Measuring mobility to monitor travel and physical distancing interventions: a common framework for mobile phone data analysis

Nishant Kishore, Mathew V Kiang, Kenth Engø-Monsen, Navin Vembar, Andrew Schroeder, Satchit Balsari, Caroline O Buckee

A surge of interest has been noted in the use of mobility data from mobile phones to monitor physical distancing and model the spread of severe acute respiratory syndrome coronavirus 2, the virus that causes COVID-19. Despite several years of research in this area, standard frameworks for aggregating and making use of different data streams from mobile phones are scarce and difficult to generalise across data providers. Here, we examine aggregation principles and procedures for different mobile phone data streams and describe a common syntax for how aggregated data are used in research and policy. We argue that the principles of privacy and data protection are vital in assessing more technical aspects of aggregation and should be an important central feature to guide partnerships with governments who make use of research products.

Lancet Digital Health 2020; 2: e622–28

Published **Online** September 1, 2020 https://doi.org/10.1016/S2589-7500(20)30193-X

Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T H Chan School of Public Health, Boston, MA, USA (N Kishore MPH, Prof C O Buckee DPhil)**; Center for Population Health Sciences, Stanford University School of Medicine, Stanford, CA, USA** (M V Kiang ScD)**; Telenor Research, Oslo, Norway** (K Engø-Monsen PhD)**; Department of Emergency Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, USA** (S Balsari MD)**; India Digital Health Network, Lakshmi Mittal and Family South Asia Institute, Harvard University, Cambridge, MA, USA** (S Balsari)**; Camber Systems, Washington, DC, USA** (N Vembar PhD)**; and Direct Relief, Santa Barbara, CA, USA** (A Schroeder PhD)

Correspondence to:
Prof Caroline O Buckee, Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T H Chan School of Public Health, Boston, MA 02115, USA
cbuckee@hsph.harvard.edu

Introduction

Data from mobile phones are being used around the world as part of the COVID-19 response, to map population movement, set parameters for disease transmission models, and inform resource allocation.[1–3] When anonymised and aggregated, these data do not reveal information about individuals but provide epidemiologically relevant estimates about population mobility—ie, the extent to which people are sheltering in place, congregating at parks, grocery stores and transit hubs, and generally moving less (or more) than usual.[3–5] These data also provide vital insights into travel patterns to help better understand the effect of travel restrictions and the risk of importation from other locations and to inform spatial epidemiological models.[6–8] These analyses can be used to identify neighbourhoods or communities that could become hotspots for community transmission or that might need additional support to practise physical distancing, or as part of surveillance more generally.[9]

Mobile phone data, although ubiquitous, have their biases and limitations.[4,5] Although one data stream might be more representative of a younger and more affluent population, another data stream could under-represent those living in rural areas. Popular analytical reports from large data providers show physical distancing (mobility) metrics, for example; however, the underlying data they represent and the aggregation methods used are typically not readily available.[1,10,11] This scant transparency makes it hard to know the representativeness and limitations of these data before using them for modelling. A common framework is needed to analyse the characteristics of these disparate data and their outputs, to allow for better comparison across mobility metrics and easier interpretation.

In this Viewpoint, we outline considerations for analysing aggregated data from mobile phones, including representativeness, situational context, and methods of aggregation. We then define the analytical pipelines used to construct nine metrics that can be applied towards measuring physical distancing interventions and modelling the spread of COVID-19 and other infectious diseases.

Data pipelines and processing

Data types

Two main types of data can be distinguished, based on how they are gathered. First, mobile operator data are obtained routinely in the form of either call detail records (CDRs) or a continuous stream of network signalling data. CDRs provide a cell tower identifier for calls, texts, or other uses associated with a SIM card, whereas network signalling data give continuous information about the cell tower that a handset is connected to as long as it is on. In both cases, the cell tower provides an approximate location for the user at the time of the call or text, with precision of the location dependent on the density of cell towers in the area. Second, global positioning system (GPS) traces are data obtained from smartphones and provide granular location data for the device over time. These data are precise with respect to location but are sometimes limited in terms of coverage and representativeness, particularly in low-income settings where many people do not have smartphones.[12] CDR data are generally more representative of the underlying population than are GPS traces (which are dependent on smartphones) because of the near-universal penetration of standard mobile phones. This notion generally remains true even in highly developed settings, because GPS data are typically only captured for a subset of the population that uses a particular application and provides consent to share location services. Ownership of more than one SIM and limited granularity of data do restrict the application of CDR for mobility analysis.[13,14]

For our analysis, we focus on GPS-derived metrics, which have been primarily used for population-level mobility analyses. Location data can, however, also be derived by triangulation from cell towers, from Bluetooth interactions, and from IP addresses via wi-fi connections.

GPS data sources can also be complex, with multiple overlapping applications, providers, publishers, and aggregators (appendix p 4).

See **Online** for appendix

Prerequisites to data sharing

Data providers are governed by national regulators who determine what CDRs (and sometimes GPS trace data) can be used for, and whether aggregated data can be shared with researchers or policy makers. For example, in most European countries, location data can only be used by the operator when they are made anonymous or with the consent of the individual, in accordance with General Data Protection Regulations. Establishing permission to use the data at all is generally the most lengthy step in the analytical pipeline. Discussions with regulators should, therefore, begin early; legal frameworks vary across jurisdictions and delays are to be expected, even during public health emergencies.[15,16]

For the **General Data Protection Regulations** see https://gdpr-info.eu

GPS trace data are routinely gathered by data publishers for commercial reasons. Publishers sell the data to brokers or aggregators, who might be able to provide insight into the representativeness of the data and their generalisability to the population. Partnership with data aggregators can reduce the technical burden of working with raw GPS trace data, which are typically massive, noisy, and require validation, and preclude the need for negotiating with every upstream data publisher. For example, Camber Systems (Washington, DC, USA), Cuebiq (New York, NY, USA), SafeGraph (San Francisco, CA, USA), and Facebook (Menlo Park, CA, USA) aggregate and preprocess data, providing the aggregates rather than raw trace data to researchers. The aggregation and preprocessing maintains privacy across the analytical pipeline. Data-use agreements and compliance with university ethics processes are also essential prerequisites for data sharing with researchers.[17]

The benefits of using personal data should outweigh the risks to privacy, even during a pandemic.[18] To date, epidemiological or clinical justification has not been satisfactorily shown to override privacy and ethical considerations in several settings where individuals have been deidentified by using mobile phone data.[19,20] Privacy should be preserved through statistical thresholds, differential privacy, and appropriate security controls with all parties in agreement on the principle of privacy protection.[21–23]

Representativeness of data

Data providers must be clear about the representativeness of their data for epidemiological research. There are at least three important considerations. First is market share: what fraction of the population are represented in these data? Second is demographic representativeness: who are the people generating the data, with respect to age groups, sex, race and ethnicity, and socioeconomic status, compared with the overall population? Third is geographical representativeness: generally, these data will be most representative of urban populations; understanding how well they represent rural communities is important to communicate to users of aggregated data.

Outliers should be discarded. Devices with an implausible number of calls, short duration between calls, or implausibly fast travel patterns are probably machines and do not represent human behaviour. The exact parameters for discarding outliers will depend on the population and the operator. Iterative communication between providers and data consumers is vital on these issues.

Summary measures of representativeness (or imbalance) can be used to compare data across different providers but are almost never shared currently. Providers could compare their internal data about user characteristics to a shared, public gold standard (such as the US Census). These measures would allow researchers to formally compare the representativeness of different providers. Importantly, this process preserves the privacy of users and providers by ensuring individual-level data are never shared.

Establishing baselines

An important component of almost all analyses will be a baseline against which to compare changes in travel and, in many cases, a home location for particular devices.[24] For example, in the COVID-19 context, mobility data are useful if compared against prephysical distancing or, now, comparing post-relaxation mobility with lockdown or prelockdown averages. Baselines can be established by making sure that the data analysis reaches back in time before the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) outbreak started and physical distancing interventions were put in place. Comparing data with the same time window in previous years accounts for important seasonal mobility patterns; however, most companies do not store data for such a long time because of data retention policies.[25,26] Clear communication of the baseline is important, including the uncertainty associated with it, so that decision makers can make sense of the changes they see in mobility.

Spatial and temporal aggregation

To preserve privacy, data should be aggregated (appendix pp 2–3). However, policy makers and researchers will typically request high-resolution disaggregated data. To strike a balance, data must be optimised to an actionable spatial boundary, such as an administrative zone or grid square, and on a timescale that can provide epidemiologically relevant information. The scale of administrative boundaries will depend on whether the location in question is rural or urban; cities sometimes need smaller spatial scales. Standard grid squares are, therefore, useful for some research questions. Timescales will also vary and the research questions should be considered; diurnal variation might be very useful to plot commute patterns, for example, whereas seasonal variation could show migration patterns related to agriculture, holidays, educational terms, or even weekday or weekend flows.[27]

The spatial scale needed will also depend on the policy or research objective. For example, city administrators might be interested in identifying hotspots of congregation or patterns in specific types of activity (eg, visits to grocery stores, transit hubs, and schools), whereas state governments might be more concerned with travel networks across administrative boundaries. If aggregators can dedicate resources in response to a crisis, they might consider generating analyses at varying spatial resolutions specific to each use case.

Disaggregation of data by sociodemographic information, such as age, gender, and race or ethnicity, should only be considered if there is a compelling reason. For example, it could be reasonable in the COVID-19 pandemic to disaggregate data by age to identify mobility patterns among people older than 60 years. Similarly, spatial aggregation must ensure that the numbers of unique devices do not fall below thresholds that enable re-identification of individuals or groups.[21,28] In general, grid cells with fewer than five to 20 users should not be shared with external parties.[29] Resetting the cohort of the analysis daily also protects privacy.[22,30,31]

Differential privacy should be applied to data releases. Differential privacy applies noise to the statistical computations to preserve individual and group privacy. It has been used by the US Census,[32] Apple,[33] Google,[34] and other organisations. Selecting appropriate parameters is implementation-dependent but must be undertaken with care.[29,35]

In large and quickly moving crises, the number of requests for data can scale up much more quickly than aggregator-specific capacity can respond. In this case, it is important that aggregators generate standard pipelines that provide data or metrics aggregated to an optimised spatial resolution, which advance analytical objectives without doing harm. When in doubt, automate lower resolution and maintain a human in the loop to provide access to higher resolution while maintaining privacy budgets.[36]

Mobility metrics and their relation to physical distancing and COVID-19 response

Metrics

Common metrics could be useful either for observing population mobility under different types of physical distancing interventions or as inputs for mechanistic models of disease spread (appendix p 1). Nearly all metrics calculated by CDR can be calculated with GPS traces; therefore, we have separated the metrics into those that can be calculated by both and those which are exclusive to GPS trace data.

Baseline metrics

Six baseline metrics are defined first.

- Population: a description of the unit of measurement that contributes data to the analysis such as unique users or unique devices. These are designated as *i* and *j*.
- Spatial resolution of categorisation: a description of the dimensions of the user locations that are used for categorisation. For example, unique locations that are visited by a user might be defined by tile grids, tower catchment areas, or GPS radii from points of interest for internal analysis. These are designated as *a* and *b*.
- Spatial resolution of aggregation: a description of the size of the regions of interest for which data are aggregated before being shared. For example, the metrics calculated for the population can be aggregated across all users in a region of interest such as neighbourhood or county. These are designated as *A* and *B*.
- Temporal resolution of categorisation: a description of the time bins that are used to categorise every user's location. For example, the modal location *a* that a user *i* logs data in every hour. These are designated as *t*.
- Temporal resolution of aggregation: a description of the time window for which data are aggregated for all users *i* in region *A*. For example, we might be interested in the average numbers of locations *a* (defined at top location per *t* min bin) over all space visited by users *i* in region *A* over the course of an 8 h time window. These are designated as *T*.
- Temporal thresholds: rather than calculating locations by top location *a* in every time bin *t*, providers may decide to calculate locations *a* as those where the user *i* spends at least a certain amount of time. This threshold is then designated at *T**.

Using these definitions, a provider might generate for every individual *i* a set of locations at spatial scale *a* for every time bin *t* in the time window *T*. The values for *a* can be directly mapped to larger regions of aggregation *A*. For example, for a given user, one can calculate their location as defined by presence in a 600 m × 600 m Bing Tile (*a*: zoom level 16)[37] for every 30 min segment (*t*) over the course of 24 h (*T*). Every Bing Tile is also mapped onto a county *A* for which all data are aggregated. We define this set of time-specific locations as

$$M_{iT} = \{a_{it1}, a_{it2}, \ldots, a_{itn}\}$$

for which a_{itn} refers to a specific bin *t* in time window *T*. Not all users will provide enough information to generate a full set M_{iT}; therefore, the provider should be transparent about any interpolation or imputation steps taken to make the set more robust.

Because of differences in spatiotemporal resolution, the definition of a stay location *a* will differ between CDR and GPS. For CDR, a stay location represents the tower, grid, or administrative region where a user's mobile phone was located. For GPS, depending on the spatiotemporal resolution, some preprocessing might take place. Typically, we can define a stay location as an area of size a within which all movement occurred for at least T^* time units (eg, a circle with a radius of 25 m within which all movement remained for at least 30 min). Different thresholds T^* can be used (eg, 5 min, 1 h, etc); the proper threshold will depend on both technological and computational constraints and the relevant question.

These key values can change depending on the metric, because of privacy-preserving objectives, or owing to computational limitations. However, the rationale of these values should be clearly communicated.

Metrics applicable to both CDR and GPS traces

Population distribution and dynamics

For most epidemiological analyses and modelling work, an estimate of the population residing in a specific region at a particular time provides an estimate of the denominator, which is the number of unique i that spend most of their time in a given area. To estimate this quantity, we assign every user to a home location, which is typically either the night-time location of the user or the area where the user spends the most time in the set of locations M_{iT}. In the case of continuous GPS data collection, one can also use location at midnight local time, or a range around that time, if data availability is not continuous. The sum of unique users in every region A is then used as the population estimate for that time window T. This value is denoted as

$$N_{AT} = \sum_{i} x_{iAT}$$

for which $x_{iAT}=1$, if the mode of M_{iT} for time T is in A and 0 otherwise.

Number of significant locations

The average number of significant locations provides an indication of how many distinct places users spend a substantial amount of time. Normal human mobility entails very few significant locations: usually, home, work, or school. However, varying shelter-in-place orders can result in different types of behaviours. For example, strict never-leave-home orders would result in a reduction of significant locations to 1, whereas less strict orders might result in increased numbers of local significant locations as individuals attempt to leave their homes more frequently but briefly.

To calculate the average number of significant locations for a specific population, we use the set of time-varying locations for every user i (M_{iT}) and create a subset of unique locations. Once the set of significant locations for every user has been estimated, the average across a region, grid, or other area of interest can be estimated as the sum of total unique locations visited by user i whose home location is in region A divided by the total number of users whose home location is in region A.

$$\overline{M_{AT}} = \frac{\sum_{i \in A} card(M_{iT})}{N_{AT}}$$

Transition between regions

This value provides an estimate of mobility between locations, which can be used in models to estimate the spatial spread of SARS-CoV-2. Transition matrices should include number, index, or proportion of unique users i who move from region A to region B. Users should contribute only once to the transition matrices within every time window to ensure that the numbers represent unique users and not trips between regions. As the time window considered decreases, researchers will be better able to understand within-day heterogeneity in movement between regions. These values can be used to calculate the percentage change in the total number of trips that occur between and within regions of interest, providing metrics of intraregional and inter-regional mobility.

It is important to note that this metric will vary with spatial scale and the time window considered. We recommend that T is, at most, 4 h for assessment of local travel networks, particularly those that might not cross time zones. Smaller time windows are unable to capture long-distance travel that takes more time than the length of the time window; however, it does allow for better understanding of within-day fluctuations of mobility. Larger time windows could be warranted for assessment of long-distance travel networks across time zones (eg, long-distance interstate travel) because the degree of displacement will typically not be clearly captured within shorter time windows.

This metric can be calculated by splitting a time window T into two halves ($T1$ and $T2$) then simply calculating the mode for each set (M_{iT1} and M_{iT2}), resulting in a_{iT1} and a_{iT2}. We assume that, for some time window T, the user transitioned from the modal location in $T1$ to the modal location in $T2$. If the user did not transition between locations, the matrix will include these counts on the diagonal. These vectors of transition are then summed across all users. Vectors with transitions that do not meet a minimum threshold are dropped. To aggregate to a larger spatial resolution, we map a to its corresponding A and sum the transition values for unique pairs of A and B.

Distances travelled

This metric measures the amount of movement occurring within a population and is calculated for all start and stop points of the vectors of travel for both CDRs and GPS traces. It is important to note that, for CDRs, this metric will be constrained by cell tower transitions and could have problems whereby two towers might route a call if a subscriber is between them and be upwardly biased in their estimates. As such, these data should not be interpreted as movement patterns. For GPS traces, distance travelled can be simple Euclidean or haversine distances calculated between points of the trail. For cases when car travel is obvious, various routing engines can be used to calculate on-ground transit distances. To make the traces useful, depending on the data source, care should be taken to remove points that represent impossible travel (accounting for different modes of transportation) and data that might be imputed and, thus, not representative of reality.[38] Average and total distances of these vectors are then weighted by the number of users who made the transition, providing a total or average distance moved by users in a given region.

$$\overline{D_i} = \sum_{l=2}^{n} |a_{il} - a_{il-1}|$$

for which

$$|a_{il} - a_{il-1}|$$

is the distance between location a_{il} and a_{il-1}. Another option would be to directly provide the distances between these locations as a value for each pair of locations, allowing researchers to aggregate and calculate as they see fit.

Radius of gyration

This metric provides a summary of travel that incorporates both the number of trips and the distance of every trip.[5,39] To calculate the radius of gyration for user i, first calculate the root mean squared distance of a user's movement across space over a given time window from their centre of gravity.

$$r_g(i) = \sqrt{\frac{1}{n}\sum_{l=1}^{n} |a_{il} - \bar{a}|^2}$$

for which the centre of gravity is

$$\bar{a} = \frac{1}{n}\sum_{l=1}^{n} a_{il}$$

Then for every user, generate their home region A as the region in which they spend the most time in their location set M_{iT}. Then, aggregate this value across a population in a given region and provide an average and percentiles.

Regularity of movement

In general, human mobility is highly predictable.[40–42] This predictability is important for urban planning, traffic forecasting, and public health. A formal measure of (un)predictability is location entropy. A low location entropy means an individual's time spent at their significant locations is highly predictable. Conversely, high location entropy suggests that predicting an individual's location is difficult. Therefore, the lowest location entropy would be achieved by a user who spends the exact same amount of time in the same places in every time window.

Using the set of locations defined above (M_{iT}), the Shannon entropy of user i can be calculated as

$$S_i^{rand} = \log_2 L_i$$

for which L_i is the number of distinct locations in M_{iT}. This measure assumes a person's location is uniformly distributed among all L_i observed distinct locations in M_{iT}. The uncorrelated Shannon entropy of the user i is

$$S_i^{unc} = -\sum_{k=1}^{L_i} p_k log_2 p_k$$

for which p_k is the frequency of the user's visit to their kth location (k is the index of all locations that the user visits). Other, more elaborate measures of location entropy have also been found to describe movement predictability well.[41]

GPS trace metrics

Average co-location with individuals in other regions

This value provides an indication of how much contact individuals from one region have with individuals from other regions. This analysis is restricted to GPS trace data but provides the most direct measure of contact between different populations. It is important that users who form the population for this metric meet a minimum threshold for data contributed during time T. First, for every user, calculate their set of locations for every time segment t in time window T.

$$M_{iT} = \{a_{it1}, a_{it2}, \ldots, a_{itn}\}$$

Depending on the completeness of the GPS traces, interpolation or imputation of user locations might need to be considered. Second, calculate every user's home region A as the region in which the user spends most of their time at night in a time window T. For situations when comparing two different regions A and B, calculate the probability of co-location as

$$\frac{\sum_{j\in B}\sum_{i\in A} y_{ij}}{N_{AT}\times N_{BT}\times \frac{T}{t}}$$

for which

$$y_{ij} = \sum_{l=1}^{n} \{a_{itl} = a_{jtl}\}$$

describes the total number of co-locations in a region of size a that users i and j have over the course of T. N_{AT} is the total population of users whose home location is in region A for time T, N_{BT} is the total population of users whose home location is in region B for time T, and T/t is the total number of time bins t that exist in time window T. For situations when comparing the same region, calculate the probability of co-location as

$$\frac{\frac{\sum_{j \neq i \in A} \sum_{i \in A} y_{ij}}{2}}{N_{AT} \times \left(\frac{N_{AT}-1}{2}\right) \times \frac{T}{t}}$$

This method has been led and implemented by Facebook's Data for Good team and provides direct measures of probability of contact between different regions in their population. For their metric, Facebook define a as Bing Tiles at zoom level 16, A as a county in the USA or an administration level 3 equivalent spatial area, t as 5 min, and T as a week. Unlike other metrics provided by Facebook, this one is independent for every time window and, therefore, does not have a baseline.

For more on the **Data for Good team** see https://dataforgood.fb.com

Measures of staying put

This metric is a direct measure of how much time people are spending in one location versus moving around and is relevant to measuring the effect of shelter-in-place policies and other strict lockdowns. This metric should be inversely related to the measures of mobility above (average distance travelled and radius of gyration). To calculate this measure, generate the full set of unique locations M_{iT} for every user i for a given time window T. Then assign every user to a home region A as the region a_{iT} where the user spends the most time in a time window T. Finally, count the number of unique locations for every user i in region A and calculate the proportion of all users i in region A who reported only one unique location during the time window T.

Measures of travel to points of interest (geofenced locations)

This metric will provide an indication of the nature of the travel that is being undertaken.[1,10] First, define a set of locations of size a in region A that are categorised as being locations of interest. These might include (but are not restricted to) parks, commercial areas, or grocery stores. These locations could either be grouped together in categories or specify different points of interest. Three steps are needed to calculate this metric. First, generate the full set of unique locations M_{iT} for every user i in a time window T. Second, assign every user to a region A based on the location a_{it} that the user i spends the most amount of time in. Third, for all users i in region A, calculate the proportion who visited a geofenced location in a time window T.

Conclusions

The COVID-19 pandemic has accelerated the use of aggregated mobility data from mobile devices, although without a universal governing framework for its application. Such data provide valuable insights, but without expertise and diligence it is easy to misinterpret these data, or cause harm, even if inadvertent.

As the COVID-19 pandemic continues, the metrics of interest and how they are used will also change. For example, our threshold of an optimal change in radius of gyration in response to a non-pharmaceutical intervention will be different now than when monitoring the same region for spikes in mobility 3 months from now.

We share this framework to advance a common language of comparison across these vast datasets. A shared language will allow us to synchronise future analysis with the limitations of every metric. Together, considerations provide insights for policy makers and could inform epidemiological models about physical distancing and the spatial spread of COVID-19. Combined with clinical and public health data, these metrics will have an important role in planning rollbacks of distancing because they help estimate the effect of various rollbacks on actual mobility patterns on the ground and, as a result, on epidemic spread.

Contributors

NK and COB had the idea for this Viewpoint, created all figures, and contributed to writing the manuscript. MVK, KE-M, NV, AS, and SB provided substantial subject matter expertise and contributed to writing the manuscript. COB and SB provided supervision and guidance.

Declaration of interests

NV is an employee of Camber Systems, a data analysis firm that produces statistical analyses of mobility data; future work by this company might use these metrics or similar metrics for commercial products. All other authors declare no competing interests.

Acknowledgments

Research reported in this Viewpoint was supported by the National Institute on Drug Abuse of the National Institutes of Health (K99DA051534 to MVK) and a National Institute of General Medical Sciences Maximizing Investigator's Research Award (R35GM124715-02 to COB). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1 Fitzpatrick J, DeSalvo K. Helping public health officials combat COVID-19. April 3, 2020. https://blog.google/technology/health/covid-19-community-mobility-reports/ (accessed Aug 7, 2020).

2 Facebook. Data for Good. https://dataforgood.fb.com/ (accessed Aug 7, 2020).

3 Palmer JRB, Espenshade TJ, Bartumeus F, Chung CY, Ozgencil NE, Li K. New approaches to human mobility: using mobile phones for demographic research. *Demography* 2012; **50:** 1105–28.

4 Wesolowski A, Buckee CO, Engø-Monsen K, Metcalf CJE. Connecting mobility to infectious diseases: the promise and limits of mobile phone data. *J Infect Dis* 2016; **214** (suppl 4): S414–20.

5 Williams NE, Thomas TA, Dunbar M, Eagle N, Dobra A. Measures of human mobility using mobile phone records enhanced with GIS data. *PLoS One* 2015; **10:** e0133630.

6 Wesolowski A, Eagle N, Tatem AJ, et al. Quantifying the impact of human mobility on malaria. *Science* 2012; **338:** 267–70.

7 Wesolowski A, Qureshi T, Boni MF, et al. Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proc Natl Acad Sci USA* 2015; **112:** 11887–92.

8 Jia JS, Lu X, Yuan Y, Xu G, Jia J, Christakis NA. Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature* 2020; **582:** 389–94.

9 Kishore N, Mitchell R, Lash TL, et al. Flying, phones and flu: anonymized call records suggest that Keflavik International Airport introduced pandemic H1N1 into Iceland in 2009. *Influenza Other Respir Viruses* 2020; **14:** 37–45.

10 SafeGraph. US consumer activity during COVID-19 pandemic: the impact of coronavirus (COVID-19) on foot traffic. Aug 6, 2020. https://www.safegraph.com/dashboard/covid19-commerce-patterns (accessed Aug 7, 2020).

11 UnaCast. Social distancing scoreboard. https://www.unacast.com/covid19/social-distancing-scoreboard (accessed Aug 7, 2020).

12 Fraiberger SP, Astudillo P, Candeago L, et al. Uncovering socioeconomic gaps in mobility reduction during the COVID-19 pandemic using location data. *arXiv* 2020; published online June 26. https://arxiv.org/abs/2006.15195v2 (preprint).

13 Wesolowski A, Eagle N, Noor AM, Snow RW, Buckee CO. Heterogeneous mobile phone ownership and usage patterns in Kenya. *PLoS One* 2012; **7:** e35319.

14 Buckee CO, Wesolowski A, Eagle NN, Hansen E, Snow RW. Mobile phones and malaria: modeling human and parasite travel. *Travel Med Infect Dis* 2013; **11:** 15–22.

15 Lomas N. Israel passes emergency law to use mobile data for COVID-19 contact tracing. March 18, 2020. https://social.techcrunch.com/2020/03/18/israel-passes-emergency-law-to-use-mobile-data-for-covid-19-contact-tracing/ (accessed Aug 7, 2020).

16 Kim MS. South Korea is watching quarantined citizens with a smartphone app. March 6, 2020. https://www.technologyreview.com/s/615329/coronavirus-south-korea-smartphone-app-quarantine/ (accessed Aug 7, 2020).

17 Buckee C, Engø-Monsen K. Mobile phone data for public health: towards data-sharing solutions that protect individual privacy and national security. *arXiv* 2016; published online June 2. https://arxiv.org/abs/1606.00864v1 (preprint).

18 Knight W. Phones could track the spread of Covid-19: is it a good idea? March 15, 2020. https://www.wired.com/story/phones-track-spread-covid19-good-idea/ (accessed Aug 7, 2020).

19 de Montjoye Y-A, Hidalgo CA, Verleysen M, Blondel VD. Unique in the crowd: the privacy bounds of human mobility. *Sci Rep* 2013; **3:** 1–5.

20 Rocher L, Hendrickx JM, de Montjoye Y-A. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun* 2019; **10:** 1–9.

21 Sweeney L. Simple demographics often identify people uniquely. Jan 1, 2000. https://kilthub.cmu.edu/articles/Simple_Demographics_Often_Identify_People_Uniquely/6625769 (accessed Aug 7, 2020).

22 Andrés ME, Bordenabe NE, Chatzikokolakis K, Palamidessi C. Geo-indistinguishability: differential privacy for location-based systems. Proceedings of the 2013 ACM SIGSAC conference on computer and communications security; Berlin, Germany; November, 2013: pp 901–14. https://doi.org/10.1145/2508859.2516735.

23 Xiong P, Zhu T, Pan L, Niu W, Li G. Privacy preserving in location data release: a differential privacy approach. In: Pham D-N, Park S-B, eds. PRICAI 2014: trends in artificial intelligence. Cham, Switzerland: Springer International Publishing, 2014: 183–95.

24 Massaro E, Kondor D, Ratti C. Assessing the interplay between human mobility and mosquito borne diseases in urban environments. *Sci Rep* 2019; **9:** 16911.

25 Wesolowski A, Metcalf CJE, Eagle N, et al. Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. *Proc Natl Acad Sci USA* 2015; **112:** 11114–19.

26 Wesolowski A, zu Erbach-Schoenberg E, Tatem AJ, et al. Multinational patterns of seasonal asymmetry in human movement influence infectious disease dynamics. *Nat Commun* 2017; **8:** 1–9.

27 zu Erbach-Schoenberg E, Alegana VA, Sorichetta A, et al. Dynamic denominators: the impact of seasonally varying population numbers on disease incidence estimates. *Popul Health Metr* 2016; **14:** 35.

28 Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. Proceedings of the 2008 IEEE Symposium on Security and Privacy; Oakland, CA, USA; May 18–21, 2008: pp 111–25. https://doi.org/10.1109/SP.2008.33.

29 Maas P. Facebook disaster maps: aggregate insights for crisis response & recovery. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; Anchorage, AK, USA; July, 2019: p 3173. https://doi.org/10.1145/3292500.3340412.

30 Ding Z, Wang Y, Wang G, Zhang D, Kifer D. Detecting violations of differential privacy. In: Proceedings of the 2018 ACM SIGSAC conference on computer and communications security; New York, NY, USA; January, 2018: pp 475–89. https://doi.org/10.1145/3243734.3243818.

31 ElSalamouny E, Gambs S. Differential privacy models for location-based services. *Trans Data Privacy* 2016; **9:** 15–48.

32 United States Census Bureau. Disclosure avoidance and the 2020 Census. June 19, 2020. https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html (accessed Aug 7, 2020).

33 Apple. Differential privacy. https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf (accessed Aug 7, 2020).

34 Statt N. Google is open-sourcing a tool for data scientists to help protect private information. Sept 5, 2019. https://www.theverge.com/2019/9/5/20850465/google-differential-privacy-open-source-tool-privacy-data-sharing (accessed Aug 7, 2020).

35 de Montjoye Y-A, Gambs S, Blondel V, et al. On the privacy-conscientious use of mobile phone data. *Sci Data* 2018; **5:** 180286.

36 Dwork C, Roth A. The algorithmic foundations of differential privacy. 2014. https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf (accessed Aug 26, 2020).

37 Brundritt R, Cai S, French C. Bing Maps Tile System. Feb 28, 2018. https://docs.microsoft.com/en-us/bingmaps/articles/bing-maps-tile-system (accessed Aug 7, 2020).

38 Rhee I, Shin M, Hong S, Lee K, Kim SJ, Chong S. On the Levy-walk nature of human mobility. *IEEE ACM Trans Netw* 2011; **19:** 630–43.

39 González MC, Hidalgo CA, Barabási A-L. Understanding individual human mobility patterns. *Nature* 2008; **453:** 779–82.

40 Song C, Qu Z, Blumm N, Barabási A-L. Limits of predictability in human mobility. *Science* 2010; **327:** 1018–21.

41 Lu X, Wetter E, Bharti N, Tatem AJ, Bengtsson L. Approaching the limit of predictability in human mobility. *Sci Rep* 2013; **3:** 1–9.

42 Qin S-M, Verkasalo H, Mohtaschemi M, Hartonen T, Alava M. Patterns, entropy, and predictability of human mobility and life. *PLoS One* 2012; **7:** e51353.