

COMPUTATIONAL EPIDEMIOLOGY

Mobility data as a proxy for epidemic measures

A new study uses longitudinal mobility data to identify how individuals behave at different stages of the COVID pandemic, elucidating benefits and challenges of using this type of data for decision-making by epidemiologists and policy-makers.

Nishant Kishore

In the initial aftermath of the SARS-CoV-2 pandemic, an avalanche of mobility data was made available by a variety of providers. These data ranged from aggregated measures at low spatial resolutions to highly granular transition and movement data^{1–4}. Given the paucity of human behavior data, these mobility datasets were used as proxies for epidemiologically relevant measures, such as the contact rate between individuals and the general flow of populations between locations of interest⁵. In some cases, these metrics were integrated into dashboards and aggregated mobility metrics were used by researchers, journalists and policy-makers to evaluate the regional successes or failures of non-pharmaceutical interventions^{6,7}. However, the link between mobility metrics and key epidemiological parameters of interest is not always straightforward, and the different stages of the pandemic are rarely taken into account. Writing in *Nature Computational Science*, Roman Levin and colleagues⁸ use longitudinal and spatially granular mobility data to identify clusters of census block groups that may be at higher risk for transmission and exportation of SARS-CoV-2 during varying stages of the early epidemic in the United States.

While mobility metrics can potentially behave as proxies for epidemiological measures, such as contact rate and effective reproductive number, the intensity and direction of this relationship can change by epidemic stage⁹. For example, early in the pandemic, before mask mandates were widely adopted, a measure of the proportion of individuals in a county who spent time outside their home was a useful proxy for potentially contagious contacts. However, the link between traveling outside home and having a transmissible contact weakened as masking and social distancing mandates were put into place. This time-varying relationship can be especially challenging for policy-makers who might use the decrease in proportion of individuals spending all their time at home as a

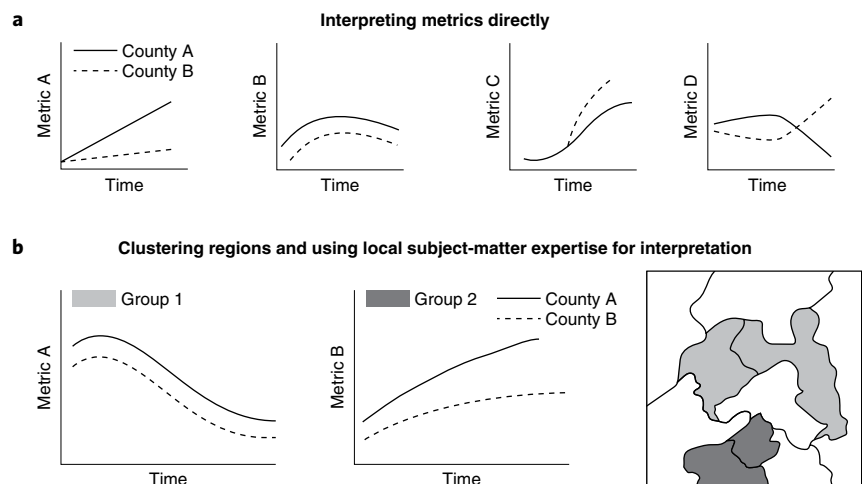


Fig. 1 | Operational difficulty in interpreting mobility metrics directly rather than using a robust clustering procedure. **a**, Evaluating metrics directly is challenging due to the difficulty in identifying causal links between metrics and public health parameters of interest. Additionally, comparing metrics directly is complicated by varying data coverage and representativeness. **b**, Using a robust clustering procedure, policy-makers are able to group spatial areas that behave similarly, together, and subsequently rely on local subject-matter expertise to interpret these metrics.

warning for a future increase in the effective reproductive number.


To overcome these challenges, Levin and colleagues define clusters of regions where individuals behave similarly using the entire time-series of data available rather than cross-sectional periods of interest (Fig. 1). In addition, they make use of a manifold learning technique to reduce the dimensionality of their longitudinal mobility data. As shown by the authors, these methodological advances allowed them to identify populations that moved early in the pandemic, which would have been missed through standard approaches. While previous approaches have identified an exodus of individuals from urban centers during the initial implementation of non-pharmaceutical interventions (NPIs)¹⁰, the approach presented by Levin and colleagues is able to differentiate between sub-regions (such as regions primarily occupied by college students) that moved

at different periods of the epidemic, likely experiencing different risk of transmission and exportation of SARS-CoV-2. This is especially useful for future research where sub-county-level populations with varying vaccination coverages may react differently and at different times to potential future NPIs, without needing to interpret mobility metrics directly.

While these mobility data have a lot of potential (as shown by Levin and colleagues), it is important to remember that they were not primarily created for use in public health or epidemiological research. As their availability and use increase, several key issues must be addressed to ensure the validity and integrity of these data. First, the data-generation process should be better elucidated, as there is often little information on the entire data-generation pipeline through which human behavior is translated into the metrics provided to

researchers and policy-makers. Without a keen understanding of what exactly is being measured, and from whom the data is being collected, the process of scientific inquiry can result in biased or potentially harmful policy. Second, researchers should define the purpose of these metrics, as not all of the metrics are useful proxies of transmission related to human behavior. Collaboration between researchers to compare the utility of these metrics and to define a standardization of epidemiologically relevant spatial and temporal scales is key for formalizing the use of mobility metrics in public health research. Third, the population should be well-defined. Data quality can change significantly when moving along the urban/rural or wealthy/poor gradient. Limited access to raw data and the data generation pipeline are imperative to robustly evaluate the validity and representativeness of these metrics. Gold standards are needed to appropriately compare data quality between providers, and standard measures of representativeness are needed to understand populations covered by different sources. Finally, privacy should be baked into the pipeline. As the public has become more aware of the data being collected by their devices, providers have incorporated differential privacy into their pipeline to ensure that individuals or vulnerable sub-groups are not identifiable. Further research is needed to understand how varying methods and noise introduced by differential privacy may (or may not) bias public health parameters of interest.

As with any other method, the usability of this framework is dependent on the validity of the data. There is often a long and opaque pipeline that connects human movement to the metrics that are provided for use by researchers and policy-makers¹¹. Each actor in this pipeline (publishers, providers and aggregators) is able to buy, clean and transform data without providing enough transparency into their processes, since the pipeline is their proprietary added value and the primary endpoint is often performance for commercial purposes. The types of activities that are captured can vary by application (for instance, the behavior captured while using a food delivery app is likely to be different than the behavior captured while using a dating app), and since these data are often bought and sold numerous times, it can be nearly impossible to understand overlapping samples of a population, methods used by various actors or even the heterogeneous use of smartphones and applications by populations of interest. Without information on coverage, representativeness and the data generation process, it is difficult to determine the correct course of action if one provider states that a metric increases in a location over a certain period of time while another states that the same metric decreases. Levin and colleagues are able to overcome some of these challenges by not interpreting these metrics directly but rather using manifold learning to identify time-varying regional clusters. Importantly, this allows policy-makers to use local subject-matter expertise to identify salient

characteristics between regions and design appropriate interventions rather than being forced to interpret the metrics directly. 

Nishant Kishore  

Center for Communicable Disease Dynamics,
Department of Epidemiology, Harvard T.H. Chan
School of Public Health, Boston, MA, USA.

 e-mail: nkishore@g.harvard.edu

Published online: 22 September 2021
<https://doi.org/10.1038/s43588-021-00127-7>

References

1. Fitzpatrick, J. & DeSalvo, K. Helping public health officials combat COVID-19. *Google* (3 April 2020); <https://blog.google/technology/health/covid-19-community-mobility-reports/>
2. Jackman, M. Using data to help communities recover and rebuild. *Facebook* (7 June 2017); <https://newsroom.fb.com/news/2017/06/using-data-to-help-communities-recover-and-rebuild/>
3. Snaith, B. & Thereaux, O. Mobility data sharing during the Covid-19 pandemic – Research from Cuebiq and The GovLab. *Open Data Institute* (24 March 2021); <https://theodi.org/article/mobility-data-sharing-during-the-covid-19-pandemic-research-from-cuebiq-and-govlab/>
4. *US Consumer Activity During COVID-19 Pandemic* (SafeGraph, 2021); <https://www.safegraph.com/data-examples/covid19-commerce-patterns>
5. Badr, H. S. et al. *Lancet Infect. Dis.* **20**, 1247–1254 (2020).
6. Chan, J. et al. CMDN user feedback case studies. *CrisisReady* (5 May 2021); <https://www.crisisready.io/publications/cmdn-user-feedback-case-studies/>
7. Fuller, A. & Hobbs, T. D. Rural Americans stopped staying in. Then Covid-19 hit. *The Wall Street Journal* (24 November 2020); <https://www.wsj.com/articles/rural-americans-stopped-staying-in-then-covid-19-hit-11606244401>
8. Levin, R., Chao, D. L., Wenger, E. A. & Proctor, J. L. *Nat. Comput. Sci.* <https://doi.org/10.1038/s43588-021-00125-9> (2021).
9. Kishore, N. et al. Preprint at *medRxiv* <https://doi.org/10.1101/2021.04.15.21255562> (2021).
10. Kishore, N. et al. *Sci. Rep.* **11**, 6995 (2021).
11. Kishore, N. et al. *Lancet Digit. Health* **2**, e622–e628 (2020).

Competing interests

The author declares no competing interests.