Linking Smartphone GPS Data with Transport Planning: A Framework of Data Aggregation and Anonymization for a Journey Planning App

Yang Wang*[1] and David Philip McArthur†[1]

[1] Urban Big Data Centre, University of Glasgow, Glasgow, G12 8RZ, United Kingdom

Summary

With the proliferation of GPS tracking data provided by smartphone apps, it is desirable to develop a data processing and anonymizing framework to transform raw GPS data into a suitable format for transport planning models. The paper aims to describe an effort to address such issues by map matching and aggregating the GPS information derived from a journey planning app. The effectiveness and flexibility of such a framework is demonstrated by an analysis of speeding and waiting time patterns in England and Wales by tracking 120 users for a year.

KEYWORDS: GPS, Transport Modelling, Map Matching, Anonymization, Aggregation

1. Introduction

Transport planning data are normally derived from surveys or travel diaries which are both time-consuming and expensive, often resulting in data being collected infrequently (Leiman et al., 2006). The proliferation of smartphone apps offers new opportunities to address the challenges by feeding automatically-generated personal travel logs and the associated insights to planners in order to better understand issues such as congestion, the quality of transport services and to facilitate evidence-based decision making (Fan et al., 2013, Nitsche et al., 2012).

The Catch! (Citizens At The City's Heart) project‡ provides a journey planning app which employs live travel feeds and shared travel experiences. Users of Catch! also aim to benefit transport planners by providing rich and granular data. Transformations of crowdsourced smartphone GPS signals to formats compatible with existing methods/models is a significant part of the project. This paper summarises the procedures that have been adopted to clean, aggregate and anonymize the data which is consequently organized into a minute-by-minute speed and junction monitoring dataset.

2. Data Aggregation and Anonymization Framework in Catch!

2.1. Data Cleansing and Journey Detection

Here, we describe the steps undertaken in Catch! towards a solution for spatio-temporally associating GPS data with road links and intersections (using OpenStreetMap (OSM) road network). The tracking records of 120 users gathered over the course of a year between August 2014 and August 2015 in England and Wales from the app helps to illustrate the methods and results.

---

* Yang.Wang@glasgow.ac.uk
† David.Mcarthur@glasgow.ac.uk

The GPS records contain Latitudes/Longitudes and timestamps of users' movements. The app also detects the travel mode based on a proprietary algorithm developed our project partner, TravelAi, using the smartphone's sensors. Tracking edges, Mondays-Thursdays (100,465 travel edges), Fridays (28,681 travel edges) and Weekends (42,743 travel edges), are generated through a 'smoothing' process where edges with unrealistic speeds and travel modes are ignored.

2.2. Map Matching Travel Journeys to Road Network

Map matching is a process of finding the road segment used by a user, given their GPS points (White et al., 2000). With challenges of low sampling frequencies, gross outliers and an inaccurate sequence of locations, it is known that good map matching results would facilitate accurate analysis of traffic flows, making recommendations for journeys, and detecting travel frequencies. We tested several geometric, topological and recent advanced map matching libraries and methods. Among them, GraphHopper and Barefoot, both in Java and providing online and offline matching, have shown significant advantages over the others we considered[§].

We collect a set of 'ground truth' data in a travel diary for a particular user to systematically compare the performance of GraphHopper and Barefoot. The issue of associating map matching results with road links is addressed by analysing a) the collections of intersections in PostGIS (denoted by method_1), b) the percentage of overlapping in a thin buffer in ArcGIS (denoted by method_2), and c) the resemblance of b) in PostGIS (denoted by method_3). We define straightforward accuracy (Equation 1), not matched (Equation 2) and redundancy map matching rates (Equation 3). Note that redundancy rates are not necessary less than 1 based on its definition.

Let $C(m)$ be the number of elements in map matching output road links, $C(g)$ be the number of elements in ground truth road links,

$$A = \frac{\cap\,(C(m), C(g))}{C(g)} \tag{1}$$

$$N = \frac{C(g) - \cap\,(C(m), C(g))}{C(g)} \tag{2}$$

$$R = \frac{C(m) - \cap\,(C(m), C(g))}{C(g)} \tag{3}$$

Figure 1 shows the fitted Kernel Density (KDE) trends for the probability distribution of the accuracy, not matched and redundancy map matching rates associated with the combinations of approaches. It is evident that there is an improvement on all the rates when adopting approach c) as illustrated with red trend lines. Barefoot provides slightly better map matching performances.

[§] We also tested two Python libraries 'Mapillary map matching' (https://github.com/mapillary/map_matching) and the code provided in (Westra, 2015). Neither show as good map matching results as the two Java libraries so are not included in this paper.
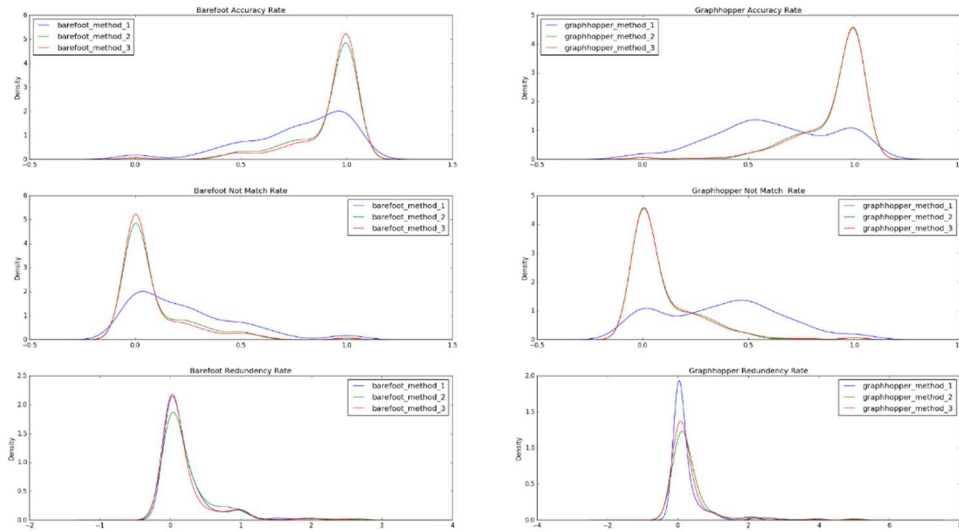
Figure 1. Measuring Results of Map Matching

2.4. Anonymization and Aggregation in Catch!

With highly sensitive spatio-temporal information about users and the requirement of local authorities for detailed travel patterns, Catch! demands privacy protection procedures which minimize the chances of an individual being identified while maximizing the usefulness of the data. We acknowledge the fact that such procedures should be compatible with transport modelling practices (Sila-Nowicka & Thakuriah, 2016). Publishing off-line trajectory data in the literature often involves clustering, generalisation and suppression based methods. We follow an approach by which k GPS trajectories are clustered to their associated road link. This approach offers the chance to share 'raw' GPS records but significant information loss was noticed during experimentation. We then generalize tracking information, such as average speed, orientation, wait time on certain road links and intersections, to relax k-anonymity by putting detailed individual tracks cloaked using average numbers.

3. Result

We aggregate mode of transport, direction, max, min and median travel speeds to each road link, if a) it is map matched to any journey b) there are GPS tracks within its 30 metre buffer[**], temporally on minute-by-minute scale. For every road intersection, we associate additional 'wait' information, if the speed is below 0.5km/h for 'pedestrian' or 5km/h for other modes. A map of the spatial distribution of average speed for travel by 'car' is shown by the upper map of Figure 2. The upper row of plots in Figure 2 shows the median speed distribution of all the roads used, the bottom row of plots illustrates an aggregation of speeding (journey speed is over 1.5 times of the max speed of the road link provided by OSM) on different categories of roads. At weekends, trunk, primary roads and motorways have higher chances of speeding, especially around 10:00-15:00 followed by the 15:00-20:00 time slot. The map in Figure 3 shows a zoomed-in view of the wait time on road intersections in London. Our framework can locate the roads with the highest wait time in a given period shown as the bottom plot of Figure 3.

4. Conclusion

This paper demonstrates the design and results of an aggregation and anonymization approach for the

---

[**] We tested 10-50m buffers around road links and intersections. Comparison of the median speeds and wait time shows a stable trend around 30m.

Catch! project. The framework provides not only the facility to extract useful speed, wait time and other information while protecting user privacy, but also the flexibility to adjust the analysis on both spatial and temporal scales. Although processing on a journey basis takes a longer computation time compare with other approaches, it provides accurate measurements. The initial application of this framework reveals useful patterns about roads and road intersections. The output can easily be linked to other information for more complicated analysis.

5. Acknowledgements

Thanks to all the partners of Catch! project.

6. Biography

Dr. Yang Wang is a research associate in Urban Big Data Centre, University of Glasgow. Her interests include mapping, processing and analysing crowdsourced data from location based social network and GPS in topics of urban and transportation planning.

Dr. David Philip McArthur is Lecturer in Transport Studies at the Urban Big Data Centre, University of Glasgow. He is a regional scientist specialising in modelling the relationships between transport infrastructure and labour markets.

References

Fan, Y., Chen, Q., Liao, C. F., & Douma, F. (2013, January). UbiActive: A smartphone-based tool for trip detection and travel-related physical activity assessment. In *Annual Meeting of the Transportation Research Board*.

Leiman, J. M., Bengelsdorf, T., & Faussett, K. (2006). Household Travel Surveys: Using Design Effects to Compare Alternative Sample Designs. In *Transportation Research Board 85th Annual Meeting* (No. 06-0965).

Nitsche, P., Widhalm, P., Breuss, S., & Maurer, P. (2012). A strategy on how to utilize smartphones for automatically reconstructing trips in travel surveys. *Procedia-Social and Behavioral Sciences*, *48*, 1033-1046.

Sila-Nowicka, K., & Thakuriah, P. (2016). The trade-off between privacy and geographic data resolution. a case of GPS trajectories combined with the social survey results. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 535-542.

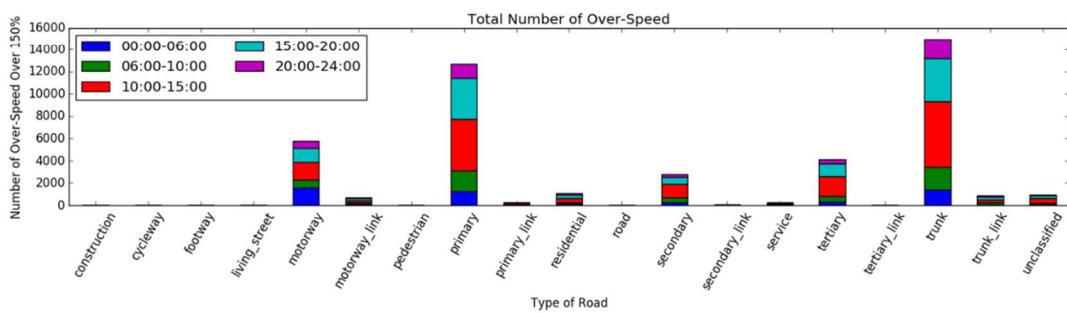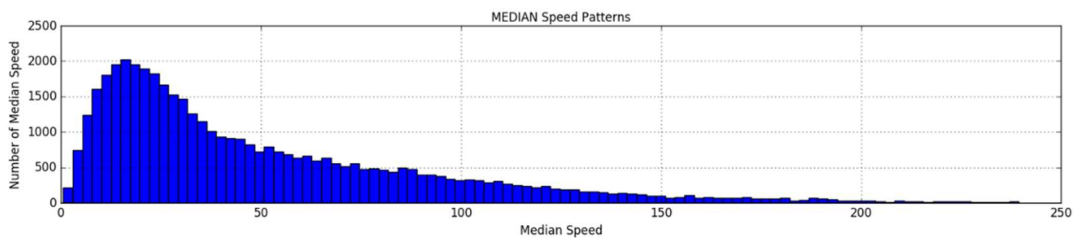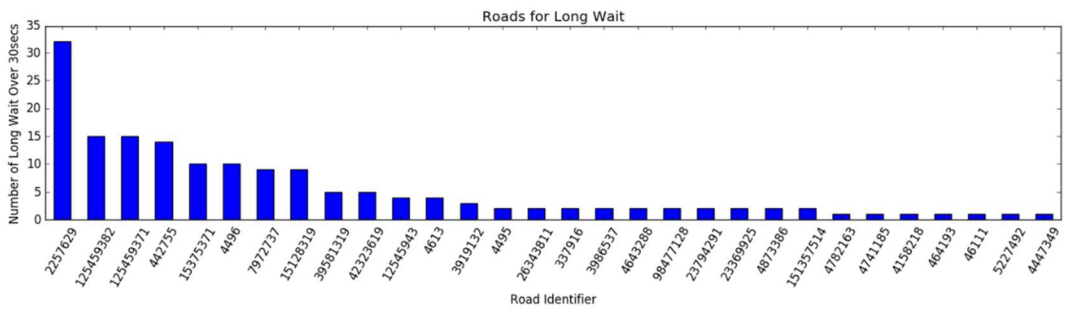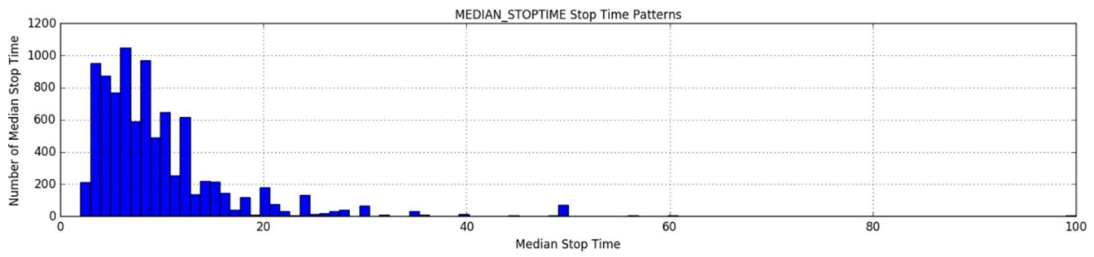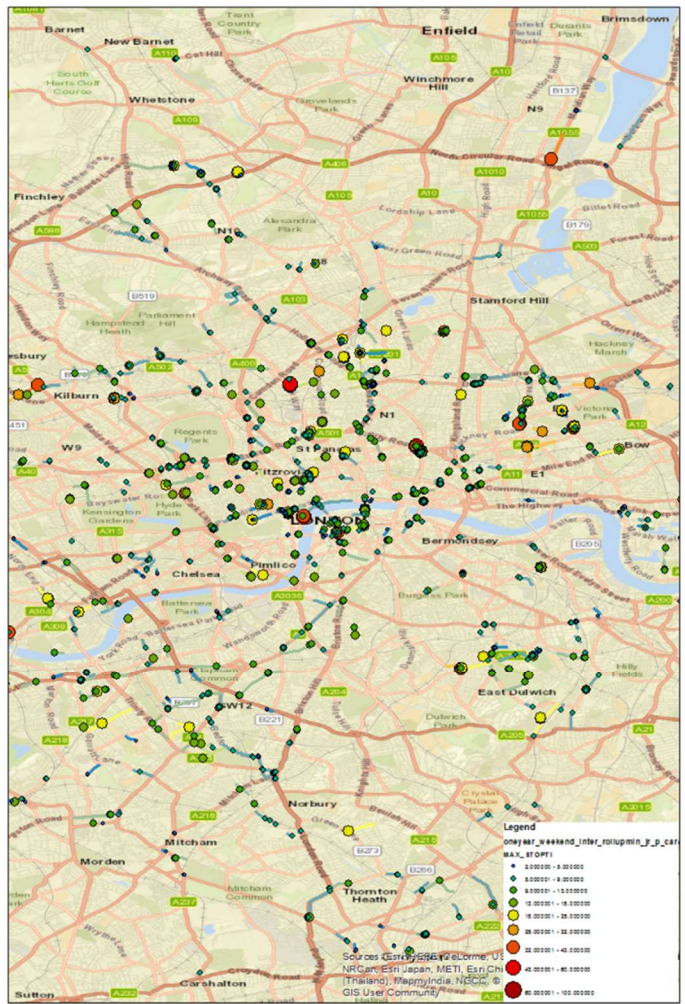Westra, E. (2015). Python Geospatial Analysis Essentials. Packt Publishing Ltd.

Figure 2. Road Speed Patterns for Weekends

Figure 3. Wait Time Patterns for Weekends