Findings

Mucci, Richard Alexander, and Gregory D. Erhardt. 2022. "Ride-Hailing Data Suppression and Exclusion Strategies Can Lead to Biased Outcomes." *Findings*, May.

TRANSPORT FINDINGS

Ride-Hailing Data Suppression and Exclusion Strategies Can Lead to Biased Outcomes

Richard Alexander Mucci[1] [a], Gregory D. Erhardt[1]

[1] Civil Engineering, University of Kentucky

Keywords: TNC, suppression, suppressed data, Chicago, Ride-hailing

https://doi.org/10.32866/001c.34191

Findings

The Chicago ride-hailing data set is one of the few data sets in the United States containing details of individual ride-hail trips. To protect privacy, locations and times are aggregated, and locations are further suppressed when the frequency of trips is low. Most researchers using this data remove the trips with suppressed locations or external destinations from their analysis. This research finds that when suppressed and external trips are excluded, the trip length, cost, and distance are all underestimated, as are trips in low-income neighborhoods. Future research should consider including these trips at a more aggregate spatial resolution.

1. Questions

Research on ride-hailing is limited because data is not widely available (Calderón and Miller 2021; Dean and Kockelman 2021; Erhardt et al. 2021; Henao and Marshall 2018). The City of Chicago publishes one of the few data sets to include the location, timing, and cost of ride-hailing trips, as well as an indication of whether those trips are pooled or private ("Chicago Data Portal" 2020).

The data include one record for each ride-hailing trip in Chicago as reported by the providers. To prevent individuals from being identified, reported origins and destinations are aggregated to census tracts, trip times are aggregated to 15-minute windows, and trip costs are reported to the nearest $2.50. In addition, when a census tract pair has two or fewer trips within a 15-minute window, the census tract for both trip ends is suppressed and reported for larger community areas. Figure 1 shows the 866 census tracts in Chicago, the 77 community areas in Chicago, and the 2,257 census tracts in Cook County but outside of Chicago. Trip ends outside the city boundary do not include the Community Area but do include the census tract if it is within Cook County.

Previous research using the Chicago Transportation Network Provider (TNP) data excluded trips with either end outside Chicago (13.72% of trips) and internal trips with the census tract suppressed (20.21%) (Dean and Kockelman 2021; Ghaffar, Mitra, and Hyland 2020; Soria, Chen, and Stathopoulos 2020; Yan, Liu, and Zhao 2020). We aim to answer the question: how do the characteristics of these external and suppressed trips in the Chicago TNP data

a email: alex.mucci@uky.edu

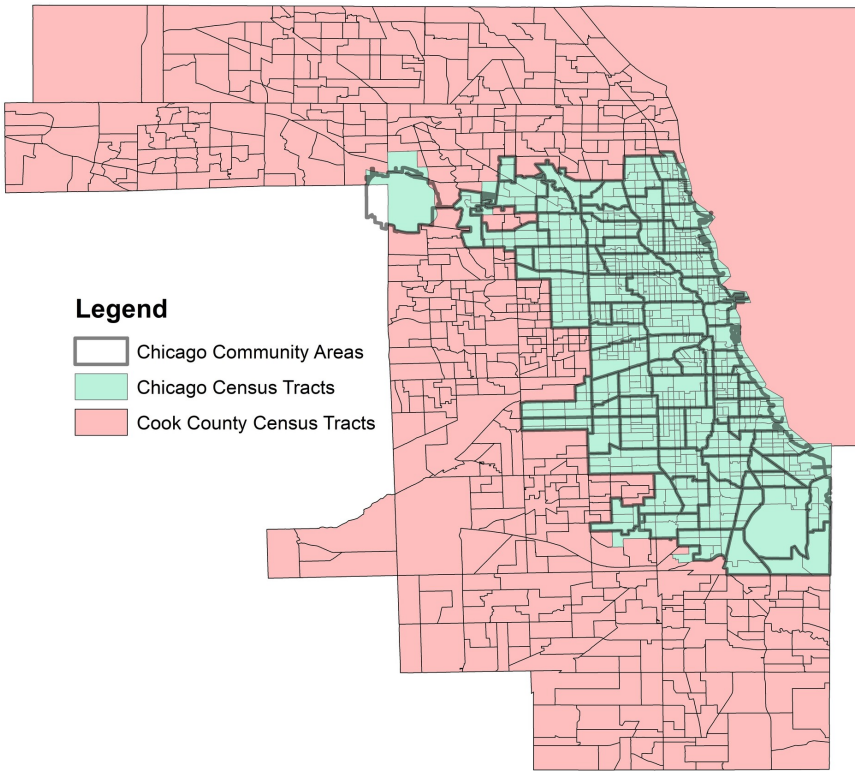

Figure 1. Census tracts and community areas in Chicago and Cook County

set differ from the internal unsuppressed trips? The answer will help other researchers assess whether those records can safely be excluded from their analyses and highlight the issue for users of other data sets that may use similar privacy protection strategies.

2. Methods

To evaluate this question, we analyzed weekday trips in the Chicago TNP data from November 2018 through February 2020. We excluded the 1.2% of internal trips lasting more than 80 minutes, costing more than $50, or longer than 40 miles because these values exceed the longest trip that could reasonably be made within the city limits. We classify the remaining trips into three categories:

- Internal unsuppressed (66.07%): The trip record includes both the pickup and drop-off census tract, so we know that both trip ends are within Chicago and at least two other trips occur between the same census tracts in the same 15-minute period.
- Internal suppressed (20.21%): The trip record includes both the pickup and drop off community area, but not census tract. We know that both trip ends are within Chicago and fewer than two other trips occur between the same census tracts in the same 15-minute period.

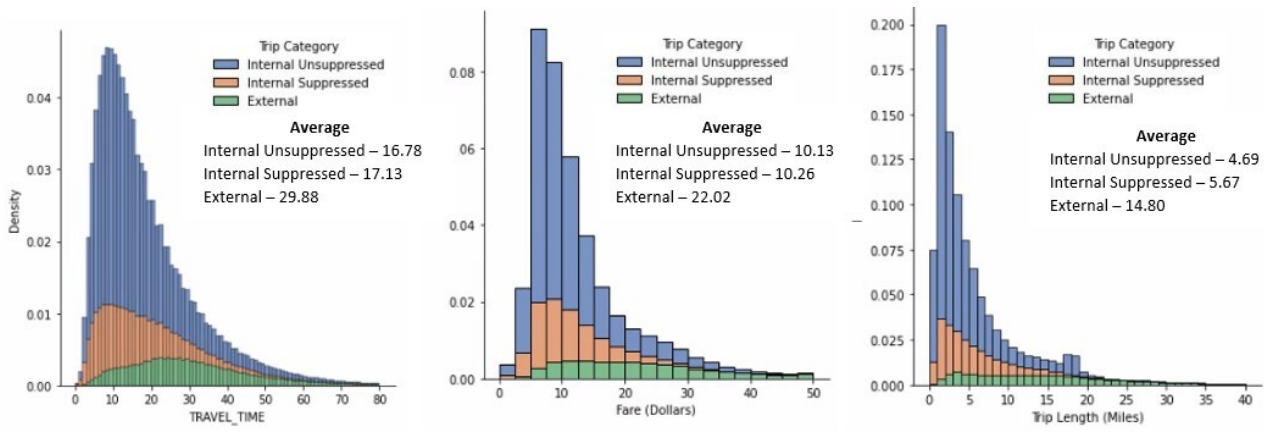

Figure 2: Trip Time, Fare and Distance

Figure 2. Trip Time, Fare and Distance

- External (13.72%): At least one trip end is missing a community area, indicating that it is outside the Chicago city limits. This group includes three subcategories:
    - Internal-external unsuppressed (8.76%): The internal end records the census tract.
    - Internal-external suppressed (4.90%): The internal end records the community area but not the census tract.
    - External-external (0.06%): Both ends are outside Chicago.

In the findings, we summarize key attributes for these three categories of trips.

3. Findings

Figure 2 shows the distribution of travel time, fare, and trip length, segmented by category. A two sample t-test found that the differences between internal suppressed and internal unsuppressed trip attributes are statistically significant. The differences between external and internal unsuppressed trip attributes are statistically significant as well.

Figure 3 shows the share of trips occurring in each hour for the three categories of trips. External and internal suppressed trips stay consistent throughout a day without AM and PM peaks, unlike internal unsuppressed trips. There is a higher share of trips suppressed overnight.

Figure 4 shows the distribution of ride-hailing trips by community area, including: a) the unsuppressed internal trip pickups throughout the study period, c) the share of internal trip pickups that are suppressed, and d) the share of pickups that are an internal-external trip. Figure 4 also shows b) the median household income as reported in the 2018-2020 American Community Survey (ACS). We average three years of ACS data because it spans the same period that the ride-hailing data covers. We observe that the internal unsuppressed

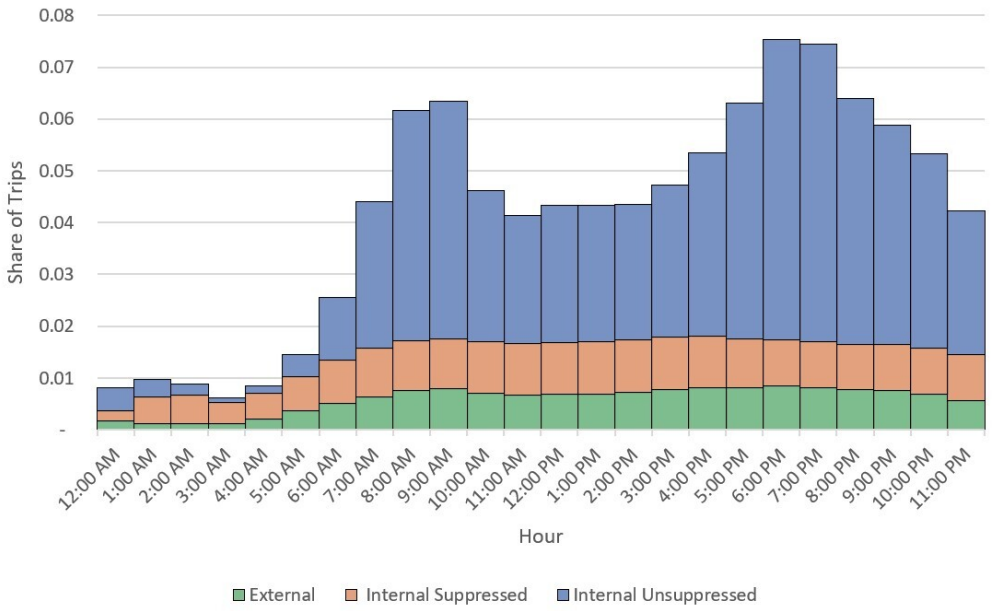

Figure 3: Hourly Breakdown of Suppression

Figure 3. Hourly Breakdown of Suppression

trips are most concentrated in the downtown area and on the north side, which are also the wealthiest parts of Chicago. In contrast, the internal suppressed trips consist of more than half of all trips in the south and west parts of Chicago, which tend to be lower income. External trips consist of more than half of all trips at locations along the edge of the city and at O'Hare airport.

Broadly, these results show that trips are more likely to be suppressed when they occur in locations and times of day with infrequent ride-hailing use. Figure 5 further illustrates this effect by comparing the frequency of pickups within a community area to the share of pickups that are suppressed. This outcome is an intended result of the data suppression strategy.

We agree that it is necessary to protect user privacy. However, dropping these trips will exaggerate the differences between frequent and infrequent trips, potentially underestimating trip lengths, underestimating overnight trips, and underestimating trips in lower density areas. The effect of such bias might be modest, but it could be important if it is correlated with variables of interest, or if it serves to exclude low-income or minority populations. In the future, analysts should consider whether suppressed and external trips can be included in their analysis, and data managers should consider data suppression strategies that could mitigate these biases.

Submitted: February 24, 2022 AEST, Accepted: April 05, 2022 AEST

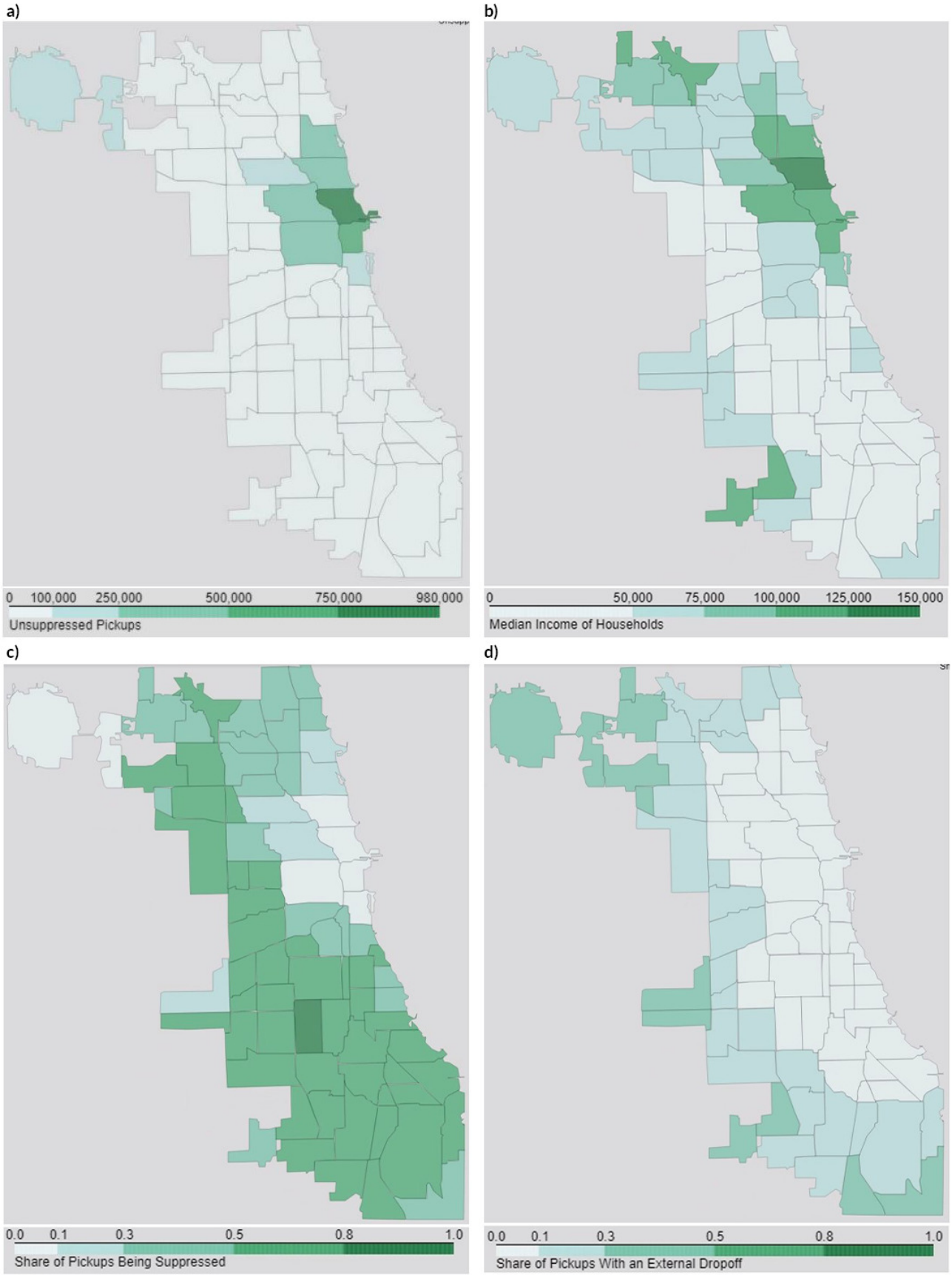

Figure 4. Visualizing Suppression, External Trips, and Household Income

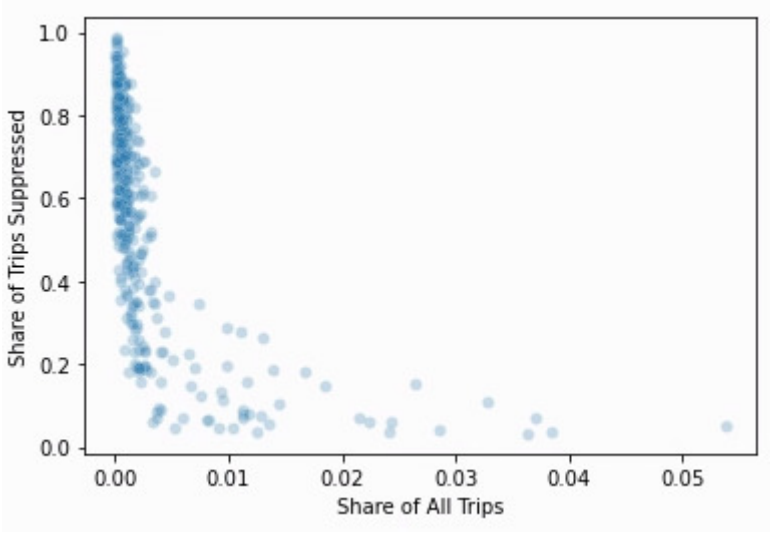

Figure 5. Frequency of Trips vs. Suppression Scatter Plot

REFERENCES

Calderón, Francisco, and Eric J. Miller. 2021. "Modelling Within-Day Ridehailing Service Provision with Limited Data." *Transportmetrica B: Transport Dynamics* 9 (1): 62–85. https://doi.org/10.1080/21680566.2020.1784809.

"Chicago Data Portal." 2020. November 1, 2020. https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p.

Dean, Matthew D., and Kara M. Kockelman. 2021. "Spatial Variation in Shared Ride-Hail Trip Demand and Factors Contributing to Sharing: Lessons from Chicago." *Journal of Transport Geography* 91 (February): 102944. https://doi.org/10.1016/j.jtrangeo.2020.102944.

Erhardt, Gregory D., Richard Alexander Mucci, Drew Cooper, Bhargava Sana, Mei Chen, and Joe Castiglione. 2021. "Do Transportation Network Companies Increase or Decrease Transit Ridership? Empirical Evidence from San Francisco." *Transportation*, February. https://doi.org/10.1007/s11116-021-10178-4.

Ghaffar, Arash, Suman Mitra, and Michael Hyland. 2020. "Modeling Ridesourcing Trip Generation: Chicago Case Study." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3571040.

Henao, Alejandro, and Wesley E. Marshall. 2018. "The Impact of Ride-Hailing on Vehicle Miles Traveled." *Transportation*, September. https://doi.org/10.1007/s11116-018-9923-2.

Soria, Jason, Ying Chen, and Amanda Stathopoulos. 2020. "K-Prototype Segmentation Analysis on Large-Scale Ridesourcing Trip Data."

Yan, Xiang, Xinyu Liu, and Xilei Zhao. 2020. "Using Machine Learning for Direct Demand Modeling of Ridesourcing Services in Chicago." *Journal of Transport Geography* 83 (February): 102661. https://doi.org/10.1016/j.jtrangeo.2020.102661.