Estimating traffic flows from vehicle trajectories based on sparse mobile phone geolocation data

(article starts on next page)

Estimating traffic flows from vehicle trajectories based on sparse mobile phone geolocation data

ROOS TEEUWEN[1,*] AND JORGE GIL[1]

[1] *Department of Architecture and Civil Engineering, Chalmers University of Technology, Gothenburg, Sweden*
[*] *roos.teeuwen@chalmers.se*

1. INTRODUCTION

Empirical traffic flow data are key in transport, mobility, and spatial planning. They can help understand congestion, traffic safety, environmental exposure and risks, among others. At scale, they can support data-driven decision making, helping to decide where interventions are most needed.

However, existing traffic flow data from sensors or traffic counts [1] lack spatio-temporal coverage and granularity. Other data, e.g. from navigation API's, are proprietary, commercial or limited-access, and unavailable to decision-makers. Large mobile phone traces data recently emerged as a promising source to capture dynamics at scale given their size, granularity, and coverage. They have been used to analyse travel demand (origin-destination), activity-locations, and individuals' activity spaces. Yet, despite their potential for exploring trajectories and traffic flows [1], dynamic applications other than understanding pedestrian routing behaviour [2] remain unexplored.

This study aims to explore how traffic flows with high spatial and temporal coverage and granularity can be estimated from vehicle trajectories based on sparse mobile phone geolocation data. We develop a methodology to create trajectories and flows from raw location data and test how various parameters affect the results. We contribute our methodology, code and data to allow for replication in other studies, and reflect on directions for future development.

2. DATA AND TOOLS

Raw mobile phone geolocation data from Sweden for 2024, containing high-resolution coordinates from GNSS (e.g. GPS) and other sources, timestamps and device ids, was acquired from the aggregator company *Pickwell*. We extract points in Stockholm (1292 million data points; from 3.2 million devices) and Gothenburg (670 million; 1.4 million), Sweden, including a 1 km buffer zone bounding box. We also collect three open ground truth data sets of average daily vehicle flows: (1) weekday flows in Stockholm, 2024, at 366 local and few major roads in Stockholm (obtained from *Stockholm Municipality*); (2) flows in Gothenburg, 2023, at 176 local and few major roads (*Gothenburg Municipality*); and (3) flows in and around Gothenburg, 2023, at 245 highway locations (Swedish Transport Administration *Trafikverket*). Lastly, we collect open data of roads accessible to vehicles and their speed limits (*Trafikverket*).

We process and analyse our data using open-source tools: *PostgreSQL* with *PostGIS* and *Python*. Data is collected and processed according to the European Union's General Data Protection Regulation and the data management plan has been approved by the Data Protection Office of the authors' institution. Our code and data are openly available on GitHub and Zenodo.

3. METHODOLOGY

A. From points to trajectories

We order all data points per device based on timestamp, and calculate the distance (d [m]), time difference (t [s]), and speed (s [km/h]) to the previous (i.e. incoming) and next data point (i.e. outgoing). To form trajectories, we exclude stationary points ($s \leq 3$ or $NULL$ and $d \leq 100$ for both incoming and outgoing points), isolated points ($s \leq 3$ and $d > 100$), singular points ($s = NULL$), and errors ($s > 250$). We then define trajectories as any set of points from the same device involving movement between stays (i.e. stationary), beginning with a *start* ($s_{incoming} \leq 3$ or $NULL$, and $d_{incoming} \leq 100$, and $s_{outgoing} > 3$ or $s_{outgoing} \leq 3$ but $d_{outgoing} > 100$), and assign those a common trajectory id.

These trajectories still require cleaning. We combine trajectories of the same device with a short stay in between ($t < 120$ and $d < 100$) (e.g. at a traffic light). We split trajectories in case of sudden jumps, i.e. when the next point is far away ($d > 1000$), the speed is too fast ($s > 250$), or too much time has passed ($t > 120$). In case a trajectory jumps far away and then back ($d \leq 100$) within reasonable time ($t < 120$) we only remove the jump. Lastly, we drop subsequent points with identical coordinates. We then create our final trajectories by linking points with the same unique trajectory id into linestrings, and calculate characteristics of the trajectory including the number of points, average distance between points and average speed.

B. Estimating flows per road segment

To estimate relative vehicle flows per road segment, we first match our trajectories to the road network using the *LCSS* map-matching algorithm [3]. The LCSS algorithm is a computationally efficient algorithm that iteratively matches (segments of) trajectories until an acceptable similarity fit has been achieved. We use the LCSS algorithm implementation from the *mappymatch* Python library, setting the similarity score parameter to 0.95 [3] and the other parameters to their default values. A limitation for our use case may be that the LCSS algorithm was developed for high-resolution data, while we use sparse data. However, since the algorithm recursively reduces the resolution of trajectories for computational efficiency as long as it yields an acceptable fit, in practice it handles trajectories of varying resolutions.

We match our trajectories with the collected road network, that we preprocess to resemble the data structure of *osmnx* networks [4] to allow integration with the LCSS algorithm as implemented in the mappymatch Python library. First, we build a data set of unique *nodes* as start or end points of the original edges, with their coordinates, and give each node a unique id as an artificial *osmid* attribute. Second, we build a data set of *edges*, looking up the osmid of their start and end points of every road segment, that we store as u and v attributes, respectively, and a unique *key* in case multiple edges connect the same pair of nodes. To resemble a *simplified* osmnx network, we drop the geometry of straight edges and maintain only the geometry of curved edges (i.e. having additional coordinates not captured by its nodes). We also add the reverse of every edge, since the original road network was non-directed.

After map-matching, we count the number of unique trajectories per road segment, and normalize these values to obtain the estimated relative vehicle flows within each city.

C. Evaluating suitability of trajectories

We compare our estimated flows to three ground truth data sets to evaluate the suitability of the generated trajectories. We join the midpoint of every ground truth road segment to the nearest (within 10 m) map-matched road segment. We then calculate the Spearman's ρ correlation between the observed traffic flow and the flow estimated from the mobile phone geolocation data, using a significance threshold of 0.01. We use varying sets of trajectories as input for the flows to evaluate how correlations are affected by the sample size and characteristics (i.e., number of points, average speed and resolution) of trajectories, to get insight into the most suitable trajectories for estimating flows. A smaller data sample has advantages in terms of map-matching and analysis efficiency, and reduced mobile data costs.

Robustness and sample size. To evaluate the robustness of the data, we take 10 random samples of 100,000 trajectories as input, test how successfully they are map-matched, and how resulting flows correlate with the ground truth data. To test the effect of sample size, we gradually increase the sample from 100,000 to 1,000,000 trajectories, concatenating the random samples and dropping duplicates, to see how correlations change.

Trajectory characteristics. To evaluate effects of trajectory characteristics, we use samples of 100,000 map-matched trajectories that fulfil various quality criteria: have a minimum number of points, gradually increasing from 2 to 10; a certain average speed, from 0 to 20 km/h, i.e., as a potential means to filter out non-motorized traffic; a maximum average distance between points, from 1000 to 100 m, i.e., as a measure of resolution. We then test how different thresholds correlate with the ground truth.

4. RESULTS

From our raw data, we generated 85 million trajectories in Stockholm and 46 million in Gothenburg. On average, trajectories span 0.9 km and last 1.9 min, with a resolution of 246 m between points. From these trajectories, we have estimated flows for 204,032 road segments in Stockholm and 127,290 in Gothenburg.

Robustness and sample size. All 10 random samples of 100,000 trajectories resulted in significant and comparably strong correlations (*rho* between 0.633 and 0.660, median 0.647, p-value < 0.01), and regardless of the sample, 73% of trajectories could successfully be map-matched to the road network, indicating stable outcomes when using different random samples of this size. Unmatched trajectories include trajectories at sea (e.g. ferries), at the borders of the study area, and to some extent at highways (e.g. due to the large spacing between nodes at these roads).

When increasing sample size from 100,000 to 1,000,000 input trajectories, we observed a negligible increase in correlation strength from 0.644 to 0.656 (p-value < 0.01). Simultaneously, the share of road segments that received an estimate increased from 54% to 81%, but this increase flattens with increasing sample size. Understanding these effects and how they hold for smaller and larger sample sizes warrants further research.

Trajectory characteristics. When selecting trajectories with a minimum amount of trajectory-points as input for estimating flows, correlation strengths varied between 0.478 and 0.636 (p-value < 0.01), but no pattern could be identified.

When setting a minimum threshold for trajectories' average speed (figure 1), comparison with the Stockholm data set showed decreasing correlations with increasing speed, from 0.600 when including trajectories of any speed to 0.390 when using a 20 km/h threshold (p-value < 0.01). For Gothenburg's local and major roads, we observed a slight increase from 0.574, to 0.636 at 7.5 km/h (p-value < 0.01), before decreasing again for higher speeds. For Gothenburg's highways, we observed a rise in correlation strengths from moderate, to strong when selecting trajectories over 10 km/h (0.629, p-value < 0.01), to very strong when selecting trajectories over 20 km/h (0.809, p-value < 0.01). This suggests that average trajectory speed is a promising indicator to improve estimates at highways, while it may not be a useful sole indicator to distinguish vehicle trajectories from other modes within the city, given speed variance between road types, and due to multi-modal trips, crossings, traffic lights, congestion, etc.

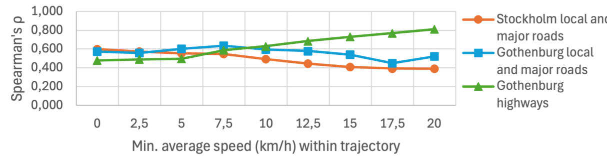

Fig. 1. Correlation strength differing with average speed.

When decreasing the allowed average distance between trajectory-points (i.e. increasing resolution), we found a decline in correlation strength with all three ground truth data sets for the most high-resolution trajectories. At high resolutions of 200 and 100 m on average, correlations are moderate for Stockholm, weak or insignificant for Gothenburg's local and major roads, and insignificant or even negative for Gothenburg's highways. This might indicate that high-resolution trajectories are not represented in every road type, requiring further investigation.

5. CONCLUSION

This paper introduces a methodology to estimate traffic flows from empirical vehicle trajectories based on sparse mobile phone geolocation data. We intend to open up new possibilities for data-driven transport, environmental and planning research and practice by providing extensive and frequent traffic flow estimates in entire metropolitan areas. Our approach yields strong correlations with ground truth data. For highway locations in Gothenburg, selecting trajectories of at least 20 km/h on average results in very strong correlations, while no subset of trajectories could be identified that outperforms a random sample of the data for the other ground truth data sets. Future work remains to understand where, when, and why deviations remain and how estimates within cities can be improved further, for instance by exploring other indicators than speed to identify vehicular trajectories, by ensuring a larger share of trajectories are map-matched, or by investigating how sparse mobile phone geolocation data can serve to estimate flows for different times, days and seasons.

Funding. This research is funded by Chalmers University of Technology Transport Area of Advance under grant agreement no. 2024-0299.

Disclosures. The authors declare no conflicts of interest.

Data availability. Estimated flows and ground truth data underlying the results presented in this paper are available on Zenodo. Geolocation points and trajectories cannot be made public due to privacy restrictions.

REFERENCES

1. J. C. Herrera, D. B. Work, R. Herring, *et al.*, "Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment," Transp. Res. Part C: Emerg. Technol. **18**, 568 (2010).
2. C. Bongiorno, Y. Zhou, M. Kryven, *et al.*, "Vector-based pedestrian navigation in cities," Nat. Comput. Sci. **1**, 678 (2021).
3. L. Zhu, J. R. Holden, and J. D. Gonder, "Trajectory Segmentation Map-Matching Approach for Large-Scale, High-Resolution GPS Data," Transp. Res. Rec. J. Transp. Res. Board **2645**, 67 (2017).
4. G. Boeing, "Modeling and Analyzing Urban Networks and Amenities With OSMnx," Geogr. Anal. p. gean.70009 (2025).