Journal of Urban Economics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Urban Economics

journal homepage: www.elsevier.com/locate/jue

JUE Insight: Measuring movement and social contact with smartphone data: a real-time application to COVID-19☆

Victor Couture[a], Jonathan I. Dingel[b,*], Allison Green[c], Jessie Handbury[d], Kevin R. Williams[e]

[a] University of British Columbia, Canada
[b] University of Chicago Booth School of Business, NBER, and CEPR, USA
[c] Princeton University, USA
[d] University of Pennsylvania and NBER, USA
[e] Yale School of Management and NBER, USA

ARTICLE INFO

JEL classification:
C8
R1
R4

ABSTRACT

Tracking human activity in real time and at fine spatial scale is particularly valuable during episodes such as the COVID-19 pandemic. In this paper, we discuss the suitability of smartphone data for quantifying movement and social contact. These data cover broad sections of the US population and exhibit pre-pandemic patterns similar to conventional survey data. We develop and make publicly available a location exposure index that summarizes county-to-county movements and a device exposure index that quantifies social contact within venues. We also investigate the reliability of smartphone movement data during the pandemic.

1. Introduction

Personal digital devices now generate streams of data that describe human behavior in great detail. The temporal frequency, geographic precision, and novel content of the "digital exhaust" generated by users of online platforms and digital devices offer social scientists opportunities to investigate new dimensions of economic activity. The COVID-19 pandemic has demonstrated the potential for real-time, high-frequency data to inform economic analysis and policymaking when traditional data sources deliver statistics less frequently and with some delay.

In this paper, we discuss the suitability of smartphone data for quantifying movement and social contact. We show that these data cover a significant fraction of the US population and are broadly representative of the general population in terms of residential characteristics and movement patterns. We use these data to produce a location exposure index ("LEX") that describes county-to-county movements and a device exposure index ("DEX") that quantifies the exposure of devices to each other within venues. These indices track the evolution of inter-county travel and social contact from their sudden collapse in spring 2020 through their gradual, heterogeneous rises over the following months. Where possible, we compare these smartphone movement data to measures of population changes, expenditure, and travel during the pandemic. We do not find evidence that the dramatic pandemic-induced changes in behavior sharply altered the reliability of smartphone data.

We publish these indices each weekday in a public repository available to non-commercial users for research purposes.[1] Our aim is to reduce entry costs for those using smartphone movement data for pandemic-related research. By creating publicly available indices defined by documented sample-selection criteria, we hope to ease the comparison and interpretation of results across studies.[2] More broadly, this

[1] The indices and related documentation can be downloaded from https://github.com/COVIDExposureIndices.

[2] Examples of research using our indices thus far include Akovali and Yilmaz (2020), Althoff et al. (2020), Brinkman and Mangum (2020), Gupta et al. (2020), Monte (2020), Rodriguez et al. (2020), Wilson (2020), and Yilmazkuday (2020a,b).

☆ We are very grateful to Hayden Parsley, Serena Xu, and Shih-Hsuan Hsu for outstanding research assistance under extraordinary circumstances. We thank Drew Breunig, Nicholas Sheilas, Stephanie Smiley, Elizabeth Cutrone, and the team at PlaceIQ for data access and helpful conversations. The views expressed herein are those of the authors and do not necessarily reflect the views of PlaceIQ, NBER, nor CEPR. This research was approved by the University of California, Berkeley Office for Protection of Human Subjects under Protocol No 2018-05-11122 and the University of Chicago Institutional Review Board under protocol IRB20-0967. This material is based upon work supported by the National Science Foundation under Grant No. 2030056, the Tobin Center for Economic Policy at Yale, the Fisher Center for Real Estate and Urban Economics at UC Berkeley, the Zell-Lurie Real Estate Center at Wharton, and the Initiative on Global Markets at Chicago Booth.

* Corresponding author.
E-mail address: jdingel@chicagobooth.edu (J.I. Dingel).

https://doi.org/10.1016/j.jue.2021.103328
Received 3 July 2020; Received in revised form 15 January 2021
Available online xxx
0094-1190/

paper provides guidance on potential benefits and relevant caveats when using smartphone movement data for economic research.

Researchers in economics and other fields are turning to smartphone movement data to investigate a great variety of social science questions. Chen and Pope (2020) use similar smartphone data covering almost 2 million users in 2016 to document cross-sectional variation in geographic movement across cities and income groups. Athey et al. (2020) use smartphone data covering more than 17 million devices spanning January to April 2017 to document experienced segregation. We focus on the distinctive advantages of the data frequency and immediacy. A growing body of both theoretical and empirical research investigates human movement, social contact, and economic activity in the context of the COVID-19 pandemic.[3] Our indices provide empirical measures of these phenomena, complementing private-sector real-time measures of social distancing and movement.[4] We describe properties of smartphone data, compare the residential distribution and movement patterns of devices to those in traditional data sources, produce publicly available indices that can be used to easily compare results across studies, and investigate potential measurement issues that arise in the context of the ongoing pandemic.

2. Data

Our smartphone movement data come from PlaceIQ, a location data and analytics firm. In this section, we describe how PlaceIQ processes devices' movements to define visits to venues, and how we select the devices, venues, and visits included when we compute our exposure indices. We then compare these devices and their movements to residential populations and movements reported in traditional data sources.

2.1. Device visit data

PlaceIQ aggregates GPS location data from different smartphone applications using each device's unique advertising identifier. The raw GPS data come as pings that register whenever the application requests location data from the device.[5] These pings are joined with a map of two-dimensional polygons, corresponding to buildings or outdoor features such as public parks, which we denote "venues." A timestamped set of pings within or in the close vicinity of a polygon constitutes a "visit."[6] Since a device's location is measured with varying precision, PlaceIQ assigns each visit an attribution score based on ping characteristics and geographic features. We retain all visits with an attribution score greater than a minimum threshold. See Appendix A.1 for details.

2.2. Sample selection

2.2.1. Devices covered

For the typical smartphone in the PlaceIQ data, we observe about six months of movements, but there is considerable heterogeneity across devices. Each Android and iOS smartphone has an identifier that uniquely identifies the device at any given time, and the device's unique advertising identifier can be refreshed by the user and may be refreshed by some system updates. Thus, the average lifespan of an advertising identifier is less than that of a physical phone. Even devices observed over a long time period may not ping regularly. Ping frequency reflects a device's applications, settings, and movements.

To focus on devices whose movements can be reliably characterized, we restrict the set of devices included in the computation of our indices to those that pinged on at least 11 days over any 14-day period from November 1, 2019 through the reporting date.[7] The earliest date for which we report our indices is January 20, 2020, so this criterion selects a set of devices based on a window of at least 80 days of prior potential activity. Later reporting dates have longer windows. Given the reduced movement associated with the COVID-19 pandemic, a criterion using a fixed window of prior potential activity would exclude devices that temporarily reduced their movements. As of December 31, 2020, 75 million devices met this device selection criterion. On any given day, about 20 million of these devices ping at least once, as depicted in Fig. B.6.

For a subset of devices, we can assign a residential location with reasonable confidence based on the duration of their residential visits since November 1, 2019. Appendix A.2 describes our home assignment algorithm. In short, we assign home locations based on where devices repeatedly spend time at night. We use Census-reported demographic characteristics for block groups, which contain about 600 to 3,000 people, as proxies for device demographics. Since many people temporarily moved to other residential locations during the pandemic, we assign a device to a block group of residence based on the block group of its first home location after November 1, 2019. As of December 31, 2020, 64 million devices have an assigned block group of residence.

In the context of the COVID-19 pandemic, a potential concern is that devices may not generate pings when sheltering in place, due to their lack of movement. Indeed, there was a general decline in the number of devices generating pings in March 2020, presumably due to pandemic-induced declines in movement.[8] When defining our exposure indices in the next section, we discuss how they are impacted by devices sheltering in place and suggest potential adjustments.

Even absent a pandemic, the number of devices appearing in the data varies meaningfully over time. This may reflect changes in smartphone ownership patterns, smartphone device settings, app usage, PlaceIQ app coverage, seasonal variation in behavioral patterns, or an Android or iOS operating system update. These are unlikely explanations for the sharp decline starting in March 2020, as that decline coincides with the COVID-19 outbreak in the United States and there has not been a major OS update or major shift in PlaceIQ app coverage since the beginning of 2020. When publishing our indices, we also publish the number of devices underlying these values so that researchers can assess when changes in the exposure indices may not reflect true changes in behavior.[9]

2.2.2. Venues covered

Venues include commercial establishments, public parks, residential locations, and polygons lacking an identified business category. When assigning devices' homes, only residential locations are relevant. When

[3] In addition to the research using our indices, see Greenstone and Nigam (2020) on the value of social distancing, Maloney and Taskin (2020) on private social distancing, Brzezinski et al. (2020) on the effect of government-ordered lockdowns, Engle et al. (2020) on correlates of observed social distancing, Farboodi et al. (2020) on optimal policy, Glaeser et al. (2020) on cases and mobility, Almagro et al. (2020) on racial disparities in cases and commuting, and Xiao (2020) on the value of contact-tracing apps.

[4] For example, Unacast reports distance traveled; Google's community mobility reports capture visits to different venue types; and SafeGraph reports time spent at and away from home. Relative to these measures, our indices are designed to summarize travel and overlapping visits relevant for COVID-19 circumstances in an IRB-approved public release.

[5] The set of applications is not revealed to us. Some applications collect location data only when in active use, while others collect location data at regular intervals.

[6] If a device pings multiple times during a visit, then we have information about visit duration.

[7] During pre-pandemic months, using lower thresholds would only modestly increase the number of devices included.

[8] Devices are less likely to ping when users shelter in place because users are less likely to open movement-related apps that use location services and the phone's operating system may pause location services to save battery life. For example, the iOS "significant-change location service" only updates the user's position when it changes by at least 500 m (Apple, 2020).

[9] For example, the number of devices drops about 10 percent during April 14–18, 2020, which presumably reflects a change in smartphone data provision rather than a common change in behavior. Such variation will be absorbed by day fixed effects in difference-in-differences research designs.

tracking devices' movements across geographic units in the LEX, visits to all such venues are informative.

When measuring potential social contact by the DEX defined in Section 3, we restrict attention to venue categories in which most venues are sufficiently small that visiting devices would be exposed to each other. In particular, we omit the categories "Residential", "Nature and Outdoor", "Theme Parks", "Airports", "Universities", as well as venues without a category identified by PlaceIQ.[10] Finally, note that PlaceIQ excludes certain venue categories for privacy reasons, such as hospitals, schools, and places of worship.

There are 750,000 venues with identified commercial categories included in our DEX calculations. Since a venue corresponds to a building, certain types of buildings can belong to multiple categories, e.g., a restaurant inside a shopping mall. Our LEX calculations include venues in unidentified categories and residential locations, for a total of 149 million venues.

The identified venues in each commercial category are not necessarily representative of all such businesses. In most categories, the coverage of chains is high, but a much smaller share of independent businesses are identified.[11] Table A.2 reports the number of venues within each venue category in the DEX. The largest category is restaurants, which has about 200,000 distinct venues.[12] There is little variation in the number of venues from January to December 2020.

2.2.3. Locations covered

We report our indices for all US states and most US counties. Many US counties have few residents and therefore few devices in the PlaceIQ data. The indices we report are restricted to counties with reasonably large device samples. To implement this restriction, we assign each device to a unique daily "residential county", where that device had the highest (cumulative) duration of time at residential locations on that date. We report our indices only for the 2,018 counties that were the residential county of at least 1,000 devices on every day from January 6, to 12, 2020. These counties account for more than 96 percent of the US residential population.

2.3. Representativeness

Smartphone data cover a significant fraction of the US population. However, differences in smartphone ownership and app use, sample selection rules specific to research applications, and the use of small geographic units may produce unrepresentative samples.[13] For example, older adults are less likely to own smartphones, making smartphone-derived samples unbalanced across age groups.[14]

In this section, we compare the residential distribution and movement patterns of devices in our sample to those in traditional data sources. This analysis requires restricting our sample to devices assigned a residential block group, which constitute about 80 percent of the devices in our sample.[15]

Panel A of Fig. 1 shows that geographic units with larger residential population have more devices in our sample residing in them. Regressing the log number of devices on the US Census Bureau's 2019 estimate of log residential population yields an R^2 of 0.96 for states and 0.95 for counties. On average, the number of devices in our sample is about one-tenth of the total population.

Panel B of Fig. 1 investigates the distribution of devices across residential block groups within each county. The panel shows the share of devices living in block groups in ten population deciles ranked by income, share white, education, and population density. For instance, the top-right chart shows that about 10 percent of devices live in each decile of a county's block group median household income distribution. Similarly, about 10 percent of devices live in each decile when we rank block groups within their county by the share of their residents who are white or college graduates. When looking at deciles ranked by population density, denser block groups are somewhat underrepresented: only about 7 percent of devices live in block groups in the highest population-density decile.

In Appendix Figure B.1, we reproduce Panel B of Fig. 1 using national population deciles instead of within-county population deciles. We find greater overrepresentation of block groups with low population densities and large shares of white residents.[16] Given that our sample is more representative within counties than across counties, we suggest that researchers focus on applications of our indices that exploit intertemporal variation within counties or make cross-county comparisons of changes over time. Applications relying on cross-county differences in levels may be prone to sample-selection biases.

Panel C of Fig. 1 depicts residential migration patterns. We compare state-to-state residential migration in 2019 in our smartphone data to state-to-state flows in the 2017–2018 Internal Revenue Service (IRS) Migration Data. To make this comparison, we restrict attention to the 5.5 million devices in the PlaceIQ data with non-missing home assignments in both the first and last week of 2019. At the state level, the two migration measures are highly correlated: regressing the PlaceIQ share on the IRS share yields an R^2 exceeding 0.8. At the county level, the correlation is weaker, with an R^2 of 0.47.[17]

Panel D of Fig. 1 examines travel from home to commercial venues by depicting the distributions of trip lengths in our smartphone data and the 2017 National Household Transportation Survey (NHTS). For the PlaceIQ data, we show trips to venues included in the DEX computation.[18] For the NHTS, we show trips within the trip-purpose categories that most closely match DEX venues.[19] The figure depicts two trip-length distributions for each data source, one for people or devices living in block groups within the top quartile of the population density distribution, and one for people or devices living in the bottom quartile. The smartphone and NHTS trip-length distributions are remarkably similar, and both show a greater propensity to make shorter trips in more densely populated areas.

Overall, the patterns documented in Fig. 1 suggest the potential of broadly representative smartphone data for use in economic research. That said, we encourage researchers using these data to evaluate the precision and representativeness of their sample in their particular context. To help researchers assess whether our indices are suitably precise for

[10] Appendix C.1 presents DEX values for two alternative sets of venues. The first includes all identified commercial establishments, weighting them by inverse area. The second measures overlapping visits to residences.

[11] See Appendix C of Couture et al. (2020) for details.

[12] Note that many of these venues, such as shopping malls, contain multiple restaurant establishments. US County Business Patterns reports there were about 570,000 establishments in NAICS 7225 in 2017.

[13] SafeGraph, another location data provider, found that about 10 percent of block groups contain 30 to 40 percent of the devices in their data, leading to "disproportionately and sometimes impossibly high" numbers of devices relative to the Census-reported residential population (Squire, 2019).

[14] The Pew Research Center estimates that 81 percent of US adults own a smartphone. That rate varies from 96 percent for ages 18–29 to only 53 percent for those over 65 years. See https://www.pewresearch.org/internet/fact-sheet/mobile/.

[15] This restricted sample is the same that we will later use to compute our indices broken down by demographic group.

[16] When examining SafeGraph data, Squire (2019) reports the opposite pattern: SafeGraph data have fewer devices in block groups with more white residents. This suggests that representativeness may vary across smartphone data providers or sample-selection criteria.

[17] We exclude the 98% of county pairs that have no migration in both the IRS and smartphone data.

[18] A trip is from home if the device's previous visit was its home within the previous hour. We estimate driving distance (trip length) as 1.5 times the straight-line distance between the home and venue.

[19] These NHTS categories are "buy goods", "buy services", "buy meals", "other general errands", "recreational activities", and "exercise". We thank Gilles Duranton for computing the NHTS values in Fig. 1.

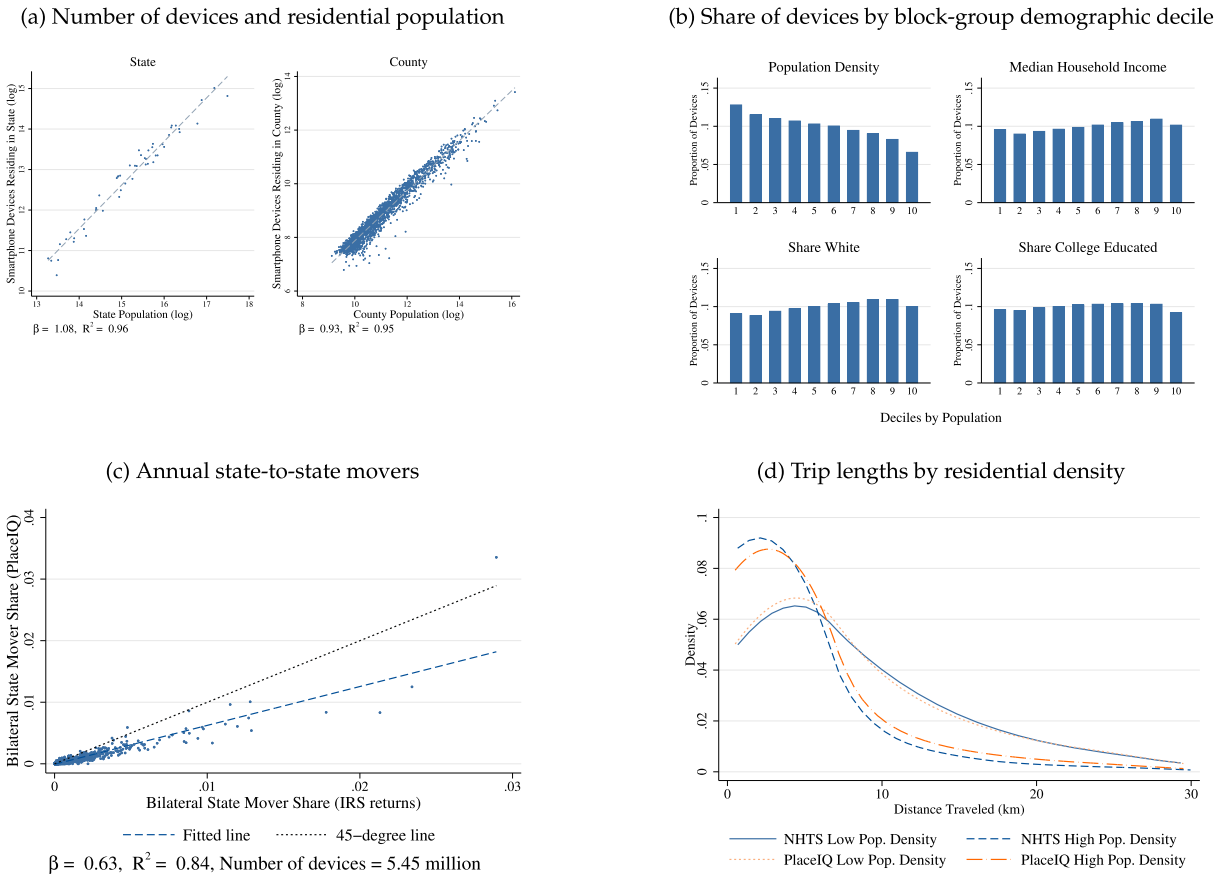

Fig. 1. Spatial and Demographic Balance of Device Populations. *Notes*: Panel A compares the number of devices residing in a geographic unit as of March 1, 2020 (vertical axis) to the Census's estimated 2019 residential population (horizontal axis) for all states, and for the 2,018 counties in the DEX and LEX. Panel B depicts the share of devices residing in block groups as of March 1, 2020 in each within-county decile of population density, median household income, share of white residents, and share of residents over 25 years with a bachelor's degree or higher. These block group characteristics are from the 2014–2018 American Community Survey. Panel C compares state-to-state residential changes in 2017–2018 IRS Migration Data to 2019 PlaceIQ data. The horizontal axis is the share of tax filers in state j who filed in state i the previous year. The vertical axis is the share of devices residing in state j in the last week of 2019 that resided in state i in the first week of 2019. Non-movers ($j = i$) are excluded. Panel D depicts a kernel density plot of trip length in kilometers, for trips from home to a commercial venue in the PlaceIQ data from November 2, 2019 through February 1, 2020 and in the 2017 NHTS, for residents of block groups in the top and bottom quartile of the population-density distribution.

their research application, we publish the underlying number of devices for each index, day, and geographic unit.

3. Exposure indices

In this section, we describe the location exposure index, which measures movement between counties or states, and the device exposure index, which measures average exposure of devices to each other within commercial venues.

3.1. Notation and preliminaries

We use the following notation when defining the LEX and DEX. Let i index devices, j index venues, g index geographic units (counties or states), and t and d index dates. Let $p_{ijt} \in \{0,1\}$ and $p_{igt} \in \{0,1\}$ equal one if device i pinged in venue j or geography g, respectively, on date t. Define $p_{it} \equiv \max_g p_{igt}$ as an indicator that equals one if device i pinged in any geographic unit on date t. Let $r_{igt} \in \{0,1\}$ equal one when device i resided in g at date t, where we assign residence based on the geographic unit in which the device spent the most time in residential venues on that date.[20]

[20] In the event of a tie, the geographic unit of residence is assigned based on visits to non-residential locations.

Next, we define sets of devices and venues based on these indicators. Let $\mathcal{I}_{j,d} \equiv \{i : p_{ijd} = 1\}$ and $\mathcal{I}_{g,d} \equiv \{i : p_{igd} = 1\}$ denote the sets of devices that pinged in venue j or geographic unit g, respectively, on date d. Let $\mathcal{G}_{g,d} \equiv \{i : r_{igd} = 1\}$ denote the set of devices that reside in geographic unit g on date d. Let $\mathcal{J}_{i,d} \equiv \{j : p_{ijd} = 1\}$ denote the set of venues where device i pinged on date d.

3.2. Location exposure index (LEX)

The LEX is a matrix that answers the following query: Among smartphones that pinged in geographic unit g' on date d, what share of those devices pinged in geographic unit g at least once during the previous 14 days? We report the LEX as a daily $G \times G$ matrix, in which each cell reports, among devices that pinged on day d in the column location g', the share of devices that pinged in the row location g at least once during the previous 14 days (conditional on pinging anywhere during the previous 14 days). Thus, each element of this matrix is

$$\text{LEX}_{gg'd} \equiv \frac{\sum_{i \in \mathcal{I}_{g',d}} \mathbf{1}\left\{\sum_{t=d-14}^{d-1} p_{igt} > 0\right\}}{\sum_{i \in \mathcal{I}_{g',d}} \mathbf{1}\left\{\sum_{t=d-14}^{d-1} p_{it} > 0\right\}} = \frac{\sum_i \mathbf{1}\left\{i : \left(p_{ig'd} = 1 \;\&\; \sum_{t=d-14}^{d-1} p_{igt} > 0\right)\right\}}{\sum_i \mathbf{1}\left\{i : \left(p_{ig'd} = 1 \;\&\; \sum_{t=d-14}^{d-1} p_{it} > 0\right)\right\}}.$$

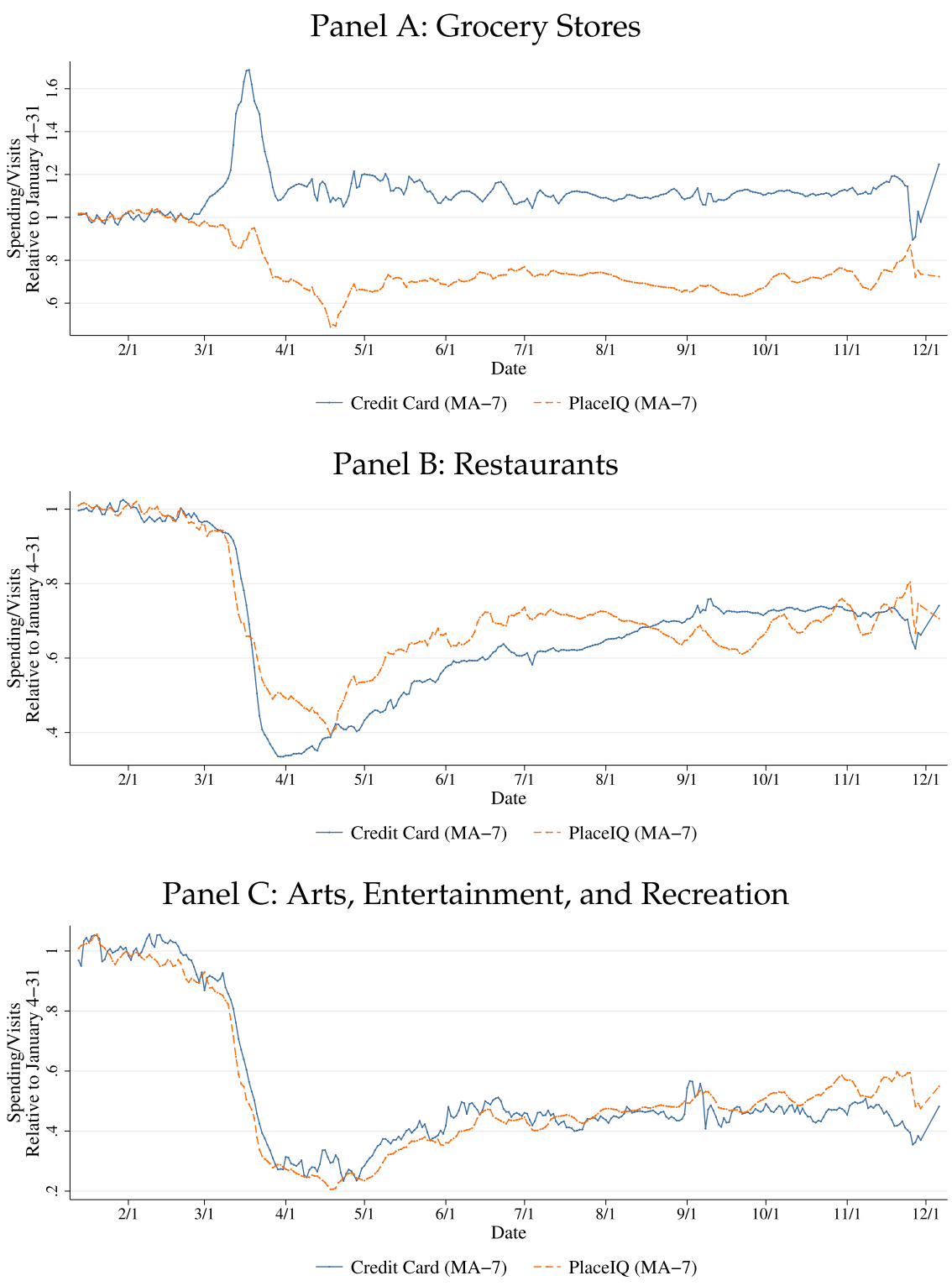

Fig. 2. Smartphone visits and Affinity expenditures *Notes*: This figure depicts total smartphone visits to grocery stores, restaurants, and arts, entertainment, and recreation (A&E). A&E includes visits to movie theaters, museums, nightclubs, bars, theme parks, and theatres. Credit card data for the same categories comes from Affinity Solutions (Chetty et al., 2020). Both series depict 2020 values relative to 2019 values normalized to the January 4–31 average and smoothed using a 7-day moving average.

We define the LEX to summarize people's movements with pandemic-related applications in mind. The index describes the share of people in a given location who have been in other locations during the prior two weeks. Thus, if COVID-19 cases surge in county g, $\text{LEX}_{gg'd}$ describes the potential exposure of county g' to the infectious disease via prior human movement from county g to g' (conditional on pinging anywhere in the US in the last 14 days). We chose the 14-day period of exposure based on the incubation period commonly cited by

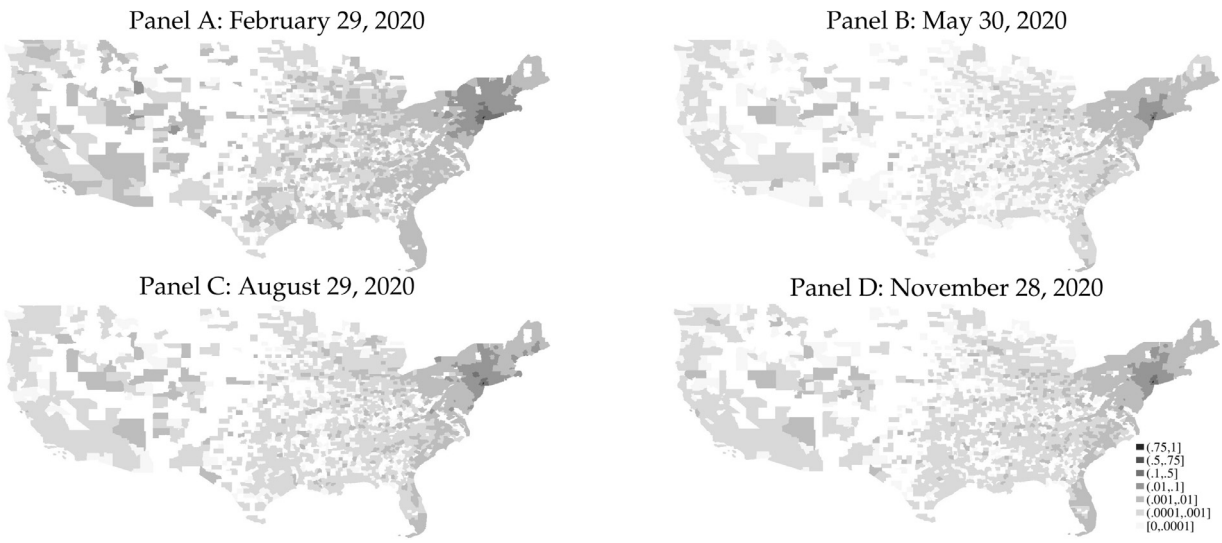

Fig. 3. County-Level Exposure to New York County (Manhattan). *NOTES*: Each panel of this figure depicts, for each of 2,018 counties, the share of devices pinging in that county that had pinged in New York, New York during the previous 14 days. The four panels depicts this for four Saturdays in 2020. Using the notation of Section 3, the four panels depict $LEX_{36061,g',d}$ for d equal to February 29, May 30, August 29, and November 28, of 2020, where 36061 is the FIPS code for New York County.

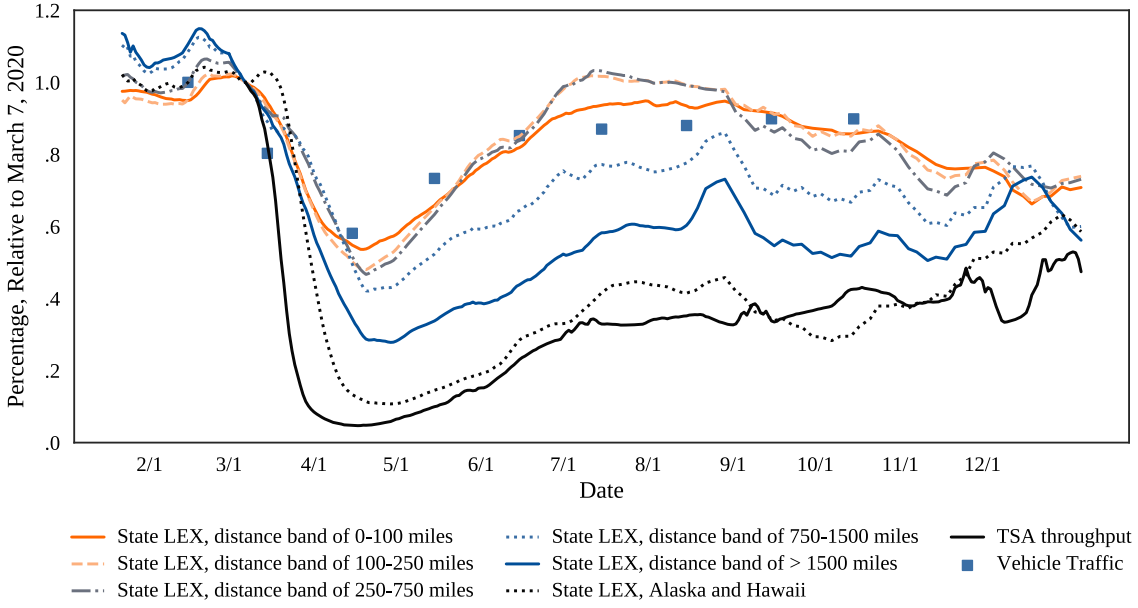

Fig. 4. State-level LEX values by distance between states. *NOTES*: This figure depicts average LEX values for pairs of states grouped by the distance between their population-weighted centroids. Each series depicts a 7-day moving average relative to its value on March 7, 2020. The Transportation Security Administration (TSA) throughput series reports the number of travelers passing through TSA checkpoints on each day. Monthly seasonally adjusted vehicle miles traveled comes from the Federal Highway Administration (series TRFVOLUSM227SFWA).

public-health authorities during the ongoing pandemic.[21] We chose to focus on all devices pinging in a given location rather than only residents because all human movement is relevant for potential disease exposure. Because a device can visit multiple locations both on a given day and during the preceding 14 days, LEX_d is not a transition matrix, its columns do not sum to one, and it is not amenable to aggregation. The temporal frequency and geographic units were selected to protect device user privacy in the context of a public data release. To complement the LEX, we also report a more aggregated statistic: the fraction of devices in geographic unit g' that in the last two weeks were in any geographic unit $g \neq g'$.

Starting in March 2020, there was a general decline in the number of devices generating pings, presumably due to individuals restricting their movements in response to the pandemic. Both the numerator and denominator of $LEX_{gg'd}$ restrict attention to devices that ping in g' on day d ($i \in \mathcal{I}_{g',d}$), so the LEX captures the locational histories of devices that are "out and about" in geographic unit g' on date d and does not capture the locational histories of devices sheltering in place and not generating any pings. This is relevant in the context of the ongoing pandemic: the index captures non-local exposure associated with "active" devices that are moving around within location g'. For applications that

[21] The CDC's COVID-19 FAQ page : "Based on existing literature, the incubation period (the time from exposure to development of symptoms) of SARS-CoV-2 and other coronaviruses (e.g. MERS-CoV, SARS-CoV) ranges from 2 to 14 days."

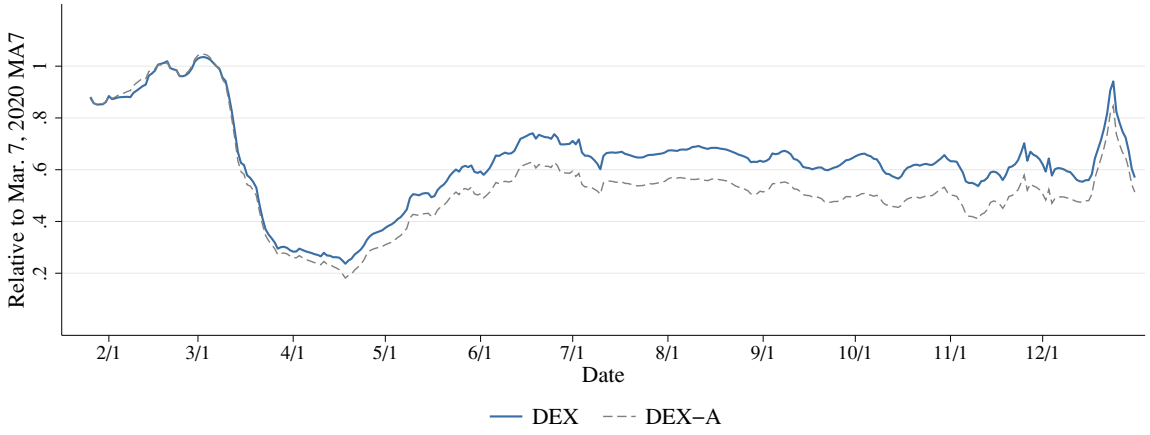

Fig. 5. DEX and DEX-A over time. *Notes*: This figure shows the population-weighted mean unadjusted and adjusted device exposure indices (DEX and DEX-A) over time. The series are smoothed using a 7-day moving average and normalized relative to their value of March 7, 2020.

require measuring exposure for the entire population of devices, including those that do not generate pings, we have published the daily number of devices that ping in each county, so that researchers can adjust their computations.

3.3. Device exposure index (DEX)

The DEX is a county- or state-level scalar that answers the following query: How many distinct devices does the average device living in g encounter via overlapping visits to commercial venues on each day? To compute the DEX, we first calculate the daily exposure set of device i as the number of distinct other devices that visit any commercial venue that i visits on date t:

$$\text{EXP}_{i,d} = \bigcup_{j \in \mathcal{J}_{i,d}} \mathcal{I}_{j,d}.$$

The DEX is then defined as the average size of the exposure set for devices that reside in geographic unit g on date d:

$$\text{DEX}_{g,d} \equiv \frac{1}{|\mathcal{G}_{g,d}|} \sum_{i \in \mathcal{G}_{g,d}} |\text{EXP}_{i,d}|.$$

As an average, the DEX can be aggregated to larger spatial units.[22] Note that the DEX values are necessarily only a fraction of the number of distinct individuals that also visited any of the commercial venues visited by a device, since only a fraction of individuals, venues, and visits are in the device sample.

We have defined the DEX to summarize social contact with pandemic-related applications in mind. The index captures overlapping visits to venues on the same day, which is relevant for potential virus exposure. We chose to define overlapping visits as visits to a venue on the same day rather than during the same hour based on both sample size and the concern that SARS-CoV-2 can persist in circulating air and on surfaces for multiple hours.

Note that devices sheltering in place would drop out of the sample used to compute the DEX if they did not generate any pings. As a result, the DEX may underestimate the reduction in exposure following the COVID-19 outbreak. We therefore implement a simple adjustment of the $\text{DEX}_{g,d}$ denominator as one means of addressing the potential sample selection problem associated with devices sheltering in place. Define a counterfactual set of pinging devices $\mathcal{G}^*_{g,d}$ such that any device in $\mathcal{G}^*_{g,d}$ but not in the observed $\mathcal{G}_{g,d}$ is sheltering in place with $|\text{EXP}_{i,d}| = 0$. The adjusted DEX is

$$\text{DEX}^{\text{adjusted}}_{g,d} = \frac{|\mathcal{G}_{g,d}|}{|\mathcal{G}^*_{g,d}|} \text{DEX}_{g,d}.$$

We assign the counterfactual set $\mathcal{G}^*_{g,d}$ to be the largest number of devices observed on any day from January 20, 2020 to February 14, 2020 in geographic unit g, so that

$$\widehat{|\mathcal{G}^*_{g,d}|} = \max_{d \in [\text{20 Jan 2020,14 Feb 2020}]} |\mathcal{G}_{g,d}|.$$

Given that $\widehat{|\mathcal{G}^*_{g,d}|}$ is an upper bound, $\text{DEX}^{\text{adjusted}}_{g,d}$ likely overestimates the drop in exposure following the COVID-19 outbreak. On the other hand, as noted above, the unadjusted $\text{DEX}_{g,d}$ likely underestimates the drop in exposure.[23] Together, these series should offer useful bounds. As mentioned before, even absent a pandemic there is meaningful variation in the number of devices in the sample that affect the DEX.

For devices that have a home assigned, we compute DEX values by the demographic characteristics of their residential block group. We only report these demographic DEX values at the state level, due to sample size and privacy considerations.

DEX by income Within each state g, we partition all census block groups into four median income quartiles with an equal number of block groups. We index these quartiles by $q \in \{1, 2, 3, 4\}$. Within each state g on each day d, we denote by $\mathcal{G}_{g,q,d}$ the set of devices i that have a home in a block group within quartile q.[24] The DEX by income is

$$\text{DEX-income}_{g,q,d} = \sum_{i \in \mathcal{G}_{g,q,d}} \frac{\text{EXP}_{i,d}}{|\mathcal{G}_{g,q,d}|}.$$

DEX by education The DEX by education is the same as the DEX by income, except that the four quartiles are based on the college share within each block group.[25]

DEX by race/ethnicity We report DEX values by racial/ethnic categories available in the Census of Population. For each $r \in$

[22] For example, metropolitan and micropolitan areas are defined as collections of counties.

[23] In practice, while the average absolute difference between the state-level unadjusted and adjusted DEX values is 7 percent, the two indices have a correlation coefficient of 0.996 in levels and 0.992 in first differences. Fig. 5 shows that the population-weighted mean values of the unadjusted and adjusted DEX track each other closely over time. The adjusted DEX should not be used when $|\mathcal{G}_{g,d}| > |\mathcal{G}^*_{g,d}|$, which will occur as social contact resumes and devices stop sheltering in place.

[24] Note that the residential block group is not necessarily within geographic-unit-of-residence g. This allows for cases where a device leaves their assigned home to shelter in place somewhere else. That is, a device can relocate, but it maintains its originally assigned demographics.

[25] The college share is the share of adults 25–65 years old with at least a four-year college degree.

{Asian, Black, Hispanic, White}, we report a weighted average of device-level exposure:

$$\text{DEX-race}_{g,d,k} = \sum_{i \in \mathcal{G}_{g,q,d}} \frac{w_{i,r}\text{EXP}_{i,d}}{\sum_{i \in \mathcal{G}_{g,q,d}} w_{i,r}},$$

where $w_{i,r}$ is the residential share of race/ethnicity r in device i's block group.[26]

4. Tracking activity during the 2020 pandemic

We present movement patterns captured by our smartphone data indices during the pandemic. These patterns generally align well with those found in other data sources when such comparisons are possible.

4.1. Comparisons to population and expenditure data

Given researchers' widespread use of smartphone data to study movement during the pandemic, it is important to assess whether the pandemic has altered the reliability of smartphone data. However, within one year of the virus spreading, there have been few opportunities to benchmark smartphone data to traditional data sources that are published less frequently and with a substantial lag. Even when traditional data are available, pandemic-induced changes in behavior may have caused smartphone movement data to diverge from the benchmark. As discussed in Section 2, sheltering in place reduces the ratio of active devices to residential population. Similarly, a shift to online shopping would alter the relationships between movement and expenditure. Nonetheless, in this section we compare the distribution of smartphone residences and visits during the pandemic to population and expenditure data.

First, we compare Census state-level population estimates for July 2018, July 2019, and July 2020 to state-level numbers of smartphones at those times.[27] Regressing the log smartphone population on the log Census population estimate for each state yields an R^2 of 0.97 for 2018 and 2019 and 0.94 for 2020.[28] Looking at the residential distribution of devices across block groups by their 2014–2018 demographic characteristics, Figure B.7 shows that the pre-pandemic tendency for devices to disproportionately reside in block groups with lower population densities and higher shares of white residents became slightly more pronounced during the pandemic. In sum, smartphone and Census residential counts diverged slightly more during the pandemic. This gap may reflect real movements not captured by Census methods rather than a decline in the reliability of smartphone data.

Second, we compare smartphone visits to expenditure data during the pandemic. Fig. 2 depicts a comparison of smartphone visits to credit card expenditure from Affinity Solutions (Chetty et al., 2020). We show data from January to December 2020 across three business categories: grocery, restaurant, and arts, entertainment, and recreation (A&E). For A&E trips and, to a lesser extent, restaurants, expenditure and smartphone visits show similar patterns of a sharp drop in late March and a slow recovery from April onward. Most at-home substitutes for A&E services belong to different expenditure categories, so the close relationship between movement and expenditure in this category is reassuring. For groceries however, changes in average grocery expenditure are unrelated to changes in smartphone visits to grocery stores. We conjecture that this divergence reflects changes in behavior, such as increased purchases of delivered groceries and greater expenditure per in-person visits.[29]

4.2. Movement between US states and counties

To illustrate the movement detail captured by the county-to-county LEX, we examine links to Manhattan (New York County), one of the early US epicenters of the pandemic. The maps in Fig. 3 depict the share of active devices in each US county that had pinged in Manhattan during the previous two weeks on the last Saturday of February, May, August and November 2020. The February panel shows a clear role for physical distance, as counties closer to Manhattan typically have a larger share of devices that have been in Manhattan during the previous two weeks, but it also makes clear that physical distance and county-to-county movements are distinct.

The LEX suggests a swift decline in travel between New York County and other counties at the pandemic's onset. From February to May 2020, Fig. 3 shows a broad decline in the share of active devices that had been in New York County during the previous two weeks. The decline was relatively greater in counties farther from New York City, making movements connected to New York County more spatially concentrated in the spring. These connections later rose, without returning to pre-pandemic levels. As noted previously, the LEX captures inter-county movement by active devices. Total inter-county movement also declined to the extent that fewer devices pinged due to not moving.

To assess the reliability of LEX values more systematically, we compare changes in state-level LEX values to measures of highway and airport traffic. We group pairs of states based on the distance between their population-weighted centroids and compute the daily mean value of $LEX_{gg'd}$ for each group. Fig. 4 depicts the mean daily LEX value using a 7-day moving average for each distance-defined group of state pairs relative to its value on March 7, 2020.

Fig. 4 shows differential declines in smartphone movements by distance that align well with the differential declines in vehicular and airport travel measures. Although the average LEX value declines for all state pairs through late April, pairs of states that are farther apart tended to exhibit larger relative declines. By mid-April, state-level LEX values at all distances were down 40 percent relative to their earlier levels. For comparison, monthly total vehicle-miles traveled, a measure that reflects both intrastate and interstate travel, fell by about 40 percent from February to April.[30] The steepest decline observed is for state pairs that include Alaska or Hawaii where across-state movements depend heavily on air travel even during the pandemic. The Alaska and Hawaii line closely tracks the decline in daily checkpoint totals at US airports reported by the Transportation Security Administration (TSA) two weeks earlier, as the LEX captures inter-state movements using a fourteen-day window. Inter-state travel at all distances began to rise in late April 2020, with short-distance travel peaking over the summer. Long-distance travel has continued to climb.

[26] To be precise, the categories "Asian," "Black," "Hispanic," and "White" are shorthand for non-Hispanic Asian, non-Hispanic black, all Hispanic, and non-Hispanic white residents. These four categories are sufficiently large to be reported for many geographic units. We only report the DEX-race for a given racial/ethnic group in states where the weighted number of devices for that group is at least 1000 devices every day from January 6, to 12, 2020.

[27] The US Census Bureau released state-level population estimates for July 1, 2020 on December 22, 2020. Population estimates for July 1, 2020 for smaller geographic units are scheduled to be published in the following six months. The Census estimates annual population changes based on births and deaths reported in vital statistics and migration evident in administrative data such as IRS tax returns and Medicare enrollments. Thus, the July 1, estimates reflect residential patterns across the various dates that households filed their tax returns.

[28] Looking at state-level population changes from July 2018 to July 2019 and from July 2019 to July 2020, we find an R^2 of 0.3 in both cases. The smartphone-data changes exhibit much greater variance than Census-data changes.

[29] Kim Severson, "7 Ways the Pandemic Has Changed How We Shop for Food," *New York Times*, 8 Sep 2020.

[30] We computed this figure using monthly seasonally adjusted vehicle-miles-traveled estimates from the Federal Highway Administration (series TRFVOLUSM227SFWA at https://fred.stlouisfed.org).

4.3. Visits to commercial venues

Fig. 5 traces the evolution of social contact over the course of the pandemic by plotting the population-weighted average of the county-level DEX values over 2020, relative to its level on March 7. Visits to commercial venues rose during February 2020, similar to behavior observed in February 2019. There is a sharp rapid decline in activity in March at the onset of the pandemic in the United States. The DEX reached a minimum in mid-April at about 25 percent of its early March level, then rose to just over 60 percent by mid-June. It remained around this level through most of the summer and autumn before rising rapidly in the final weeks of 2020.[31]

Some of this DEX variation is consistent with policy differences across jurisdictions. Appendix Figure B.5 depicts the evolution of the county-level DEX around policy events, controlling for county and time fixed effects. As in Brzezinski et al. (2020), we find that some of the DEX decline coincided with the timing of shelter-in-place orders, after which the DEX dropped by approximately 20 percent. Given the large number of potential confounding forces, these regressions are only suggestive.

The geographic and demographic detail of smartphone movement data should allow researchers to investigate important questions leveraging information not available in other data sources. For example, Fig. B.8 depicts DEX changes by educational attainment and race. This reveals limited differences in visits to commercial venues along these demographic dimensions. That may suggest a limited role for heterogeneous exposure rates within commercial venues in explaining differences across demographic groups infection and mortality rates during the pandemic.

5. Conclusion

These initial applications of our indices demonstrate the potential of smartphone movement data to quantify movement and social contact with high frequency and spatial precision. We have also articulated a number of caveats relevant for researchers using such data. We hope that our publicly available indices will support deeper and varied investigation of human movement during the ongoing pandemic.

Author statement

All authors contributed equally.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jue.2021.103328.

[31] Figure B.3 maps county-level DEX values on the last Saturday of February, May, August, and November 2020. Fig. B.4 plots the interquartile range of the DEX over time.

References

Akovali, U., Yilmaz, K., 2020. Polarized politics of pandemic response and the covid-19 connectedness across the us states. Covid Economics 57, 94–131.
Almagro, M., Coven, J., Gupta, A., Orane-Hutchinson, A., 2020. Racial disparities in frontline workers and housing crowding during covid-19: Evidence from geolocation data. Available at SSRN 3695249.
Althoff, L., Eckert, F., Ganapati, S., Walsh, C., 2020. The city paradox: Skilled services and remote work.
Apple, 2020. Getting the user's location. https://developer.apple.com/documentation/corelocation/getting_the_user_s_location/.
Athey, S., Ferguson, B.A., Gentzkow, M., Schmidt, T., 2020. Experienced Segregation. Working Paper 27572. National Bureau of Economic Research.
Brinkman, J., Mangum, K., 2020. The Geography of Travel Behavior in the Early Phase of the COVID-19 Pandemic. Technical Report WP 20-38. Federal Reserve Bank of San Philadelphia.
Brzezinski, A., Deiana, G., Kecht, V., Van Dijcke, D., 2020. The covid-19 pandemic: Government vs. community action across the united states. Covid Economics 7, 115–156. CEPR
Chen, M.K., Pope, D.G., 2020. Geographic Mobility in America: Evidence from Cell Phone Data. Working Paper 27072. National Bureau of Economic Research doi:10.3386/w27072.
Chetty, R., Friedman, J.N., Hendren, N., Stepner, M., The Opportunity Insights Team, 2020. The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data. Working Paper 27431. National Bureau of Economic Research doi:10.3386/w27431.
Couture, V., Gaubert, C., Handbury, J., Hurst, E., 2020. Income growth and the distributional effects of urban spatial sorting.
Engle, S., Stromme, J., Zhou, A., 2020. Staying at home: Mobility effects of Covid-19. Technical Report. Covid Economics: Vetted and Real Time Papers.
Farboodi, M., Jarosch, G., Shimer, R., 2020. Internal and external effects of social distancing in a pandemic.
Glaeser, E.L., Gorback, C., Redding, S.J., 2020. How much does covid-19 increase with mobility? evidence from new york and four other us cities. Journal of Urban Economics 103292.
Greenstone, M., Nigam, V., 2020. Does social distancing matter? Covid Economics 7, 1–22. CEPR
Gupta, S., Nguyen, T.D., Rojas, F.L., Raman, S., Lee, B., Bento, A., Simon, K.I., Wing, C., 2020. Tracking public and private response to the covid-19 epidemic: Evidence from state and local government actions. Technical Report. National Bureau of Economic Research.
Maloney, W.F., Taskin, T., 2020. Determinants of social distancing and economic activity during COVID-19: A global view. Technical Report 9242. World Bank.
Monte, F., 2020. Mobility zones. Economics Letters.
Rodriguez, A., Tabassum, A., Cui, J., Xie, J., Ho, J., Agarwal, P., Adhikari, B., Prakash, B.A., 2020. Deepcovid: An operational deep learning-driven framework for explainable real-time covid-19 forecasting. medRxiv.
Squire, R. F., 2019. Quantifying sampling bias in safegraph patterns. https://tinyurl.com/yb34h5p3.
Wilson, D.J., 2020. Weather, Social Distancing, and the Spread of COVID-19. Technical Report 2020-23. Federal Reserve Bank of San Francisco.
Xiao, K., 2020. Saving Lives Versus Saving Livelihoods: Can Big Data Technology Solve the Pandemic Dilemma? Technical Report. Available at SSRN 3583919.
Yilmazkuday, H., 2020. COVID-19 and Unequal Social Distancing across Demographic Groups. Technical Report. Available at SSRN 3580302.
Yilmazkuday, H., 2020. COVID-19 Deaths and Inter-County Travel: Daily Evidence from the US. Technical Report. Available at SSRN 3568838.