column 7 in table 2.

6 weights trips by their congestion factor raised to the power 0.3. Columns 7 and 8 restrict to nighttime (local sunset to sunrise) and daytime (sunrise to sunset) trips, respectively. Column 9 uses the 'narrow' fixed effects from column 7 of table 2. In all cases, results are very similar to those in table 5 column 3.

Appendix J. Additional Welfare Results

In this appendix, we express the welfare gains from a 10% reduction in uncongested speed in monetary units, and explore heterogeneity in these gains across cities. For each city, we

compute the yearly welfare gains for the average worker as follow:

$$250 \times (distance_c / speed_c) \times commute\_share_c \times wage\_commuter_c \times 0.1 \qquad \text{(J1)}$$

A number of assumptions are implicit in this equation. We only consider gains to work commuters, and assume a work year of 250 days, no crowding out from congestion, and no gains from induced demand. Finally, we assume that the value of one hour of travel time equals 100% of the hourly wage.[41] The share of commuters using motorized on-road vehicles (cars, motorcycles, on-road buses) and the average commute distance come from the Indian Census. We get city-specific commute speed at peak time from our GM data. The hourly wage comes from NSS microdata on male workers' daily earnings. We then compute a city-specific average wage of vehicle commuters by assuming that these commuters are the richest workers in their city.[42] Due to small NSS sample sizes in some cities, we limit our sample for these computations to the 100 cities with the most observations (at least 76 workers). Appendix B provides additional description of all data sources.

Computing equation (J1) for each city, and focusing on city population-weighted averages across all cities, we find that a 10% reduction in uncongested speed generates yearly gains of 1,152 INR per worker (about 16 USD). These gains accrue to the 54 percent of workers who commute by vehicle, for on average about 0.92 hour daily. These commuters save about 5.5 minutes in travel time per work day from the speed improvement, valued at 100% of the average wage of vehicle commuters of 119 INR per hour (1.7 USD). These average gains mask significant heterogeneity across cities, with an interquartile range across cities from 701 to 1,342 INR (10 to 19 USD). The largest average yearly gains per worker are 2,696 INR (38 USD) in Delhi, which has a high commute share and high commuter wages, and lowest at 361 INR (5 USD) in Bhavnagar, a city with a low commute share and low commuter wages.[43] Looking across the city size distribution, we find an unweighted average of 1,472 INR per worker (21

[41] A value of time at 50% of hourly wage is commonly used following Small and Verhoef (2007)'s evidence from rich countries. However, Kreindler (2018) finds much larger values at 400% times the hourly wage using experimental data from Bangalore. We pick a value of 100% to remain conservative while accounting for the possibility of higher valuation in developing countries, perhaps due to road or vehicle quality or other costs like fuel that are relatively more important there.

[42] We need this assumption because the NSS has no information on commutes, and a simple average would miss that a significant fraction of Indians are too poor to afford a vehicle. So if 40% of workers in a city are vehicle commuters according to the Census, then we compute commuter wage as the average wage within the top 40% of the NSS wage distribution in that city.

[43] Looking at interquartile ratios, these gains are 1.91 times larger at the 75^{th} percentile than at the 25^{th} percentile of the city distribution of gains. This variation in gains is mostly driven by variations in wages (interquartile ratio of 1.70) and commuter share (interquartile ratio of 1.48). Commute speed and distance feature less cross-city variation (interquartile ratios of 1.28 and 1.26 respectively.)

USD) in the ten largest cities in India, versus 762 INR (11 USD) for the 10 smallest cities in our sample, with most of the difference due to higher wages and higher commute shares in larger cities.

Appendix K. A ring analysis of speed in Indian cities

Although our main findings of city-level correlations in Section 6 are generally stable across a wide variety of specifications, they may be subject to bias due to omitted city-level variables. We now use within-city variation in population, area, and roads to avoid this problem and gain further insights about variation in speed.

Specifically, we divide each city in our sample into concentric rings. Among other advantages, nearly all radial trips will pass through the same rings, regardless of route. We apply the following transformation of equation (2), which uses the location of trips within cities to estimate a speed index for each ring within each city:

$$\log S_i = \alpha X_i' + \sum_r R_{rc(i)}\, share_{rc(i)}(i) + \epsilon_i\,, \tag{K1}$$

where $share_{rc(i)}(i)$ is the share of trip i which takes places within ring r of city c and R_{rc} is a speed index for ring r of city c. We consider (up to) 5 rings around each city center: 0 to 3 kilometers, 3 to 5, 5 to 10, 10 to 15, and 15 and beyond. We compute each trip's share in each ring using information from driving directions. We estimate equation (K1) using as controls log trip length, log distance to the center, time of day and day of week indicators in a manner that is consistent with our baseline index.

In a second step, we estimate the following regression:

$$\hat{R}_{rc} = \kappa_r + \beta_c + \alpha X_{rc}' + \epsilon_i\,, \tag{K2}$$

where κ_r is a ring fixed effect, β_c is a city fixed effect, and X_{rc} is a vector of city-ring explanatory variables. In our dataset, only land area, population, and roads are available separately by city-ring. Two caveats must be kept in mind. First, we only consider rings with 20,000 residents or more to avoid rings that are mostly empty. Second, we also expect some equilibrium effects across rings as, for instance, population in nearby rings may affect speed locally. Given the limited precision of our population data, detecting such effects may be out of reach here.

Table K.1: Correlates of city speed indices, rings analysis

	(1) Base	(2) No Step 1 Control	(3) < 5 km	(4) < 3 km	(5) Base	(6) <5 km	(7) Peak	(8) Peak <5 km
log ring population	-0.078^a	-0.10^a	-0.071^a	-0.070^a	-0.073^a	-0.066^a	-0.087^a	-0.079^a
	(0.013)	(0.016)	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)	(0.014)
log ring area	0.036^c	0.047^b	0.031	0.031	0.013	0.014	0.048^b	0.043^b
	(0.019)	(0.023)	(0.019)	(0.019)	(0.019)	(0.020)	(0.020)	(0.020)
asinh roads	-0.0040	-0.0062	-0.0056	-0.0053			-0.0049	-0.0066
	(0.0042)	(0.0051)	(0.0042)	(0.0040)			(0.0044)	(0.0044)
ring 2	0.091^a	0.16^a	0.062^a	0.050^a	-0.17^b	-0.036	0.095^a	0.060^a
	(0.021)	(0.024)	(0.019)	(0.018)	(0.068)	(0.058)	(0.022)	(0.020)
ring 3	0.15^a	0.29^a	0.10^a	0.087^a	-0.037	0.061	0.16^a	0.10^a
	(0.031)	(0.036)	(0.030)	(0.029)	(0.073)	(0.071)	(0.033)	(0.031)
ring 4	0.13^a	0.27^a	0.073^b	0.058^c	-0.18^c	-0.096	0.14^a	0.065^c
	(0.036)	(0.044)	(0.034)	(0.032)	(0.096)	(0.083)	(0.039)	(0.036)
ring 5	0.11^a	0.22^a	0.049	0.035	-0.31^a	-0.21^a	0.12^a	0.036
	(0.039)	(0.047)	(0.039)	(0.038)	(0.086)	(0.075)	(0.043)	(0.042)
roads per ring	N	N	N	N	Y	Y	N	N
Observations	628	628	627	627	628	627	628	627
R^2	0.45	0.62	0.35	0.32	0.49	0.38	0.48	0.36

Notes: OLS regressions with a city fixed effect and a ring fixed effect in all columns (180 cities in all regressions). The dependent variable is the city-ring fixed effect estimated as per equation (K2). Robust standard errors in parentheses. a, b, c: significant at 1%, 5%, 10%. Column 1 is our baseline estimation for which city-ring effects are estimated as described in the text. Column 2 considers city ring effects estimated with out trip controls in the first step. Columns 3 and 4 only consider trips with a length of less than 5 and 3 kilometers respectively. Columns 5 and 6 estimate separate roads effects for each ring. Columns 7 and 8 duplicate columns 1 and 3 but only consider peak-hour trips.

We report results in table K.1. The coefficient on population is -0.08 in our baseline specification, and similar in the rest of the table. We note that the population coefficients estimated in table K.1 are only about half those estimated in table 5. This may be because our measures of ring population are less precise. We also expect speed within ring to be determined by population in neighboring rings.[44] Consistent with table 5, table K.1 also reports small positive coefficients for area. On the other hand, the coefficient on roads is generally negative, albeit insignificant. Although we do not report the details here, this negative coefficient is driven mainly by the central ring when roads effects are allowed to

[44]We experimented with specifications that also included population in neighboring rings. Estimated coefficients are generally small and insignificant.

vary by ring in columns 5 and 6. Finally, table K.1 also reports that speed is generally higher in outer rings, which confirms results from section 4.

Appendix L. Transit

Table L.1 reports results analogous to table 2 columns 4 and 7 for the speed of transit trips. The main text and Appendix A discuss the transit data in detail including important caveats for interpreting these results, most notably that travel times are scheduled, not actual, and only for formal transit services.

Columns 1–4, our preferred specifications, calculate transit speed using transit time and driving distance (i.e. distance for the private vehicle route). Columns 5–8 alternatively use transit distance (i.e. distance for the transit route), which is sometimes considerably longer. We prefer driving distance because we believe that transit distance is likely to be mismeasured. Also, to the extent that the purpose of this exercise is to compare transit with driving, holding route constant allows us to focus attention on travel time.

Columns 1–2 (and 5–6) consider the full trip, 'door-to-door' (trip time), while columns 3–4 (and 7–8) exclude the portions of the trip walking to and from transit stops and waiting for a vehicle, restricting attention to the time spent in the transit vehicle, typically a bus (transit time).

Longer trips are consistently faster, as in private vehicles. Distance to the city center only has a significant relationship with speed when using transit routes. The negative coefficient is suggestive of higher frequency of service near the city center.

Route-specific variables are somewhat more difficult to assess in the context of transit, because they are likely to be mismeasured both in columns 1–4, where we measure the private vehicle route, and in columns 5–8 as noted above. Road types enter roughly as expected, though somewhat more attenuated in columns 5–8, and right turns are associated with slower bus travel. Perhaps counterintuitively, more establishments are associated with faster travel, even when limiting to transit time. More intersections are associated with slower travel when considering transit routes (but faster travel when considering private vehicle routes.) Finally, as noted in the main text, there are large differences between trip types. Most notably, circumferential trips are considerably slower than radial trips, consistent with transit networks being even more radially oriented than road networks.

Table L.1: Determinants of log trip speed for transit

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Distance	Privately-owned vehicle route				Transit route			
Duration	Trip time		Transit time		Trip time		Transit time	
log trip length	0.23^a	0.077^a	0.19^a	0.054^a	0.34^a	0.33^a	0.30^a	0.31^a
	(0.014)	(0.015)	(0.016)	(0.017)	(0.010)	(0.012)	(0.012)	(0.013)
log distance to center	-0.020	0.0086	0.0034	0.030	-0.067^a	-0.061^a	-0.044^a	-0.040^a
	(0.012)	(0.016)	(0.015)	(0.018)	(0.0074)	(0.0075)	(0.0097)	(0.0093)
Gross gradient up		-1.35^a		-1.60^a		-0.52		-0.77^b
		(0.41)		(0.41)		(0.33)		(0.36)
Gross gradient down		-0.65^c		-0.49		-0.27		-0.12
		(0.35)		(0.33)		(0.26)		(0.21)
Share primary roads		-0.088^a		-0.086^a		0.019		0.020
		(0.025)		(0.026)		(0.017)		(0.018)
Share secondary roads		-0.23^a		-0.22^a		0.0061		0.015
		(0.032)		(0.031)		(0.017)		(0.016)
Share tertiary roads		-0.35^a		-0.34^a		-0.044^c		-0.035
		(0.030)		(0.030)		(0.023)		(0.022)
Share resid. roads		-0.45^a		-0.45^a		-0.057^b		-0.059^b
		(0.027)		(0.027)		(0.022)		(0.023)
Share other roads		-0.33^a		-0.34^a		-0.065^a		-0.074^a
		(0.031)		(0.034)		(0.024)		(0.027)
Share missing roads		-0.15^a		-0.15^a		-0.048^b		-0.047^c
		(0.038)		(0.037)		(0.024)		(0.024)
log # intersections		0.14^a		0.12^a		-0.022^a		-0.037^a
		(0.014)		(0.015)		(0.0079)		(0.0082)
arsinh # right turns		-0.021^a		-0.025^a		-0.034^a		-0.037^a
		(0.0054)		(0.0051)		(0.0028)		(0.0029)
arsinh # establishments		0.038^a		0.036^a		0.032^a		0.031^a
		(0.011)		(0.010)		(0.0072)		(0.0067)
Type: circumferential	-0.22^a	-0.16^a	-0.23^a	-0.17^a	-0.067^a	-0.054^a	-0.076^a	-0.065^a
	(0.017)	(0.016)	(0.015)	(0.014)	(0.018)	(0.017)	(0.016)	(0.015)
Type: gravity	-0.12^a	-0.087^a	-0.13^a	-0.092^a	-0.033^a	-0.023^b	-0.038^a	-0.029^a
	(0.017)	(0.016)	(0.016)	(0.016)	(0.012)	(0.011)	(0.012)	(0.011)
Type: amenity	-0.11^a	-0.085^a	-0.12^a	-0.092^a	0.011	0.011	0.0055	0.0034
	(0.018)	(0.019)	(0.018)	(0.018)	(0.015)	(0.014)	(0.015)	(0.014)
Observations	1,001,723	1,001,002	1,001,723	1,001,002	1,002,145	1,001,403	1,002,145	1,001,403
R^2	0.39	0.47	0.36	0.43	0.60	0.61	0.56	0.58

Notes: 148 cities in each column. OLS regressions with weather controls and city, weekday, and time of day (for each 30 minute period) indicators using Intents weights. The transit index in columns to 4 uses distance from an analogous trip using a privately-owned vehicle. Transit distance is used in columns 5 to 8. The transit index in columns 1, 2, 5, and 6 uses total trip time. Time in transit is used in columns 3, 4, 7, and 8. Robust standard errors in parentheses. a, b, c: significant at 1%, 5%, 10%. Sample sizes for columns 2 and 6 apply to columns 2–5 and 6–7, respectively. Share of road classes are measured as a function of trip length. Motorways are the reference category. The reference category for trip type is radial trips. Weather indicators for rain (yes, no, missing), thunderstorms (yes, no, missing), wind speed (16 indicator variables), humidity (15 indicator variables), and temperature (5 indicator variables).

Table M.1: Trip statistics for US cities

	Mean	St. dev.	percentile: 1	10	25	50	75	90	99
Panel A: Sample: all trip instances (N=52,158,502)									
Speed	40.6	14.5	17.2	25.1	30.4	37.4	47.7	62.1	83.1
Duration	11.3	8.8	2.6	4.3	5.9	8.7	13.7	20.9	45.6
Duration (no traffic)	10.0	7.0	2.4	4.0	5.4	7.9	12.3	18.5	35.8
Trip length	8.6	9.5	1.4	2.1	3.1	5.2	10.3	19.1	47.0
Effective length	6.0	7.2	1.0	1.4	2.0	3.4	6.9	14.6	35.6
Panel B: Sample: all cities (N=139)									
Mean speed (distance weighted)	44.5	5.0	32.8	37.3	41.1	44.4	48.0	51.1	56.7
Mean speed	40.5	3.2	32.0	35.6	38.6	41.0	42.7	44.4	47.0
Mean duration	10.1	2.5	6.3	7.6	8.3	9.6	11.4	12.5	20.4
Mean duration (no traffic)	9.0	2.0	5.7	6.8	7.6	8.6	10.2	11.3	16.4
Mean trip length	7.5	2.4	3.8	5.1	5.7	7.2	8.9	10.5	16.6
Mean effective length	5.2	1.7	2.6	3.5	4.0	5.0	6.1	7.4	12.3

Note: Durations are in minutes, lengths in kilometers; and speeds in kilometers per hour.

Appendix M. US results

Tables M.1–M.4 report US analogs to tables 1, 2, 4 and 5. Table M.1 reports summary statistics for the US. At over 40 kilometers per hour on average, speeds are 70% higher than in our India sample. Only a small part of this gap is due to longer trips across spatially larger cities. US trips are 30% longer on average. Even with a large 20% elasticity of trip speed to trip length, differences in trip lengths would only cause US trips to be 5% faster than Indian trips. Trips are also comparably circuitous, with trip length exceeding effective (haversine) length by about 45% in both countries.

Table M.2 reports trip level regressions. While results are generally quite similar to India, we note several differences here. Trip length and city and time of day effects explain even more of the variation across trips in the US than in India. At about 0.3, the log trip length coefficient is 25–50% larger in the US. This is consistent with fast road classes, especially highways, which are more useful on longer trips, being more prevalent in the US. American highways also have a greater speed advantage over smaller roads. Downhill travel is faster in the US, unlike in India. Intersections slow US travel more, while business establishments and turns against traffic (left in the US) slow it less than in India.

Table M.2: Determinants of log trip speed in the US

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
log trip length	0.33^a	0.32^a	0.30^a	0.30^a	0.30^a	0.22^a	0.30^a	0.33^a
	(0.0061)	(0.0061)	(0.0062)	(0.0062)	(0.0059)	(0.0058)	(0.0036)	(0.0033)
log distance to center			0.075^a	0.077^a	0.080^a	0.094^a	0.077^a	0.065^a
			(0.0081)	(0.0085)	(0.0089)	(0.0091)	(0.0080)	(0.0075)
Gross gradient up						-0.42^c	-0.24^b	-0.48^a
						(0.21)	(0.12)	(0.16)
Gross gradient down						0.55	0.40	0.49^c
						(0.33)	(0.26)	(0.26)
Share primary roads						-0.29^a	-0.26^a	-0.23^a
						(0.0077)	(0.0092)	(0.0094)
Share secondary roads						-0.29^a	-0.25^a	-0.24^a
						(0.0088)	(0.012)	(0.011)
Share tertiary roads						-0.28^a	-0.25^a	-0.26^a
						(0.0080)	(0.0073)	(0.0079)
Share resid. roads						-0.36^a	-0.30^a	-0.31^a
						(0.0087)	(0.0095)	(0.0095)
Share other roads						-0.36^a	-0.35^a	-0.37^a
						(0.025)	(0.016)	(0.013)
Share missing roads						-0.28^a	-0.30^a	-0.30^a
						(0.0091)	(0.0095)	(0.011)
log # intersections							-0.10^a	-0.089^a
							(0.0062)	(0.0057)
arsinh # left turns							-0.055^a	-0.055^a
							(0.0015)	(0.0013)
arsinh # establishments								-0.043^a
								(0.0013)
Type: circumferential	0.016	0.022^c	-0.036^a	-0.035^a	-0.031^a	0.0056	0.011^a	-0.0081^b
	(0.013)	(0.013)	(0.0082)	(0.0081)	(0.0073)	(0.0044)	(0.0040)	(0.0037)
Type: gravity	0.070^a	0.076^a	-0.00066	0.00056	0.0032	0.025^a	0.027^a	0.010^a
	(0.011)	(0.011)	(0.0041)	(0.0042)	(0.0044)	(0.0036)	(0.0033)	(0.0029)
Type: amenity	0.032^b	0.038^a	-0.037^a	-0.038^a	-0.037^a	-0.011^a	-0.0071^b	-0.0026
	(0.013)	(0.013)	(0.0056)	(0.0056)	(0.0054)	(0.0037)	(0.0030)	(0.0030)
City effect	Y	Y	Y	Y	Y	Y	Y	Y
Day effect	Y	wkday	wkday	wkday	wkday	wkday	wkday	wkday
Time effect	Y	Y	Y	Y	Y	Y	Y	Y
Weight	N	N	N	intents	NHTS	NHTS	NHTS	NHTS
Weather	N	N	Y	Y	Y	Y	Y	Y
Observations	52,158,502	37,926,251	-	-	-	-	37,804,021	-
R^2	0.66	0.65	0.67	0.67	0.65	0.70	0.72	0.74

Notes: 180 cities in each column. OLS regressions with city, day, and time of day (for each 30 minute period) indicators. Log speed is the dependent variable in all columns. Robust standard errors in parentheses. a, b, c: significant at 1%, 5%, 10%. Sample sizes for columns 2 and 6 apply to columns 2–5 and 6–7, respectively. Share of road classes are measured as a function of trip length. Motorways are the reference category. The reference category for trip type is radial trips. Weather indicators for rain (yes, no, missing), thunderstorms (yes, no, missing), wind speed (16 indicator variables), humidity (15 indicator variables), and temperature (5 indicator variables).

Table M.3: Variance decompositions of our baseline speed index for the US

Sample	Cities	All trips			High peak trips		
		Uncongested speed	Congestion factor	Covariance	Uncongested speed	Congestion factor	Covariance
All	139	0.529	0.097	-0.187	0.388	0.179	-0.216
Smallest 50%	69	0.661	0.072	-0.134	0.454	0.189	-0.179
Largest 50%	70	0.496	0.112	-0.196	0.382	0.190	-0.214
Largest 25%	35	0.457	0.125	-0.209	0.382	0.183	-0.217
Largest 10%	14	0.479	0.116	-0.203	0.381	0.176	-0.221

Note: High peak hours are 4-5:30 PM.

Table M.3 shows that like in India, uncongested speed explains more variation across US cities than congestion. Unlike in India, this remains true even when restricting attention to high peak trips and the largest cities: in all ten samples shown, the variance share of uncongested speed is more than twice that of the congestion factor. The most notable difference is that the covariance term is considerably larger in magnitude in the US, at around -0.2 in nearly all samples. While cities with more congestion are slower in the absence of traffic in both countries, this correlation is stronger in the US. One possible explanation is that in both countries, features like shorter blocks and red lights slow down uncongested travel, and these features are more common in congested cities (e.g. in large and dense cities). However, there are more sources of variation across Indian cities, like paving quality, that are relatively uniform across American cities, diluting this correlation in India relative to the US.

Table M.4 reports correlates of city speed, uncongested speed and congestion, analogous to table 5. The main text notes that results for India and the US are generally similar and discusses three key differences between the two countries.

Table M.4: Correlates of city indices for the US

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent variable	Speed index			Uncongested speed			Congestion factor		
log population	-0.11^a	-0.13^a	-0.11^a	-0.074^a	-0.093^a	-0.078^a	0.038^a	0.041^a	0.037^a
	(0.025)	(0.025)	(0.024)	(0.020)	(0.020)	(0.019)	(0.0070)	(0.0068)	(0.0065)
log area	0.072^a	0.042^c	0.028	0.045^b	0.018	0.0091	-0.027^a	-0.024^a	-0.019^a
	(0.026)	(0.025)	(0.023)	(0.021)	(0.021)	(0.019)	(0.0072)	(0.0069)	(0.0065)
Elevation variance	0.0012	0.00097	0.00089	0.00034	0.00016	0.000083	-0.00086	-0.00080	-0.00080
	(0.0025)	(0.0022)	(0.0026)	(0.0019)	(0.0017)	(0.0018)	(0.0011)	(0.0011)	(0.0012)
Water length	-0.045^a	-0.038^a	-0.042^a	-0.028^a	-0.023^b	-0.026^a	0.016^a	0.015^a	0.015^a
	(0.011)	(0.014)	(0.013)	(0.0087)	(0.010)	(0.0095)	(0.0034)	(0.0042)	(0.0041)
log major roads		0.049^b	0.049^b		0.043^b	0.042^b		-0.0060	-0.0064
		(0.021)	(0.020)		(0.018)	(0.016)		(0.0062)	(0.0062)
Network		0.13^a	0.089^a		0.084^a	0.045^b		-0.046^a	-0.043^a
		(0.027)	(0.028)		(0.021)	(0.022)		(0.0078)	(0.0081)
Earnings			0.058			0.0058			-0.052^b
			(0.083)			(0.067)			(0.021)
Earnings^2			-0.016			-0.0070			0.0094^a
			(0.014)			(0.011)			(0.0035)
Observations	139	139	139	139	139	139	139	139	139
R^2	0.59	0.66	0.69	0.53	0.59	0.63	0.59	0.68	0.69

Notes: OLS regressions with a constant in all columns. The dependent variable of columns 1, 2, and 3 is the city fixed effect estimated in the specification of column 5 of table M.2. The dependent variable of columns 4, 5, and 6 is the city fixed effect of an analogous regression using uncongested speed as dependent variable. The dependent variable of columns 7, 8, and 9 is the city fixed effect of an analogous regression using the congestion factor as dependent variable. Robust standard errors in parentheses. a, b, c: significant at 1%, 5%, 10%. Log population is constructed using gridded WorldPop data. Elevation variance, water length, log major roads, and the network shape variable are as in table 5. Earnings (per capita) are computed using household income (measured in tens of thousands of dollars) and average household size from the 2015-2019 American Community Survey.

Appendix references

Ahlfeldt, Gabriel M. and Elisabetta Pietrostefani. 2019. The economic effects of density: A synthesis. *Journal of Urban Economics* 111:93–107.

Akbar, Prottoy A. and Gilles Duranton. 2018. Measuring congestion in a highly congested city: Bogotá. Processed, University of Pennsylvania.

Anderson, Simon P., André de Palma, and Jacques-Francois Thisse. 1992. *Discrete Choice Theory of Product Differentiation*. Cambridge, MA: MIT Press.

Atkin, David, Benjamin Faber, and Marco Gonzalez-Navarro. 2018. Retail globalization and household welfare: Evidence from Mexico. *Journal of Political Economy* 126(1):1–73.

Baugh, Kimberly, Christopher D. Elvidge, Tilottama Ghosh, and Daniel Ziskin. 2010. Development of a 2009 stable lights product using DMSP-OLS data. In *Proceedings of the Asia-Pacific Advanced Network*, volume 30.

Ben-Akiva, Moshe E. and Steven R. Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: MIT Press.

Boeing, Geoff. 2017. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems* 65:126–139.

Bondarenko, Maksym, David Kerr, Alessandro Sorichetta, and Andrew J. Tatem. 2020. Census/projection-disaggregated gridded population datasets, adjusted to match the corresponding unpd 2020 estimates, for 183 countries in 2020 using built-settlement growth model (BSGM) outputs. WorldPop, University of Southampton. URL `http://dx.doi.org/10.5258/SOTON/WP00685`.

Center for International Earth Science Information Network (CIESIN). 2016. Gridded population of the world, version 4 (GPWV4): Population count. NASA Socioeconomic Data and Applications Center (SEDAC). URL `http://dx.doi.org/10.7927/H4X63JVC`.

Corbane, Christina, Aneta Florczyk, Martino Pesaresi, Panagiotis Politis, and Vasileios Syrris. 2018. GHS built-up grid, derived from Landsat, multitemporal (1975-1990-2000-2014), R2018A. European Commission, Joint Research Centre (JRC).

Couture, Victor. 2016. Valuing the consumption benefits of urban density. Processed, University of California Berkeley.

Duranton, Gilles, Geetika Nagpal, and Matthew A. Turner. 2021. Transportation Infrastructure in the US. In Edward L. Glaeser and James M. Poterba (eds.) *Economic Analysis and Infrastructure Investment*. National Bureau of Economic Research, Inc.

Harari, Mariaflavia. 2020. Cities in bad shape: Urban geometry in India. *American Economic Review* 110(8):2377–2421.

Kreindler, Gabriel. 2018. The welfare effect of road congestion pricing: Experimental evidence and equilibrium implications. Processed, MIT.

Li, Yue, Martin Rama, Virgilio Galdo, and Maria Florencia Pinto. 2015. A spatial database for South Asia. Unpublished manuscript.

Pesaresi, Martino and Sergio Freire. 2016. GHS-SMOD R2016A - GHS settlement grid, following the REGIO model 2014 in application to GHSL landsat and CIESIN GPW v4-multitemporal (1975-1990-2000-2015). European Commission, Joint Research Centre.

Sheu, Gloria. 2014. Price, quality, and variety: Measuring the gains from trade in differentiated products. *American Economic Journal: Applied Economics* 6(4):66–89.

Small, Kenneth A. and Erik T. Verhoef. 2007. *The Economics of Urban Transportation*. New York (NY): Routledge.

United Nations, Department of Economic and Social Affairs, Population Division. 2019. *World Urbanization Prospects: The 2018 Revision (ST/ESA/SER.A/420)*. New York: United Nations.