

TRANSPORT FINDINGS

Uncertainty in Cell-Phone Generated Bike and Pedestrian Volumes

Lily Heidger¹, Trisalyn Nelson, Ph.D.¹, Dan Willett¹¹ Geography, University of California, Santa Barbara

Keywords: big data, validation, active transportation, GPS data, cell-phone data

<https://doi.org/10.32866/001c.145234>

Findings

Big data from mobile phones are increasingly used in transport research and planning, offering unprecedented spatial and temporal detail. However, data accuracy remains unclear. This study evaluates Replica, a dataset modelled from mobile phone GPS data, by comparing modeled volumes for motor vehicles, bicycles, and pedestrians against field counts in Santa Barbara, California. Car volumes were modeled with high accuracy ($R^2 = 0.92$), while bicycle ($R^2 = 0.23$) and pedestrian ($R^2 = 0.05$) estimates showed substantial uncertainty. When using transport data generated from mobile phone GPS, additional caution is needed for non-motorized modes.

1. Questions

Big mobility data, often generated from cell phones, are becoming increasingly more popular in transportation research. Cell phone-generated mobility data provide unprecedented spatial and temporal resolutions and extent at a fraction of the cost of traditional data collection. These datasets are typically generated as synthetic mobility records derived from multiple sources, trained on mobile location data, and validated against ground-truth observations for multi-state regions (Replica 2025). However, further data validation is necessary in advance of using data for applied science or for local decision-making. Our study asks the question: how well do modeled mobile phone tracking data, specifically Replica data, represent actual bike, pedestrian, and car volumes in Santa Barbara, California?

2. Methods

According to Replica, their transportation data are modeled from a variety of original data sources, such as mobile location data, demographic data, built environment data, economic activity data, and ground truth data from local municipalities. The scale of these models is “megaregions,” which encompass multiple states and 10 to 50 million residents, depending on the region (Massey, n.d.). Transportation data are modeled as individual trips, with each trip specifying the street network links traversed by a synthetic user. Trips are generated to represent an average weekday or weekend within a given season, such as a typical Thursday in the spring of 2021. Replica data also do not account for recreational trips, where the start and end points of the trip suggest biking or walking for exercise (Replica 2025).

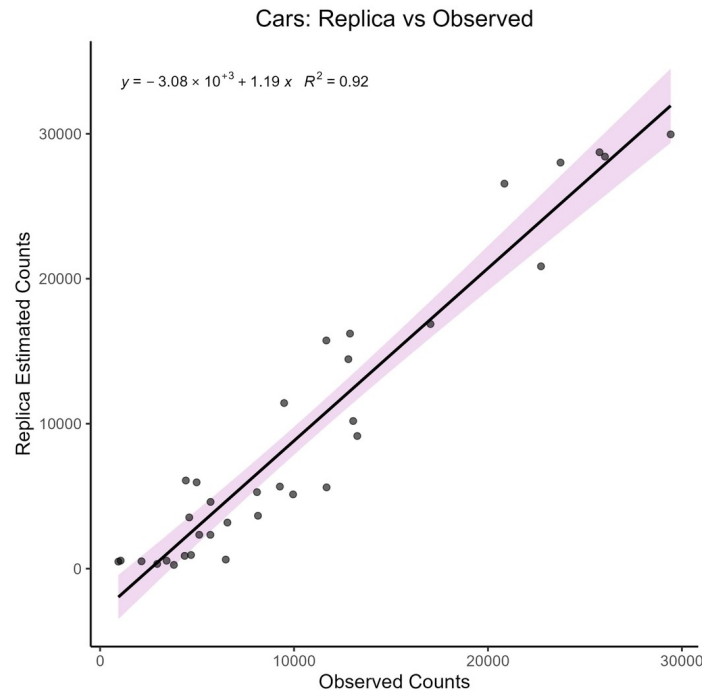


Figure 1. Linear regression of Replica car volumes and observed video count volumes.

We accessed modeled transportation volumes from [ReplicaHQ.com](https://replica-hq.com) and evaluated them against validation data provided by the City of Santa Barbara. Video counter volume data, collected at multiple intersections and analyzed manually, were cleaned and matched to Replica’s network links. For each time frame, Replica trips passing through intersections with video count data were totaled, and observations were aligned by day of the week, season, and year for comparability. In total, 33 sites across the city were matched, representing a range of car, bicycle, and pedestrian volumes.

To compare Replica counts to actual counts, we ran a linear regression for each of the modes of transportation. We estimated the R^2 value for each analysis and compared the results among transportation types.

3. Findings

Modeled car volumes were accurate, with an R^2 value of 0.94. [Figure 1](#) shows the results of the linear regression analysis. As shown in the plot, a majority of the car count sites were underestimated by Replica’s model, particularly those with lower volumes. Most of the sites with the highest volumes were more likely to be overestimated by the model.

Bike volumes had a much lower correlation coefficient than vehicles, with an R^2 of 0.23. [Figure 2](#) depicts a much larger confidence interval, illustrating a higher chance of error. For bikes, Replica’s model over- and underestimates volumes at similar rates, with slightly more than half being underestimated by the model. The variability appears random, where the model over and underpredicts randomly, not based on the site volumes. The two sites with

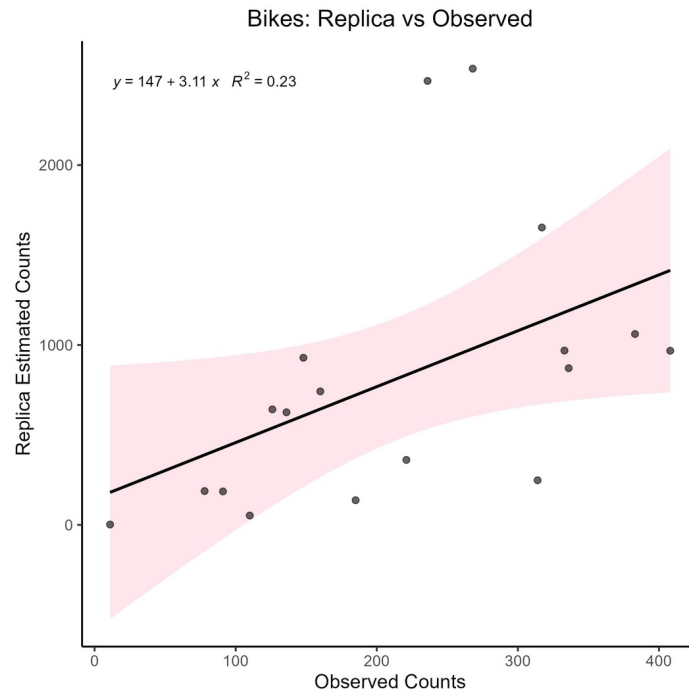


Figure 2. Linear regression of Replica bike volumes and observed video count bike volumes.

the largest residuals are overestimated by nearly an order of magnitude. While we do not show the results here, we spent significant time trying to identify patterns that could explain the differences in prediction, but the error always appeared random.

Among the three transportation modes, pedestrian volumes were modeled with the least accuracy, yielding an R^2 of 0.05 and wide confidence intervals ([Figure 3](#)). As with bicycle volumes, approximately half of the sites were overestimated and half underestimated, with no clear trend in model performance across volume levels. However, the magnitude of error was substantially greater, with the most severely overestimated site exhibiting observed volumes nearly eight times lower than the modeled estimate.

These findings highlight the critical need for improved data on bicycling and walking. Modeled car volumes likely achieve higher accuracy as a greater proportion of the cell phone data are generated through vehicle travel (Massey, n.d.). Bicycling and walking are more difficult to isolate based on space-time patterns (Lee and Sener 2020). Additionally, Replica does not model multi-modal transit trips, meaning the trips where individuals take the bus and walk to reach their destination are assigned one mode over the other (Replica 2025). This may lead to an underrepresentation of biking and walking trips near transit stops. Investment in bicycle and pedestrian count programs could improve training data and facilitate more accurate model development. In the meantime, alternative data sources will be necessary for applications such as exposure estimation. It is likely that this issue is not specific to Replica and includes other mobile location datasets, as the task of identifying different modes is challenging (Lee and Sener 2017, 2020). While

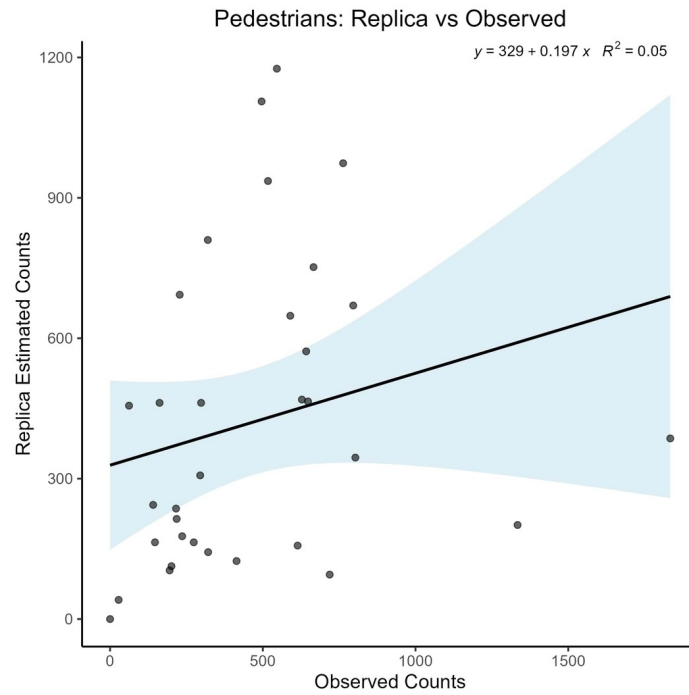


Figure 3. Linear regression of Replica pedestrian volumes and observed video count pedestrian volumes.

mobile location data with high spatial and temporal resolution are appealing, reliance on them without sufficient validation may be problematic and risk misleading decision-making, particularly in small and mid-sized cities such as Santa Barbara.

Submitted: September 10, 2025 AEDT. Accepted: October 07, 2025 AEDT. Published: October 17, 2025 AEDT.



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-SA-4.0). View this license's legal deed at <https://creativecommons.org/licenses/by-sa/4.0> and legal code at <https://creativecommons.org/licenses/by-sa/4.0/legalcode> for more information.

REFERENCES

- Lee, K., and I. N. Sener. 2017. "Emerging Data Mining for Pedestrian and Bicyclist Monitoring: A Literature Review Report." Safety through Disruption (Safe-D) National University Transportation Center (UTC) Program.
- . 2020. "Emerging Data for Pedestrian and Bicycle Monitoring: Sources and Applications." *Transportation Research Interdisciplinary Perspectives* 4:100095. <https://doi.org/10.1016/j.trip.2020.100095>.
- Massey, Lauren. n.d. "Data Quality Overview and Technical Approach." Replica Help. Accessed August 18, 2025. <http://help.replicahq.com/en/articles/6214243-data-quality-overview-and-technical-approach>.
- Replica. 2025. "Active Transportation Trips." <https://documentation.replicahq.com/docs/active-transportation>.