Analysing Connected Car Data to Understand Vehicular Route Choice

Elliot Karikari [*1], Manon Prédhumeau [†2], Peter Baudains[‡3] and Ed Manley [‘2]

[1] Leeds Institute for Data Analytics, University of Leeds
[2] School of Geography, University of Leeds
[3] ESRC Consumer Data Research Centre, University of Leeds

GISRUK 2023

Summary

This paper motivates the use of GPS traces to better understand the complexities of driver behaviour and movement, and introduces a new car trajectories dataset. This dataset, provided via the ESRC Consumer Data Research Centre, consist of connected car data for 50,000 vehicles over one month. We propose to analyse route choice through six cardinal statistical measures including travel distance, travel time, stop time, number of turns, angular deviation, and sinuosity. We report preliminary results on a 450 trips sample and aim to extend the analysis to the entire dataset to better understand how individuals navigate in their daily journeys.

KEYWORDS: Connected Vehicle, Geospatial Big Data, Driver Behaviour

1. Introduction

Understanding the complexities of human behaviour and movement are essential for advancements in various contexts, ranging from emergency services (Liu et al., 2022) to official statistics (Marchetti et al., 2016; Pappalardo et al., 2016). Traditional data collection methods, such as surveys, have been limited regarding collection frequency (Punzo et al., 2011). Though the use of mobile phones has been an invaluable source of data for studying movement behaviour, it does not offer the level of granularity provided by connected car data (Chen et al., 2019).

Connected car data can fill this accuracy gap as it supplies high-level spatial and temporal detail. It enables analysis at a national scale, providing a more representative view of population mobility than was previously possible with traditional data. The abundance of connected car data enables high-level analysis of factors influencing individuals’ routing behaviour. We can now analyse an accurate record of individuals journeys, the time spent travelling, their stops and the routes taken over large areas. By comparing these routes to other optimal routes, the effects of adopting one route over another can be further analysed. This provides new insight into the general patterns of human mobility and a better understanding of behaviour variation over space and time. Such understanding informs more nuanced recommendations to policymakers regarding urban planning and infrastructure monitoring.

Traditional approaches to route choice modelling assumes that all individuals make rational judgements with the same information concerning routes travelled (Wardrop, 1952). The availability of accurate transportation data has allowed the analysis of more realistic cognitive processes such as heuristic decision making (Golledge and Garling, 2001; Kaplan and Prato, 2012). Studies such as (Manley et al., 2015) have found that urban structures have an important effect on individual route choice, highlighting a widespread deviation between observed and more optimal routes. Other studies have also shown the influences of urban infrastructure on route choice, indicating that individuals spend lots of time in particular places and tend to visit such places at specific times (Pappalardo and Simini, 2018). While several studies have analysed route choice using GPS data, analysis at the national level is lacking, which may raise representativeness issues.

[*] e.n.karikari@leeds.ac.uk
[†] m.predhumeau@leeds.ac.uk
[‡] p.baudains@leeds.ac.uk
[‘] e.j.manley@leeds.ac.uk

In this paper, we contribute to the understanding of route choice by analysing route choice variation by time and space with a national scale dataset. To this end, we calculate six cardinal statistical measures: travel distance, travel time, stop time, number of turns, angular deviation, and sinuosity. In what follows, we describe the dataset and methods used for analysis before presenting some preliminary results. We then conclude by discussing the potential for further analysis.

2. Data and methods

2.1 Data overview

The dataset is made up of over 400 million GPS data points across the United Kingdom (UK), as shown in **Figure 1,** collected from 3rd June 2022 to 1st August 2022**.**

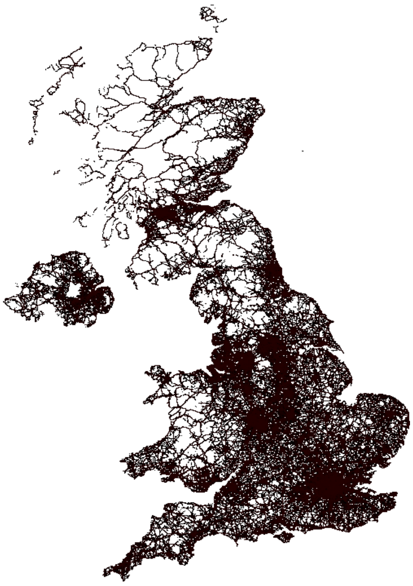
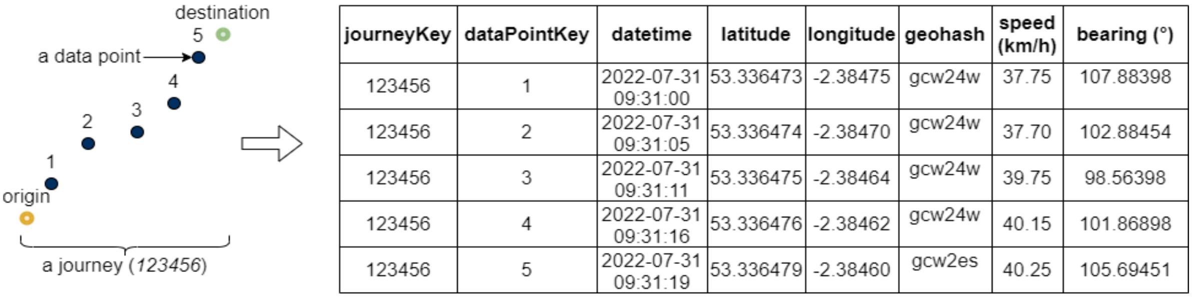


Figure 1 Distribution of connected car data.

Recordings are taken at an approximate 3-second interval and demonstrate the precision and volume of data made available by current technology. Each recording is date-stamped in the ISO8601 format to the 3-digit millisecond. Each trip consists in a set of GPS recordings and can be identified by a unique journey key, as illustrated on a dummy example in **Figure 2**. The dataset contains approximately 1,830,000 journeys. Each recording also includes information on vehicle speed, bearing angle (vehicle direction), and geographic location in the form of a geohash. To ensure privacy, the exact origin and destination of each trip are obfuscated.



journeyKey	dataPointKey	datetime	latitude	longitude	geohash	speed (km/h)	bearing (°)
123456	1	2022-07-31 09:31:00	53.336473	-2.38475	gcw24w	37.75	107.88398
123456	2	2022-07-31 09:31:05	53.336474	-2.38470	gcw24w	37.70	102.88454
123456	3	2022-07-31 09:31:11	53.336475	-2.38464	gcw24w	39.75	98.56398
123456	4	2022-07-31 09:31:16	53.336476	-2.38462	gcw24w	40.15	101.86898
123456	5	2022-07-31 09:31:19	53.336479	-2.38460	gcw2es	40.25	105.69451

Figure 2 Illustrative dummy data example

Due to the size of the dataset, this paper focus on an exploratory process undertaken on a sample of 450 trips across the North-East and North-West of the UK.

An analysis of the dataset quality indicates that the dataset has a high level of detail, with low levels of missing values, inaccuracies, and duplications. 97% of inter-event times within the sample are under 3.2 seconds, as shown in **Figure 3**.

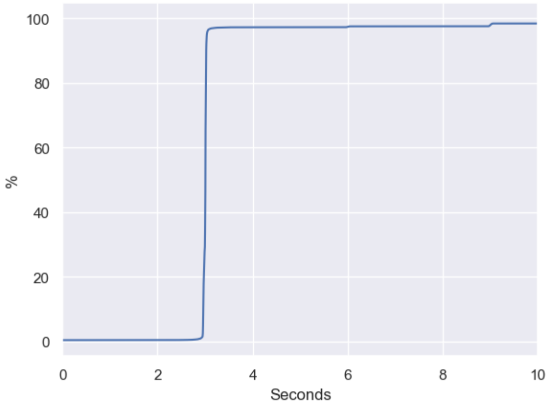


Figure 3 Cumulative distribution of inter-event time

2.2 Route analysis framework

This study aims to explore routing behaviour through the generation of descriptive statistics. No preprocessing was required as raw data is highly accurate. Scikit mobility, a Python library for human mobility analysis, was used to generate travel distance and stop time metrics (Pappalardo et al., 2019). The following statistics have been computed:

Travel Distance - This statistic measures the length of a journey (L). This is calculated in **Equation 1** as the sum of the distances between consecutive time-ordered points of the journey, where n is the number of points, and r_{j-1} and r_j are consecutive points.

$$L = \sum_{j=2}^{n} dist(r_{j-1,}\ r_j) \qquad (1)$$

Travel Time – This statistic measures the duration of a journey (D). This is computed in **Equation 2** as the difference between the timestamps of the first (t_1) and last (t_n) GPS recordings for a journey.

$$D = t_n - t_1 \qquad (2)$$

Stop Time - This statistic measures the total stop time within a journey. GPS points which were stationary (i.e. within a spatial radius of 200 m) for a minute or more were identified as stops (see Table 1 of Hariharan and Toyama, 2004 for a definition). Places of interest, identified by long stop times, are considered pull factors in analysing route choice. Shorter stops resultant from traffic are also identified.

Number of Turns - The number of turns along a journey is determined by the bearing change between two consecutive trajectory points. Bearing changes greater than 50° are considered as turns (Douglas and Peucker, 1973). This statistic helps to understand the patterns regarding routes taken, i.e. are people more likely to go along a route without many turns or not.

Cumulative Angular Deviation – The *CAD* measures the angularity of the trajectory. This is defined in **Equation 3** as the sum of the absolute differences in bearing angle between consecutive time-ordered points of the journey, where *n* is the number of points, and θ_{j-1} and θ_j are bearings of consecutive points.

$$CAD = \sum_{j=2}^{n} \left|\theta_{j-1} - \theta_j\right| \tag{3}$$

Sinuosity – The sinuosity *S* measures the trajectory efficiency, compared to a straight line. It is the ratio of the route length *L* to the straight-line distance between origin (r_1) and destination (r_n), as defined in **Equation 4**.

$$S = \frac{L}{dist(r_1, r_n)} \tag{4}$$

3. Preliminary results

Preliminary results computed on a sample of 450 trips show a summary of descriptive statistics generated per journey (**Table 1**).

Table 1 Descriptive statistics of sample (450 trips)

	Min	Max	Mean	1st quartile	Median	3rd quartile	Standard deviation
Travel distance *L* (km)	0.000	172.366	25.491	5.379	12.354	31.365	34.622
Travel time *D* (min)	0.000	101.518	12.773	3.150	7.352	16.799	14.787
Stop time (min)	1.000	43.799	6.0706	1.549	3.151	8.226	6.973
Number of turns	0.000	92.000	14.328	5.00	10.00	20.00	13.801
Cumulative Angular Deviation *CAD* (°)	0.000	15836.839	2026.964	724.390	1484.699	2776.279	1933.864
Sinuosity (one-way trips) S	1.066	249.807	7.848	1.897	3.375	3.375	7.301

The preliminary results indicate a high level of variability in trip distance and duration within the sample. The results suggest that stop time may account for nearly half of the travel time. Sinuosity was calculated excluding round trips (n=129), to avoid biasing the results due to similar origins and destinations. The descriptive statistics presented above provide insight into the data distribution, which although not representative of the entire dataset, offers valuable information for further exploration and future research.

4. Discussion and conclusion

Connected cars provide a level of data which has previously not been available within the transportation industry. The availability of digital footprints data will play an increasing role in understanding human behaviour and informing policy. In many cases, such data requires extensive processing and analysis before meaningful insights can be extracted. In this paper, we introduce a dataset from connected cars that has been made available via the ESRC Consumer Data Research Centre. We have presented an initial assessment of its quality and some preliminary results that can begin to inform new insights into human behaviour. We expect to use this dataset to identify new patterns associated with human mobility, and explore variations by day of the week, time, and UK region.

5. Acknowledgement

The data for this research have been provided by the Consumer Data Research Centre, an ESRC Data Investment, under project ID CDRC 376, ES/L011840/1; ES/L011891/1.

References

Chen, G., Viana, A.C., Fiore, M., Sarraute, C., (2019). *Complete trajectory reconstruction from sparse mobile phone data.* EPJ Data Sci. 8, 1–24.

Douglas, D.H., Peucker, T.K., (1973). *Algorithms for the reduction of the number of points required to represent a digitized line or its caricature.* Cartogr. Int. J. Geogr. Inf. Geovisualization 10, 112–122.

Golledge, R.G., Garling, T., (2001). *Spatial Behavior in Transportation Modeling and Planning* (University of California Transportation Center, Working Paper). University of California Transportation Center.

Hariharan, R., Toyama, K., (2004). *Project Lachesis: Parsing and Modeling Location Histories,* in: Egenhofer, M.J., Freksa, C., Miller, H.J. (Eds.), Geographic Information Science, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 106–124.

Kaplan, S., Prato, C.G., (2012). *Closing the gap between behavior and models in route choice: The role of spatiotemporal constraints and latent traits in choice set formation.* Transp. Res. Part F Traffic Psychol. Behav. 15, 9–24.

Liu, Y.-H., Albuquerque, O. de P., Hung, P.C.K., Gabbar, H.A., Fantinato, M., Iqbal, F., (2022). *Towards a Real-Time Emergency Response Model for Connected and Autonomous Vehicles.*

Manley, E.J., Orr, S.W., Cheng, T., (2015). *A heuristic model of bounded route choice in urban areas.* Transp. Res. Part C Emerg. Technol. 56, 195–209.

Marchetti, S., Caterina, G., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L., Gabrielli, L., (2016). *Small Area Model-Based Estimators Using Big Data Sources*. J. Off. Stat. 31, 263–281.

Pappalardo, L., Rinzivillo, S., Simini, F., (2016). H*uman Mobility Modelling: Exploration and Preferential Return Meet the Gravity Model.* Procedia Comput. Sci., The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016) / Affiliated Workshops 83, 934–939.

Pappalardo, L., Simini, F.,(2018). *Data-driven generation of spatio-temporal routines in human* *mobility.* Data Min. Knowl. Discov. 32.

Pappalardo, L., Simini, F., Barlacchi, G., Pellungrini, R., (2019). *Scikit-mobility: a Python library for the analysis, generation and risk assessment of mobility data.*

Punzo, V., Borzacchiello, M.T., Ciuffo, B., (2011). O*n the assessment of vehicle trajectory data accuracy and application to the Next Generation SIMulation (NGSIM) program data.* Transp. Res. Part C Emerg. Technol. 19, 1243–1262.

Wardrop, J.G., (1952). *Road paper. some theoretical aspects of road traffic research.* Proc. Inst. Civ. Eng. 1, 325–362.

Biographies

Elliot Karikari is a Data Scientist at the Leeds Institute for Data Analytics, University of Leeds. He has research interests in geodemographics, population, human behaviour, and health.

Manon Prédhumeau is a Research Fellow at the School of Geography of the University of Leeds. Her research interest includes agent-based modelling and simulation of human behaviour, applied to urban mobility and future transportation.

Peter Baudains is a Research Data Scientist at the ESRC Consumer Data Research Centre at the University of Leeds. He has research interests in building data-driven models of human behaviour and in developing derived data products for open use.

Ed Manley is Professor of Urban Analytics in the School of Geography, University of Leeds, and Turing Fellow at the Alan Turing Institute. His research aims to improve our analyses and models of human spatial behaviour, and understand how these behaviours shape urban systems and public health.