Improving mobility data for infectious disease research

Natalya Kostandova+ (1), Ronan Corgel (1), Shweta Bansal (2), Sophie Bérubé (1), Eimear Cleary (3), Chelsea Hansen (4, 5), Matt D.T. Hitchings (6, 7), Bernardo García-Carreras (8), Lauren Gardner (9, 1), Moritz U G Kraemer (10, 11), Shengjie Lai (3), Yao Li (1), Amanda C. Perofsky (4, 5), Giulia Pullano (2), Jonathan M Read (12), Gabriel Ribeiro dos Santos (13), Henrik Salje (13), Saki Takahashi (1), Cécile Viboud (4), Jasmine Wang (1), Derek AT Cummings* (6, 7), Amy Wesolowski* (1)

(1) Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, (2) Department of Biology, Georgetown University, Washington, District of Columbia, USA, (3) WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton SO17 1BJ, UK, (4) Fogarty International Center, National Institutes of Health, Bethesda, MD, USA, (5) Brotman Baty Institute, University of Washington, Seattle, Washington, USA, (6) Department of Biostatistics, University of Florida, Gainesville, USA, (7) Emerging Pathogens Institute, University of Florida, Gainesville, USA, (8) Department of Biology, University of Florida, Gainesville, FL, USA, (9) Department of Civil and Systems Engineering, Johns Hopkins University, Baltimore, MD, 21218, USA, (10) Department of Biology, University of Oxford, Oxford, United Kingdom, (11) Pandemic Sciences Institute, University of Oxford, Oxford, United Kingdom, (12) Centre for Health Information Computation and Statistics, Lancaster Medical School, Lancaster University, Lancaster, United Kingdom, (13) Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK

+Corresponding author: Natalya Kostandova, nkostan1@jh.edu

Standfirst

Mobility data can help reconstruct infectious disease dynamics and tailor control and elimination measures. We describe three challenges and opportunities to improve our understanding of human mobility for infectious disease research. We call for simulation and modeling, reporting guidelines, and investment in data repositories.

Main Text

Quantifying human mobility is critical to develop a more complete understanding of how travel and contacts between populations facilitate the spread of infectious pathogens. Information about human mobility, which can include both presence and absence of travel on a wide range of temporal and spatial scales, has been used in reconstructing and predicting transmission dynamics, and determining the effectiveness of control strategies for a wide range of diseases, including influenza, malaria, Ebola, cholera, and SARS-CoV-2. Since the beginning of the COVID-19 pandemic, there has been an increased availability of data describing different aspects of human mobility. These data are used in modeling (theoretical representations of the transmission system) and simulations (multiple iterations of representations of a transmission system). However, more work is needed to address the challenges and opportunities of evaluating how, when, where, and for which populations generalizable insights on mobility can be applied to various epidemiological scenarios to prepare for future pandemic and epidemic threats.

Challenges in use of mobility data

Challenge 1: Identifying which type of mobility data and metric are the best proxies of potentially infectious interactions and risk. Along the spectrum of pathogen transmission states from pre-emergence to elimination, mobility data can provide valuable information about how and where contact potentially occurs between susceptible and infected individuals (Table 1). Ideally, we would be able to quantify all interactions between susceptible and infected individuals, however in reality we often use proxy measures that describe travel or mixing patterns between individuals from different locations without knowing their disease status. These data have been applied for multiple purposes, such as using population commuting information to refine estimates of community-level vaccine coverage[1], or using mobile phone data to quantify mobility between administrative units in Brazil in a metapopulation model used to calculate schistosomiasis prevalence[2]. Given the wide range of applications, different transmission processes, and specific modeling purpose there may not be a single source of mobility data or metric that can capture the relevant infectious interactions. There is no single source 'gold standard' data set that will be universally relevant. Understanding when, where, and for whom mobility data are available will be context-specific and difficult to generalize, and hence there is no standard guidance. Ideally, each mobility data set and a range of mobility metrics would be compared against measures of disease transmission across a wide range of settings and for various pathogens. However, often the disease transmission data are sparse, biased, or unavailable, making these analyses not possible. Further, many analyses are ecological where a general, population-level metric of mobility is used. This can further complicate our ability to associate disease and mobility and potentially result in ecological bias[3]. While ideal datasets used in analyses depend on the questions being asked and the populations of interest, a better understanding of how to map mobility data to a wide range of transmission-relevant behaviors in more diverse populations and geographies is needed.

Challenge 2: Harmonization, integration, and availability of multiple data sources. Often, 'mobility' is considered a catch-all term and is used as a proxy measure for the connectivity between locations, or to encompass behaviors that may expose a susceptible individual to an infectious individual. However, different mobility datasets report a wide range of mobility-related behaviors, including the number of trips between pairs of locations, clustering of individuals in a specific location, percentage of devices staying home, and contact rates, among others. The way data are collected (e.g., self-response on a questionnaire, tracking GPS location on a mobile phone) also varies across data sets. Often, the temporal and spatial aggregation processes used for different datasets vary in their approach, and many datasets are censored based on a minimum population number for a particular spatial resolution to minimize privacy concerns. These differences make it challenging to harmonize and compare across data types, while the choice of appropriate data source should be determined by research objective. For many purposes, however, no single dataset may be able to provide appropriate level of coverage or granularity of population mobility data. However, with growing data availability in the public domain, it may be possible to combine different datasets to create single estimates of mobility or for example translate estimates from contact data to flows. Nevertheless, to date, there has been limited research and evaluation of which statistical and mathematical approaches, including simulation, should be used to integrate and evaluate disparate data sets.

Challenge 3: Accounting for sampling and measurement bias. There remain gaps in our understanding of mobility across locations and populations. The data collection and aggregation processes may result in datasets that are not representative of the population of interest, as underscored by many recent studies that quantify the representativeness of different data sources and the impact of spatiotemporal sampling accuracy on type of mobility captured. For example, few data sets include information on the movement of children or the very elderly or describe patterns in many low- and middle-income settings[4]. Data collected by mobile phone providers require mobile phone ownership and use, while network penetration rates vary by geographical location. Some mobility data obtained via social media applications rely on a combination of internet use, smartphone access, and opting in to share location information. These may result in sampling bias and issues of representativeness, when access to and use of technologies used to create the data sets is associated with mobility patterns. Often there is limited accompanying metadata to describe differences in mobility across groups or individuals, such as by age or socio-economic status[5], which prevents correcting these biases. However, these concerns need to be balanced with data protection measures to preserve the privacy of individuals. Finally, data may only capture one aspect of travel, such as trip counts, and ignore other aspects, such as trip frequency or duration.

Opportunities for future uses

Despite these challenges, researchers have multiple opportunities to leverage existing resources to provide a step change connecting mobility and transmission.

Opportunity 1: Use simulation and modeling to better understand data needs and uses across pathogens. Simulation and modeling can help us to better understand how various mobility data sets may result in different spatio-temporal disease dynamics. Assessing multiple competing models of mobility in infectious disease models can help evaluate the utility of data over a null model or assumption, i.e., including no connectivity between populations or a basic spatial interaction model [preprint[6]]. Further evaluation of a hierarchy of model complexity, i.e., from basic assumptions of no mixing to simple non-parameterized spatial interaction models to parameterized mobility models, can provide insight into how and what aspects of human behavior drive transmission. Finally, this approach can serve as an additional methodology for propagating uncertainty throughout transmission-modeled simulations by allowing for the evaluation of uncertainty due to model misspecification. Using simulation to road-test mobility data and develop better-informed estimates of connectivity driving disease transmission can help build a more generalized understanding of human mobility for future applications. For example, simulations can be used to explore the impact of data censoring and aggregation on modeled transmission dynamics, guiding data needs and requirements for particular use cases. As an illustration, in Figure 1, we display the results from a simple model, where we explored the impact of demographic bias, censoring, and temporal and spatial aggregation on estimated arrival times in a simulation (see Github for details). Explorations can be conducted to examine how these factors would impact disease inference using simulation as a guiding principle.

Opportunity 2: Reporting guidelines to increase the interpretation, standardization, and reuse of existing mobility data sets. Interpreting results from models that integrate mobility data, and determining how these data may be relevant for future applications, can benefit from a framework that makes it easy to understand how and what type of data were used. However, there are no established guidelines or systematic structure for reporting how mobility data are pre-processed, incorporated into infectious disease models, and reported in the results of analyses. This makes it difficult not only to understand the way mobility data were used and to compare results across published research, but also to reproduce results. Here, we propose a specific set of reporting guidelines to be included by researchers using mobility data (Supplementary materials). These guidelines cover key components of mobility data origin, analysis, and use, including a description of raw mobility data, pre-processing by the data provider, processing of the data carried out by the researcher, how mobility data were integrated into transmission analyses, and the results that should be presented. Use of these guidelines will allow for consistency in presenting mobility data, easier communication of results, and improved understanding of the quality and scope of mobility data used. Guidelines can also increase reproducibility and enable the integration of multiple data sets, which can, in turn, provide opportunities for identifying and accounting for bias due to sampling or measurement between datasets. Building a framework off the FAIR principles could further help mitigate future problems.

Opportunity 3: Investment in data repositories to ensure continuity of data access. Following a dramatic increase in availability of infection and mobility data from various sources during the SARS-CoV-2 pandemic, data access has since become increasingly limited or costly. For example, updates to Data for Good at Meta: Mobility Data, a public data source providing county-level information on relative travel, were discontinued as of December 31, 2020. Similarly, Google COVID-19 Mobility Reports are no longer updated as of October 15, 2022, while SafeGraph stopped providing social distancing metrics, including percent time staying at home, in April 2021. There is a need to maintain the availability and ease of access to historical data, especially as new opportunities and developments in data generation, synthesis, and analysis emerge. This may be of particular importance with further development of machine learning approaches, which may provide opportunities for creation and availability of large-scale synthetic trajectory datasets without the costs and risks to privacy associated with more conventional sources of data. The use of repositories and specialized R packages could facilitate easy access. However, continued access requires advocacy with governments and other decision-makers. This must address the barriers to making the data available, such as funding human resources necessary to collect, maintain, and manage the data, and should use evidence to make the case for why data should continue to be collected. To avoid inappropriate allocation of resources, we must identify the most important data to collect to prioritize. Investment in data repositories would allow for better assessment of data gaps to target empirical resources to populations that are under-represented in datasets but key to understanding transmission dynamics. Central repositories could also pave the way to expanding data and resource access beyond a select set of research groups, which has historically been the case due to proprietary access.

Conclusion

Since the beginning of the COVID-19 pandemic, there was an unprecedented increase in access to, and use of, data describing human mobility patterns to evaluate and guide public health interventions. Leveraging these resources requires addressing challenges of access, transparency, and reproducibility, as in genomic data analyses and the push for open and easily accessible data. Further methodological challenges to concretely tie together mobility and transmission that could be strengthened using simulation, reporting guidelines to improve standardization, and sustainable data repositories. These steps are crucial in leveraging the exciting promise of mobility data to better understand and mitigate future disease outbreaks and epidemics in an increasingly connected world.

References

1. Delamater, P.L., Leslie, T.F., Yang, Y.T. & Jacobsen, K.H. *Appl. Geogr. Sevenoaks Engl.* **71**, 123–132 [2016].
2. Mari, L. et al. *Sci. Rep.* **7**, 489 [2017].
3. Wakefield, J. *Annu. Rev. Public Health* **29**, 75–90 [2008].
4. Wardle, J., Bhatia, S., Kraemer, M.U.G., Nouvellet, P. & Cori, A. *Epidemics* **42**, 100666 [2023].
5. Lenormand, M. et al. *Sci. Rep.* **5**, 10075 [2015].
6. Pullano, G., Alvarez-Zuzek, L.G., Colizza, V. & Bansal, S.2023.11.22.23298916 [2023].doi:10.1101/2023.11.22.23298916
7. Ramadona, A.L., Tozan, Y., Lazuardi, L. & Rocklöv, J. *PLoS Negl. Trop. Dis.* **13**, e0007298 [2019].
8. Tegally, H. et al. *Cell* **186**, 3277-3290.e16 [2023].
9. Kondo, K. *Sci. Rep.* **11**, 18951 [2021].
10. Ruktanonchai, N.W. et al. *Malar. J.* **15**, 273 [2016].

Competing interests:

The authors declare no competing interests.

Table 1. Use cases of mobility data in public health response.

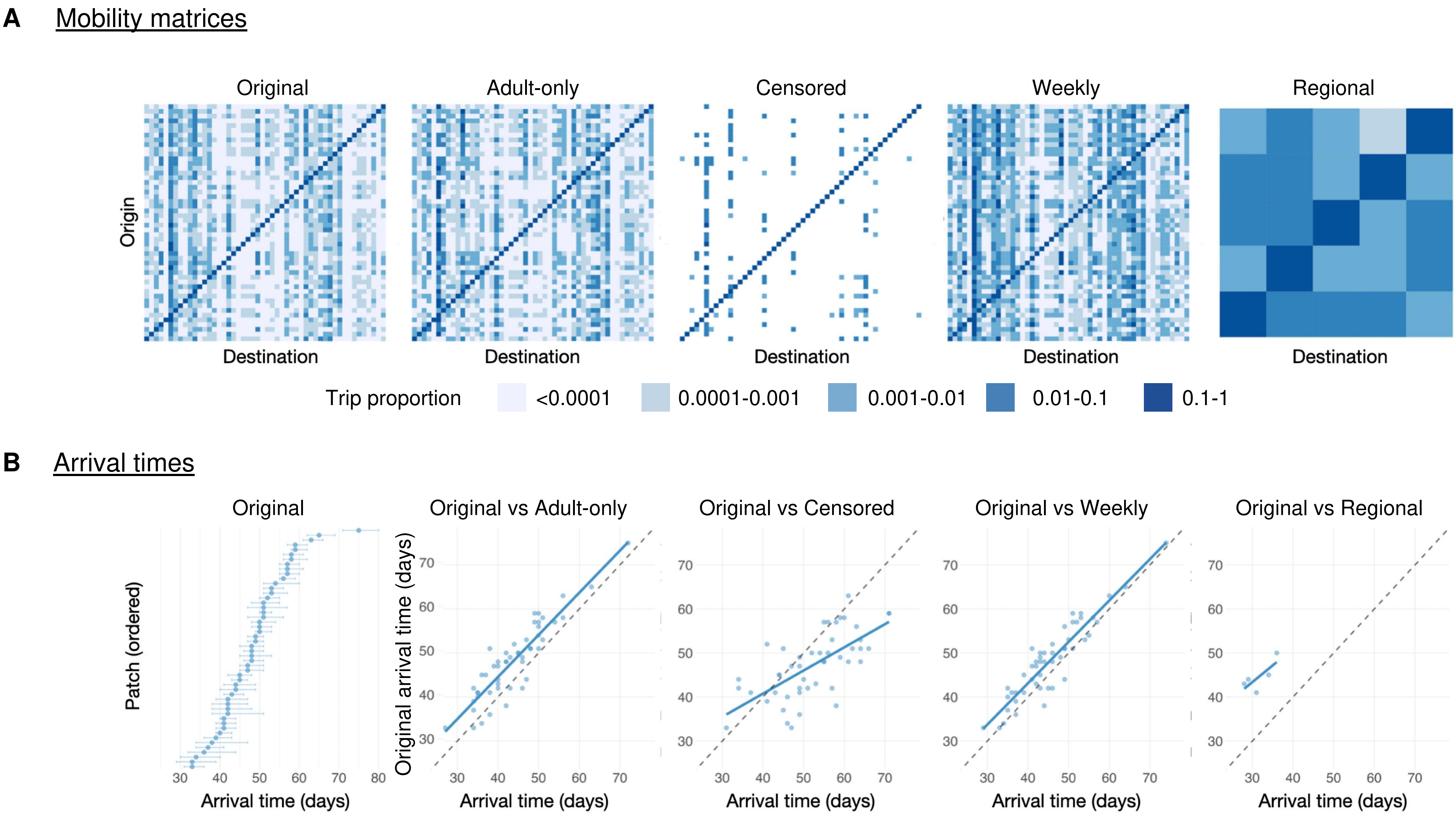
How can information on human mobility be used in public health response?					
Phase	**Main question(s)**	**Mobility data needs**	**How could the mobility data be used**	**Key unknowns**	**Example of use case**
1. Prior to an outbreak or in preparation	Given an outbreak, where will the pathogen spread?	Ideally access to multiple data sets given the high uncertainty in the population at risk or transmission patterns. Priority towards data describing international travel and from locations that are likely sources of initial spread. Data that could describe general movement patterns, since risk factors may not be identified, are preferable to understand how the mobility data could be integrated and used with disease surveillance network information	Mobility could be integrated into modeling frameworks to estimate introduction rates to other locations, given a range of initial starting locations.	What is the risk of importations to other locations?	Assessing risk of dengue by location to guide development of early warning and response systems[7]. Mobility data used: Twitter data on mobility between pairs of neighborhoods Method used: Regression analysis to assess predictors of dengue incidence. Number of cases modeled using Bayesian spatio-temporal modeling framework assuming a Poisson distribution.
2. Emergence	What is the order and timing of locations for the initial spatial	More specific temporal and spatial information for areas already	Using case data where the pathogen has emerged, estimate spatial spread to other	Who is infected? Which locations are infected?	Identifying potential secondary hubs for viral transmission of SARS-CoV-2[8].

	spread of a pathogen?	impacted and likely to spread; also requires information on the magnitude and frequency of travel between specific origins and destinations such as location data from mobile phones, social media, or GPS loggers.	locations using the frequency and number of trips to locations at risk.	How will different mobility behaviors drive transmission? How will superspreading or skewed transmission events impact the spatial spread	Mobility data used: Air passenger volumes between international airports / countries Method used: Reconstruction of dispersal patterns of variants of concern using phylogenic and phylogeographic methods. Use of regression to test association between travel volume, COVID-19 case and death counts, and international travel ban on estimated mean monthly exports of the virus.
3. Control	What is the amount of mixing or coupling between different populations to determine how interventions may or may not be impacted? How do different contact rates impact the effectiveness of interventions?	Validated mobility data against historic outbreaks, mapping of behaviors onto transmission inferred from disease data as well as detailed information about interventions that were deployed. Mixing and contact rates between populations, time spent at locations, origin-destination trip counts, duration, impact of	Inference against data to identify which data sources, mobility behaviors are most relevant for different diseases transmission	How do different mobility measures map onto behaviors that continue to drive transmission?	Demonstrating heterogeneity of SARS-CoV-2 transmission in rural and urban prefectures of Japan[9]. Mobility data used: Monthly flows between regions in Japan obtained from location data of mobile phone users Method used: Simulations using a spatial compartmental model to analyze effects of restricting mobility between regions in Japan

		interventions (aimed at mobility) on changes in mobility such as travel surveys, mobile phone and social media data, transportation.			on spatial spread of disease
4. Elimination	Where are locations less likely to be able to eliminate due to importation from endemic areas? Are there priority locations where additional monitoring and surveillance should be conducted?	Multiple types of data sets can be used that ideally would map specific movement patterns to infection status/risk of introduction events (e.g., travel surveys of cases); focus on local and international travel patterns.	Evaluation of various data sets against actual resurgence events and uncertainty in reestablishing transmission	Uncertainty of reestablishing local transmission	Identifying areas for targeted vector control in context of near-elimination of malaria[10]. Mobility data used: Mobile phone data (Haiti); census microdata from El Salvador, Costa Rica, Haiti, and Nicaragua Method used: After validating census data in estimating population movement, used census data to predict exportation and importation of malaria cases.

Figure 1. Illustration of the effect of incomplete data availability on dynamics of predicted epidemics. With different mobility matrices, the predicted dynamics of the modeled epidemics will also change; however, because the "Original" matrix is usually unobserved, the implications of sampling bias, censoring, and spatial and time granularity on predicted dynamics are also often not considered. Detailed information on the model and parameters used are available on Github.

A) Mobility matrices, namely origin-destination matrices of the amount of travel between locations, derived from a gravity model for a simulated population, comprised of adults and children with different gravity model parameters. Original matrix is that derived from the gravity model with no censoring or spatial or temporal aggregation. Additional variants include adult-only mobility matrix, censored mobility matrix excluding routes with fewer than 900 trips, weekly mobility matrix, and simulated five cluster regional mobility matrix. B) Simulated arrival times using the various mobility matrices to quantify connectivity between patches in a metapopulation using original mobility matrix, as well as comparisons of arrival times when using original mobility matrix against adult-only, censored, weekly, and regional mobility matrices.


A
Mobility matrices
Original
Adult-only
Censored
Weekly
Regional
Origin
Destination
Destination
Destination
Destination
Destination
Trip proportion
<0.0001
0.0001-0.001
0.001-0.01
0.01-0.1
0.1-1
B
Arrival times
Original
Original vs Adult-only
Original vs Censored
Original vs Weekly
Original vs Regional
Patch (ordered)
Original arrival time (days)
30
40
50
60
70
80
Arrival time (days)
Arrival time (days)
Arrival time (days)
Arrival time (days)
Arrival time (days)