Device exposure index based on PlaceIQ data*

Victor Couture
Berkeley Haas

Jonathan I. Dingel
Chicago Booth

Allison Green
Princeton

Jessie Handbury
Wharton

Kevin Williams
Yale SOM

January 7, 2021

We use PlaceIQ data that describe smartphone devices "pinging" in a given geographic unit on a given day. We use these data to compute a device exposure index ("DEX"). We define the daily exposure of device i as the number of distinct devices that visit any commercial venue that device i visits on that day. The location-level DEX is a daily measure that answers the following query: on average, what is the exposure of devices residing in a given location?

We are making these indices publicly available to all researchers in the context of the spread of COVID-19. The notes below precisely define how we compute these indices so that researchers can use them appropriately.

Notation

- Index dates by d and t.
- Index devices by i.
- Index venues by j.
- Index geographic units by g. The number of geographic units is G. In the posted data, the geographic unit g indexes is a state, a county, or a CBSA.
- Let $p_{ijt} \in \{0,1\}$ be an indicator that is equal to one if device i pinged at venue j on date t.
- Let $\mathfrak{p}_{it} \in \{0,1\}$ be an indicator that is equal to one if device i pinged anywhere on date t, $\mathfrak{p}_{it} \equiv \max_j p_{ijt}$.
- Let $r_{igt} \in \{0,1\}$ be an indicator for device i geographic-unit-of-residence at date t. We set $r_{igt} = 1$ for the geographic unit g in which device i spent the most time at residential locations on date t.[1]
- Define the set of devices i that pinged at venue j on date d by $\mathcal{I}_{j,d} \equiv \{i : p_{ijt} = 1\}$.
- Define the set of venues j where device i pinged on date d by $\mathcal{J}_{i,d} \equiv \{j : p_{ijt} = 1\}$.
- Define the set of devices i that have geographic-unit-of-residence g on date d by $\mathcal{G}_{g,d} \equiv \{i : r_{igt} = 1\}$.

*We are very grateful to Hayden Parsley and Serena Xu for outstanding research assistance under extraordinary circumstances.

[1]The PlaceIQ data identifies visits to residential location separately. In the event of a tie in duration, the device is assigned to the geographic unit in which it made more distinct visits to residential locations. In the event of a tie in distinct visits, the device is assigned to the geographic unit in which the device had the highest duration of time at non-residential locations. In the event of a tie in non-residential duration, the device is assigned to the geographic unit in which it made more distinct visits to non-residential locations.

Definition of device exposure index

Device Exposure Set definition: The set of devices that visited any venue that device i visited on date d.

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

DEX definition: The average size of the exposure set for devices with geographic-unit-of-residence-of-residence g on date d.

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

Devices covered

We restrict the set of devices included in our calculations. Not all devices in the raw data ping regularly. Ping frequency reflects a device's applications, settings, and movements. To focus on devices whose (non-)movements can be reliably characterized (particularly after the onset of COVID-19), we restrict the set of devices included in our calculation of $\mathbf{DEX}_d$ to those that pinged on at least 11 days over any 14-day period from November 1, 2019 through date d. In order to be included in our calculation on date d, a device must satisfy the following inclusion criterion:

$$\exists d' : \left(\text{November 1, 2019} \leq d' \leq d - 13 \;\&\; \sum_{t=d'}^{d'+13} \mathfrak{p}_{it} \geq 11 \right)$$

Since the earliest date for which we report $\mathbf{DEX}_d$ is January 20, 2020, this criterion selects a set of devices on the basis of more than two months of prior activity. This window should be sufficient to select devices that ping regularly. For later dates, this criterion selects devices based on 14-day periods within a longer window of potential activity. Given the reduced movement associated with the ongoing pandemic, we worry that a criterion with a fixed window of potential activity would exclude devices that temporarily reduced their movements. Our criterion's growing window continues to include these devices based on their movements observed earlier.

Device coverage over time

In the context of the COVID-19 pandemic, a potential concern about the $\text{DEX}_{g,d}$ defined above is that devices may not generate pings when "sheltering in place", due to lack of movement. Both the numerator and denominator of $\text{DEX}_{g,d}$ restrict to attention to devices that pinged on day d. Since early March 2020, there has been a general decline in the number of devices generating pings, presumably due to individuals restricting their movements in response to the pandemic. We note that if sheltering-in-place devices with zero exposure drop out of our sample, $\text{DEX}_{g,d}$ may underestimate the reduction in exposure following the COVID-outbreak. We report the number of pinging devices with geographic-unit-of-residence in each geographic unit on each date (the denominator of $\text{DEX}_{g,d}$) so that researchers can account for this decline. Below, we propose one simple adjustment of the $\text{DEX}_{g,d}$ that addresses this potential bias.

Venues covered

We restrict the set of venue categories included in our calculations. We do not have data on venues' square footage for this release, so we restrict attention to categories of venues that should be small enough such that visiting devices are indeed exposed to each other. In particular, we omit the categories "Nature and Outdoor", "Theme Parks", "Airports", "Universities", as well as any location whose category is unidentified by PlaceIQ. Given the presumption that devices sheltering-in-place at their residence do not generate pings, we also omit the "Residential" category. Finally, note that certain categories of venues are excluded from the PlaceIQ data for privacy reasons, such as hospitals, schools, and places of worship.

The 25 categories that remain account for 900,000 venues that are overwhelmingly commercial venues. In most categories, the coverage of branded commercial venues is high, but we observe fewer independently operated establishments. For instance, the largest category is restaurants with 200,000 distinct locations corresponding to 370,000 establishments.[2] [3]

DEX by Demographic Groups

For a smaller set of devices, we can infer a device permanent block group of residence with high confidence. We assign a permanent block group of residence to a device based on the duration of its residential visits since November 1, 2019.[4] We can then assign demographic characteristics to each device based on 2014-2018 ACS data on the demographic composition of their permanent block group of residence. We only produce the demographic DEX at the state level and CBSA level, due to sample-size and privacy issues.

DEX by income Within each state/CBSA 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/CBSA g on each day d, we denote by $\mathcal{G}_{g,q,d}$ the set of devices i that have permanent residence in a block group within quartile q.[5] The DEX by income is then:

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

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.[6]

DEX by race/ethnicity We index each race/ethnicity by $r \in \{\text{Asian}, \text{Black}, \text{Hispanic}, \text{White}\}$. We define $w_{i,r}$ as the share of race/ethnicity r in the block group in which device i has its permanent residence. The

[2] We cannot distinguish movement within a building, so for some location types like malls, all of the establishments within the building are assigned to a single location identifier. We are able to see each establishment within the building, but we are not able to distinguish which of those establishments the person visited. For the purpose of calculating the DEX, we treat these buildings as a single venue.

[3] The US Census reports there were about 670,000 restaurants in the United States in 2017.

[4] We designate a device's weekly residence as the residential building the device spends the most time at night conditional on the device pinging on three different nights in a week. A device's permanent block group of residence is the census block group containing the device's earliest assigned weekly residence after November 1, 2019.

[5] Note that the permanent block group of residence is not necessarily within geographic-unit-of-residence g. This allows for cases where a device leaves their permanent residence to shelter-in-place somewhere else. For example, if a person permanently resides in a block group in New York corresponding to the bottom income quartile, and they move to Pennsylvania to shelter in place, that person is still assigned to the first income quartile but their state/CBSA of residence changes to Pennsylvania.

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

DEX by race/ethnicity is then:

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

Definition of adjusted DEX

We suggest a simple adjustment of the $\text{DEX}_{g,d}$ denominator as one means of addressing the potential sample selection problem associated with sheltering-in-place behavior described above. This adjustment is appropriate if sheltering-in-place devices with zero exposure drop out of our sample due to lack of movement, so that $\text{DEX}_{g,d}$ underestimates the decline in average device exposure following the COVID-19 outbreak.[7]

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:

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

The $\text{DEX}^{\text{adjusted}}_{g,d}$ that we report uses the largest number of devices observed on any given day from January 20th to February 14th in geographic unit g as the value of $|\mathcal{G}^*_{g,d}|$, 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, the unadjusted $\text{DEX}_{g,d}$ likely underestimates this drop in exposure, as explained above. We recommend that researchers employ both indices when investigating hypotheses concerning device exposure.

Locations reported

Many US counties have few residents and therefore few devices in the PlaceIQ data. The DEX matrices we report are restricted to counties with reasonably large device samples. In particular, we only report the DEX for counties and CBSAs with at least 1,000 devices every day of the week of January 6-12, 2020. Thus, we report DEX values for 2,018 of the more than 3,000 counties and 904 of the 946 CBSAs in the United States.

In a few states and CBSAs, the weighted number of devices $\sum_{i \in \mathcal{G}_{g,q,d}} w_{i,r}$ is low for some racial/ethnic group. 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 1,000 devices every day of the week of January 6-12, 2020.

Device and venue coverage over space

The full PlaceIQ device sample is meant to be nationally representative. However, our sample selection criteria mean that the devices used to compute exposure indices are unlikely to be fully representative of the residential population or daytime population on a county-by-county basis. A similar caveat applies to

[7] Alternative explanations for the decline in our device sample could include broad changes in smartphone device settings, changes in app usage, changes in PlaceIQ app coverage, or a major Android or iOS operating system update. As the timing of decline coincides with the COVID19 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, we consider these alternative explanations to be unlikely.

the set of venues in each county. Thus, while we expect within-county comparisons of the DEX over time to be informative, cross-county comparisons of DEX levels may reflect cross-county differences in device and venue coverage, rather cross-county differences in exposure.

A reminder

Phone movements and human movements are correlated but not synonymous. These data offer the opportunity to use the former as a proxy for the latter, with the necessary caveats.