Clustering reveals key behaviours driving human movement network structure

Hamish Gibbs[1*], Rosalind M. Eggo[2], and James Cheshire[1]

[1] Department of Geography, University College London, London, United Kingdom
[2] Centre for Mathematical Modelling of Infectious Disease, London School of Hygiene and Tropical Medicine, London, United Kingdom

* Corresponding author: Hamish.Gibbs.21@ucl.ac.uk

Abstract

We analysed daily travel patterns from in-app GPS data in the United Kingdom to identify characteristic modes of travel behaviour, and the relative importance of different behaviours for the topology of the overall travel network. We clustered the detailed travel trajectories and identified four characteristic modes of travel that we named: *regular-travel*, and *stay-at-home* which represented the majority of travel days, and two 'long tailed' travel modes: *long-distance*, and *away-from-home*, which represented ~2.6% of travel days. We focused on these second two modes, in which individuals travel a long distance from home and deviate from predictable routines. We demonstrated how this relatively small portion of travel behaviour plays an outsized role in influencing the connectivity of the aggregate travel network. This analysis highlights the need to understand individual behavioural variations that can dictate the collective dynamics recorded by aggregate descriptions of human travel.

1. Introduction

Human mobility data derived from mobile phones are increasingly used to measure economic activity (1,2), predict the spread of disease (3–6), forecast travel demand (7,8), measure responses to natural disasters (9,10), and understand human social dynamics (11,12). The use of mobile phone mobility data accelerated during the COVID-19 pandemic as a result of accessible mobility indices produced by major technology companies such as Google (13–16). To minimise the risk of revealing personally disclosive information, granular mobility data are typically aggregated to infer domain-specific population-level dynamics, ranging from measures of 'activity' at specific points-of-interest (17), to measures of travel volume between pairs of locations (8,18).

Aggregating individual mobility data removes information about the unique travel behaviours of individuals (19). However, aggregation can also mask important individual differences in travel behaviour. This leads to uncertainty about the specific travel behaviours that define collective dynamics. A key unknown in the analysis of aggregated travel networks is the degree to which these networks measure a representative subset of individual behaviours, or are influenced by biassed sampling of individuals or activities.

The problem of uncertainty regarding individual behaviour in aggregated travel networks is particularly important in light of the long-tailed dynamics observed across a wide range of human mobility measures (20–25). Long-tailed distance and visitation frequency distributions mean that a relatively small amount of travel behaviour can define the collective dynamics observed in aggregated travel networks. Similar distributions are common to many other measures of human behaviour (26–29). At present, however, there is little insight into how researchers can detect individual behavioural differences that may influence aggregate mobility measures, without presenting risks to individual privacy. In this analysis, we endeavour to provide a view "underneath" an aggregated travel network, seeking to illustrate characteristics of daily travel that drive the observed travel network topology, and to explore the temporal and spatial context in which these dynamics arise.

We use passively collected in-app mobile phone location data to better understand the individual behaviours responsible for long-tailed dynamics in aggregate travel networks. We demonstrate an approach for resolving questions about the quality and representativeness of

mobile phone location data through clustering of behavioural patterns at the individual level, and aggregation within these identified clusters, as opposed to treating a dataset as a whole. This approach maintains individual privacy, while providing insight into the individual behaviours which comprise the aggregated network.

2. Results

2.1 Detecting travel behaviour from in-app location histories

We analysed anonymized GPS location data recording 12 million days of travel by 782,000 mobile devices in England and Wales collected between January 1st, 2018 and December 31st, 2019. We filtered GPS positions for high accuracy observations (Methods and Supplemental Figure 1.1), and applied a two-stage 'stop-point' detection algorithm (30) to sequences of GPS locations which identified 43 million 'visits' where mobile devices were stationary within a spatial radius of 200m for a minimum duration of 5 minutes (Figure 1a-d). We then selected days of travel with observation windows of at least 5 hours per day (Supplemental Figure 1.2), yielding a dataset recording the positions of 10.9 million days of travel by 677,00 devices (Figure 1e).

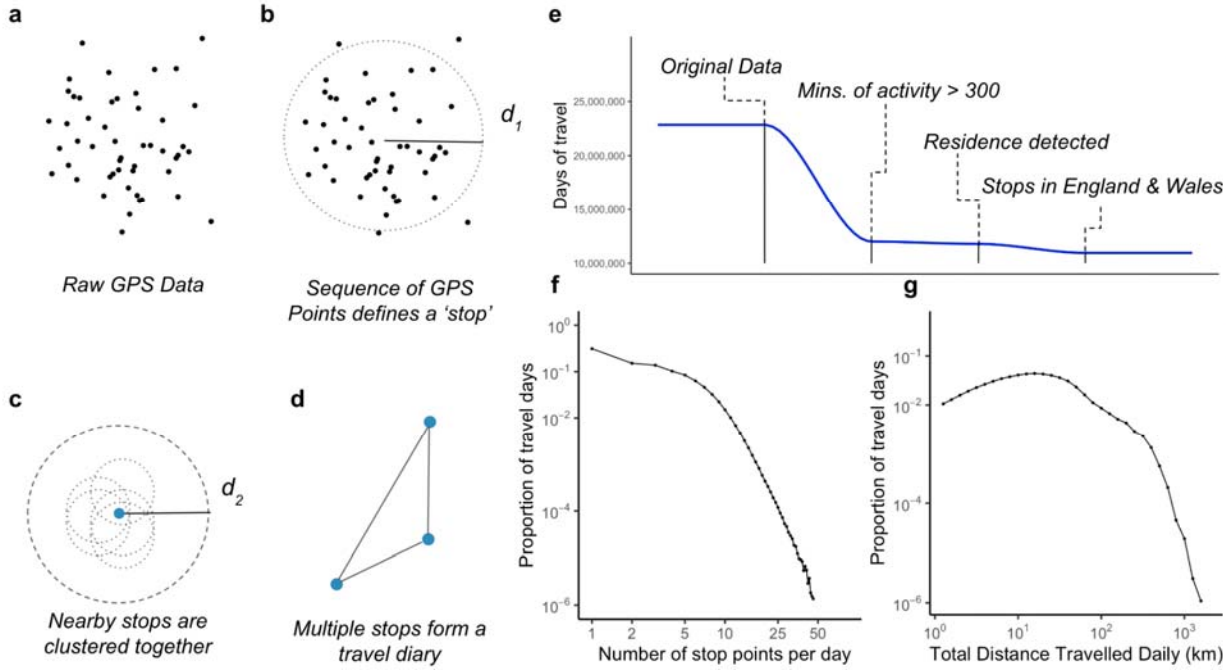

Figure 1. Overview of the travel dataset. *a-d) Diagrams illustrating the detection of stop-points from a sequence of GPS locations. Distance threshold d_1 is the maximum spatial 'roam' of a device, d_2 is the spatial threshold used to attribute nearby stops to the same location. e) The size of the mobile phone location dataset (total days of travel) through different stages of*

sample selection. f) The distribution of the number of stop-points for all days of travel and g) the total distance travelled for daily sequences of stop-points with travel greater than 1km.

In the travel dataset, individual devices were present for varying length periods. Devices are recorded in the dataset for a median of 4 days (IQR: 1, 14). To account for the varying-length sampling windows, we frame our analysis around 'days' of travel as opposed to individual devices, where one device may contribute multiple days of travel to the dataset. In the dataset, 31% of travel days recorded only one visited location, while 43% of travel days recorded a total distance travelled less than 1 km (Figure 1f,g).

2.2 Daily travel activity clustering

We identified groups of similar daily travel activity based on measures of travel distance, visitation frequency, and entropy (predictability) using k-means clustering informed by a principal component analysis of the daily travel measures (Supplemental Figure 2.1). The clustering analysis identified four clusters: long-tailed clusters (*long-distance, away-from-home*) account for only ~2.6% of the dataset while the *regular-travel* and *stay-at-home* clusters encapsulate 59% and 38% of the dataset respectively (Table 1). The long-tailed clusters were primarily differentiated by high total distance travelled (*long-distance)*, and high minimum distance from home (*away-from-home*) (Figure 2a). The distribution of total distance travelled (Figure 2b) highlights the difference in travel behaviour across clusters, with the *long-distance* cluster showing high total distance travelled, while *away-from-home* is distinguished by a high minimum distance travelled from home (Figure 2c).

Cluster	Descriptive Title	Travel days	Description
1	*Regular-travel*	59%	Variable total travel distance, low distance from home, higher likelihood to visit a new location, variable entropy
2	*Stay-at-home*	38%	Low travel distance, low likelihood of visiting a new location, low entropy, high time spent at home
3	*Long-distance*	1.6%	High distance travelled, low minimum distance from home, high likelihood of visiting new locations
4	*Away-from-home*	0.9%	High minimum distance from home (never near home location), variable total distance, high likelihood of visiting a new location

Table 1. Description of identified modes of travel. *The result of cluster analysis of travel based on distance, visit frequency, and entropy of daily travel patterns.*

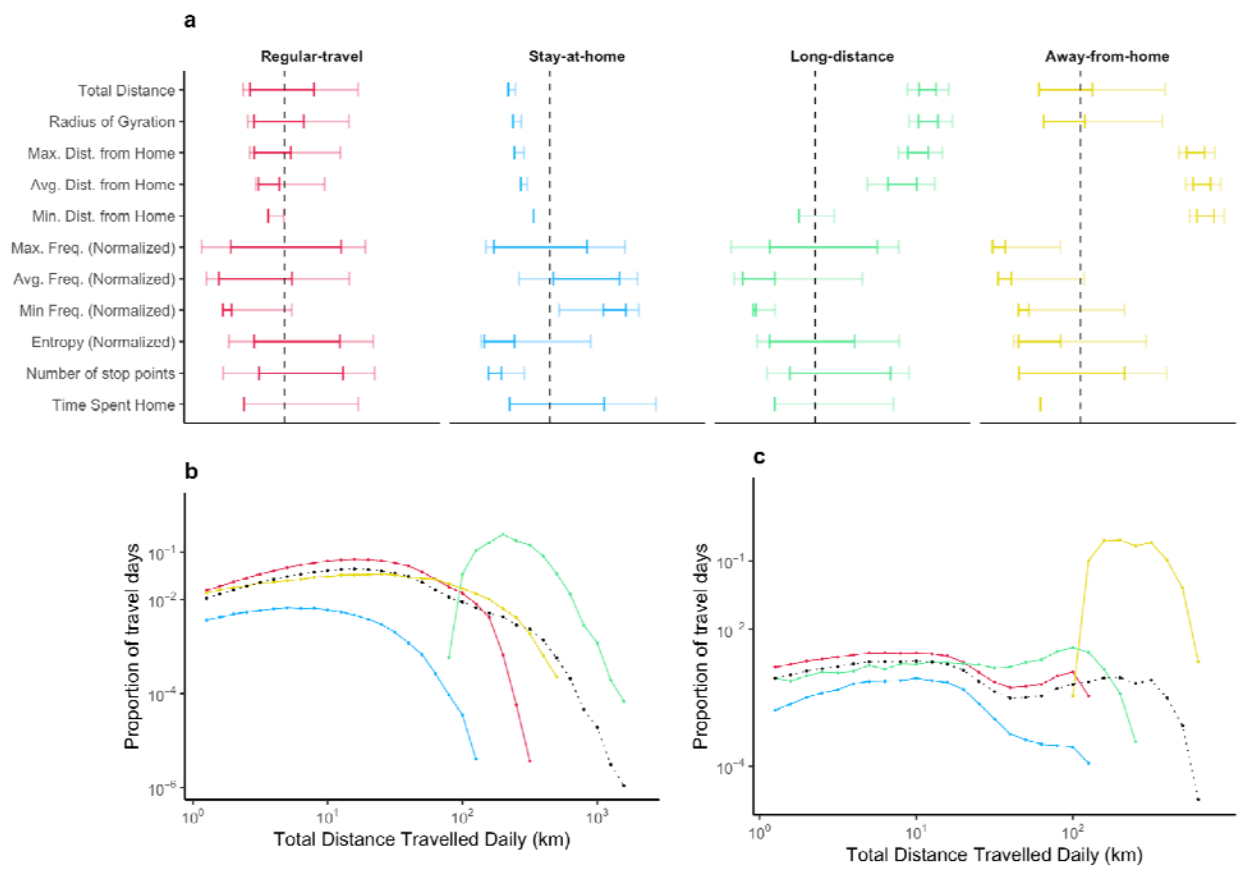

Figure 2. Clustering identifies long-distance, atypical travel. *a) The distribution of travel measures for each identified cluster. Intervals indicate 90% and 50% density intervals for observations within each cluster. Values have been transformed for comparison across measurement scales. Dotted black line indicates average for each variable. b) The distribution*

of total distance travelled for each cluster. Dotted black lines in panels b and c indicate the variable distribution for the overall dataset. c) The distribution of the minimum distance travelled from home for each cluster.

2.3 Effect of 'long-tailed' travel on network connectivity

To understand how the identified clusters of travel behaviours influence the topology of the travel network, we aggregated movements into origin-destination flows across a regular spatial grid (31). We then performed an experiment, removing one cluster at a time, to observe the resulting changes on the travel network topology (Table 2). We found that *regular-travel* formed (42%) of the unique network connections ('edges') in the aggregated network, and that *long-distance* movements, although they comprised 1.6% of all travel days, represented 35% of the unique edges in the network. By contrast, the other 'tail' cluster, *away-from-home,* represented only 1.9% of unique edges, while removal of *stay-at-home* movements had almost no influence on the network.

Cluster removed	**Edges Removed**	**Days in Removed Cluster**
None	0%	0%
Regular-travel	42%	59%
Stay-at-home	0.1%	38%
Long-distance	35%	1.6%
Away-from-home	1.9%	0.9%

Table 2. Number of edges removed by each cluster compared to the proportion of travel days. *Shows changes in the number of edges in the aggregated travel network caused by the removal of different clusters of daily travel activities.*

The number of unique edges represented by each cluster is only one means of assessing changes to the topology of the travel network. We computed additional measures to better illustrate how the aggregated travel network responded to the removal of each cluster of travel behaviours (Table 3). We found that some spatially explicit measures, such as the diameter of the network (maximum distance between any nodes), and the spatially-explicit mean shortest path length, remained similar with any cluster removed. This is because the network retained the same spatial extent despite the removal of some movement modes. The topology of the network, however, showed sensitivity to the removal of certain clusters of behaviour, in particular to *long-distance* movements. For example, the topological mean shortest path length

(measuring the number of jumps required to connect nodes), increased from 2.62 with the overall network, to 3.56, reflecting increasing difficulty traversing the network. For *regular-travel* movements, which represent the majority of travel days, topological mean shortest path length only increased to 2.75. Transitivity of the network (probability of connection between adjacent nodes) also increased with the removal of *long-distance* movements, from 0.26 to 0.38, indicating a higher likelihood of isolated communities. Conversely, with the removal of *regular-travel* movements, transitivity dropped to 0.20.

Cluster removed	**Diameter (km)**	**Mean Shortest Path (km)**	**Mean Shortest Path (Topological)**	**Transitivity**
None	750.79	228.58	2.62	0.26
Regular-travel	1048.40	231.65	2.75	0.20
Stay-at-home	760.27	228.60	2.62	0.26
Long-distance	763.40	229.09	3.56	0.38
Away-from-home	753.45	228.67	2.62	0.26

Table 3. Measures of network characteristics with different behavioural clusters removed. *A collection of measures characterising the spatial and topological changes in the travel network with the removal of different activity clusters.*

Changes in the degree and edge distance distributions provided insight into the topological changes to the network. With *regular-travel* removed there were a greater proportion of low-degree nodes while with *long-distance* movements removed, there were comparatively more nodes in the middle of the degree distribution (Figure 3a). However, the removal of either cluster reduced the distribution of high-degree nodes relative to the overall travel network. The removal of *regular-travel* movements led to fewer edges in the middle of the distance distribution, while the removal of *long-distance* movements, removed nearly all of the longer distance edges (Figure 3b). Similar to the degree distribution, the distribution of edge distances was not affected by the removal of either *stay-at-home* or *away-from-home* movements.

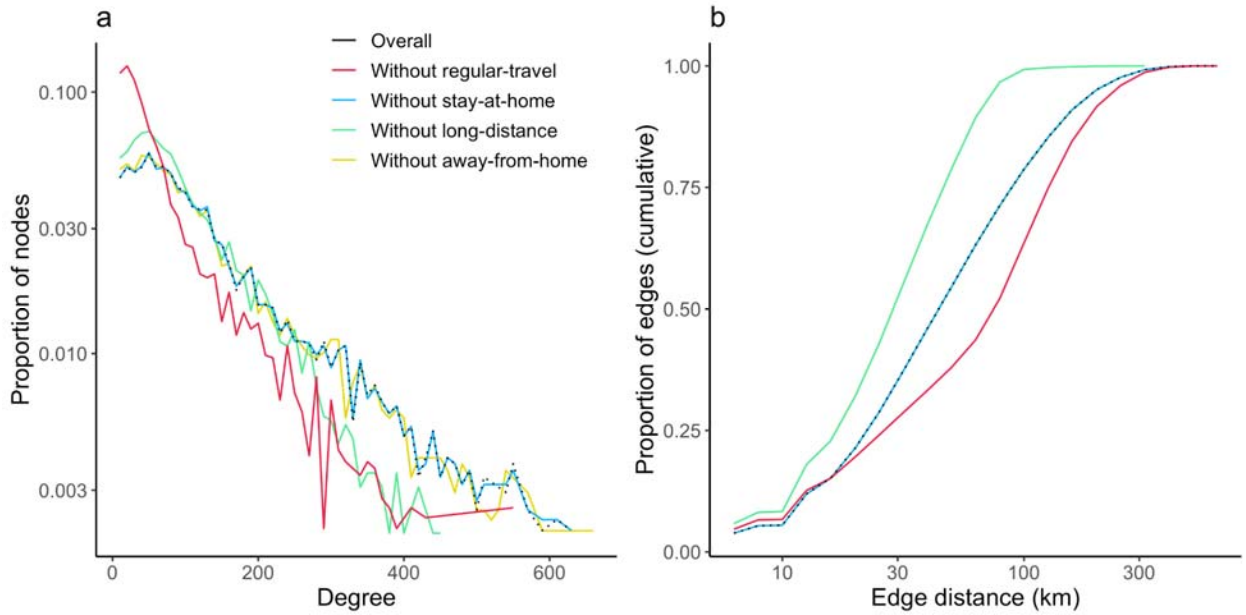

Figure 3. Degree and edge distance distribution with behavioural clusters removed. *a) The degree distribution and b) edge distance distribution for the aggregated travel network, for travel networks with each activity cluster removed one-at-a-time.*

It is worth noting that the effect on the aggregate travel network of removing each cluster was sensitive to the spatial scale of aggregation. To explore this sensitivity, we performed repeated aggregation of the travel network to varying sized spatial grids (Supplemental Section 4). We found that some network characteristics, such as the spatially-explicit mean shortest path length, varied predictably with the scale of aggregation, while others, such as the clustering coefficient were highly sensitive to the spatial scale of aggregation.

2.4 Spatiotemporal context of 'long-tailed' dynamics

We based our clustering analysis around days of travel to permit comparison of individuals with varying data collection periods. However, to determine if the small proportion of 'long-tailed' travel days we found is identifying infrequent days of travel by many individuals, or a smaller group of individuals who primarily exhibit long-tailed travel behaviour, we calculated the proportion of travel days labelled *long-distance* or *away-from-home* relative to the total number of days recorded for each device. We found a mix of these two explanations: overall, 72,900 individuals (10.7%) had at least one day of 'long-tailed' travel, while a very small number of individuals (1400, 0.2%) had >95% of their days of travel labelled as 'long-tailed' (Supplemental Figure 5.1).

The temporal distribution of long-tailed travel showed peaks in both *long-distance* and *away-from-home* travel days coinciding with national holidays in the United Kingdom and with the timing of school term dates (Figure 4). Over time, we observed a relative increase in *long-distance* travel compared to *away-from-home* travel. This may indicate sensitivity of our analysis to the types of activities included in the dataset over time, resulting of heterogeneous sampling of activities over time (Supplemental Figures 5.2, 5.3).

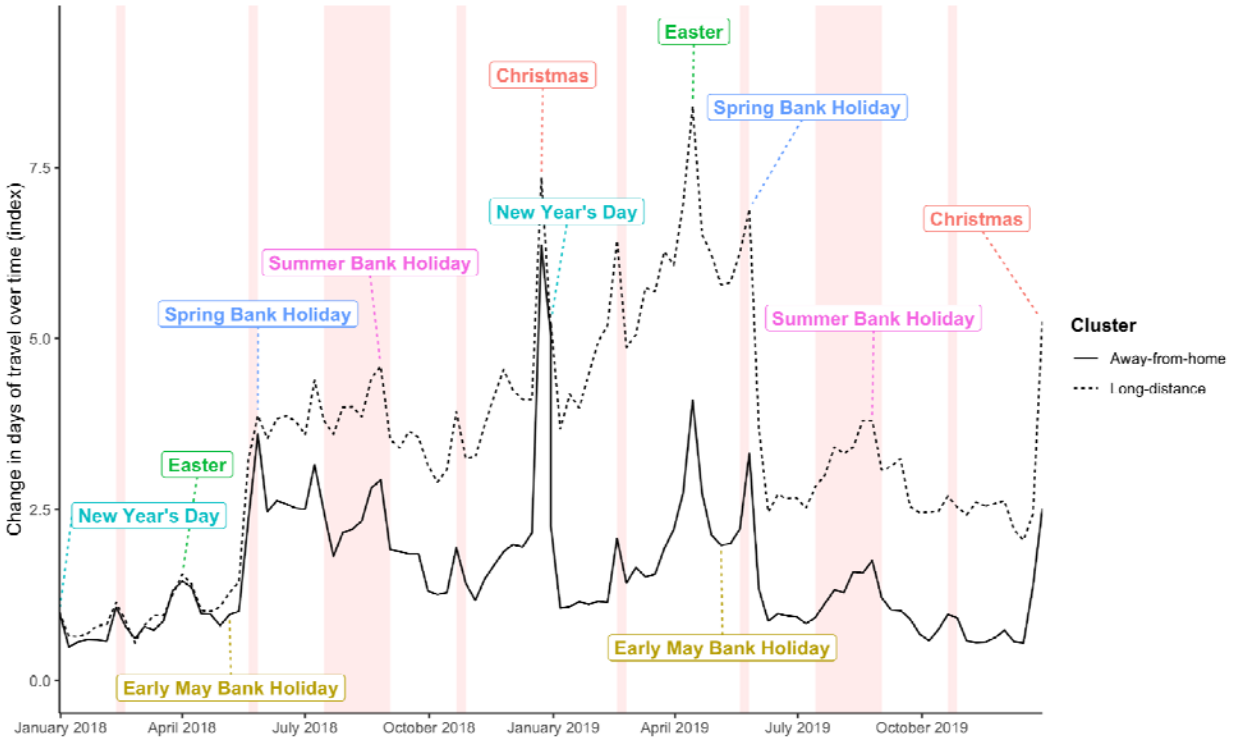

Figure 4. Timing of travel in 'long tail' clusters. *The quantity of days in 'long-tail' clusters (Long-distance and Away-from-home) per week. Annotations show the timing of key holiday dates in the United Kingdom. Shaded areas show school holiday dates: February half term, May half term, Summer holidays, October half term. Figure shows a scaled index, accounting for temporal heterogeneity in the overall volume of data collected per week. For original distribution, see Supplemental Section 5.1.*

To determine how our travel behaviour clusters mapped to official UK geodemographic categories (32), we matched the geographic location of these categories to our identified home location of a device (Supplemental Figure 6.1, Supplemental Table 6.1). We found that *long-distance* and *away-from-home* travel was notably more common for "Rural Residents", "Suburbanites", and "Urbanites" and less common for lower SES groups "Hard-pressed living" and "Constrained city dwellers", and for urban categories "Multicultural Metropolitans" and "Ethnicity Central" (Figure 5). Though official geodemographic categories provide only a coarse

summary of differences between populations, our findings point towards a difference in urban and rural patterns of travel, and the potential influence of socioeconomic characteristics on longer distance travel (see Supplemental Table 6.1). It is important to note, however, that our original dataset suffers from over- and under-sampling of specific groups, with especially high sampling of urban populations “Multicultural Metropolitans”, “Ethnicity Central”, and “Cosmopolitans” (Supplemental Figure 6.2).

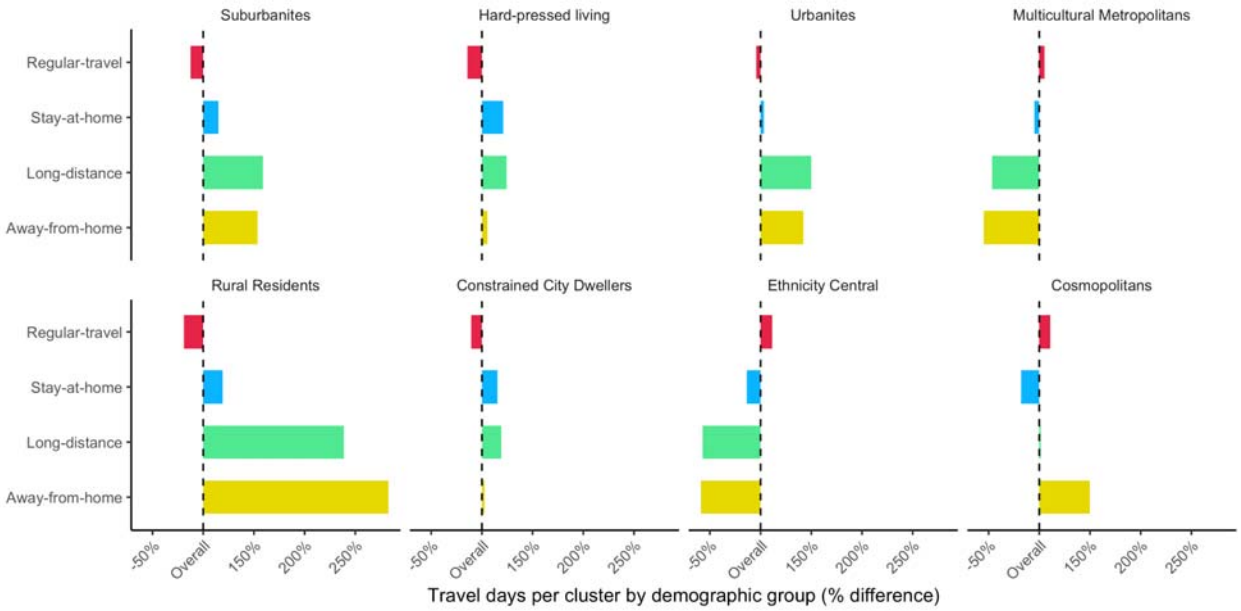

Figure 5. Association between geo-demographic categories and clusters of travel. *Likelihood of travel days categorised in a particular travel cluster, based on the geodemographic label of the home location of a device. Horizontal axis shows percent difference relative to the overall distribution of travel clusters. For maps of national distributions of geodemographic categories, see Supplemental Figure 6.1.*

3. Discussion

This analysis demonstrates how granular data are critical to identifying how differences in individual activity patterns have an outsized impact on the structure of an aggregated travel network. In particular, when days in the *long-distance* cluster were removed, the network lost most long distance connections, exhibited greater clustering, and greater topological distance between nodes. For researchers with access only to aggregate travel networks, it is crucial to understand what features of network structure are lost by aggregation.

We show how activity in the *long-distance* and *away-from-home* clusters is distributed across individuals' days of travel. Many individuals (~10%) had travel days labelled as *long-distance* or *away-from-home*, but these clusters were the dominant pattern of activity for only a small number (0.2%) of this subset of individuals. This provides evidence that important behaviours of "exploration", such as travelling long distances and visiting new locations form a component of many individuals' activity, instead of being confined to a small group of unique people (33).

In addition to identifying key behavioural modes, we also provided insight into the timing of 'long-tailed' clusters of travel behaviour. These long-tailed dynamics followed a clear temporal trend with respect to holidays and dates of school terms in the UK. We also observed patterns of increased likelihood to engage in days of 'long-tailed' travel in specific geodemographic groups, such as a higher likelihood among "Rural Residents", and lower likelihood among socioeconomically disadvantaged groups. We note, however, that our dataset represents an uneven sampling of the UK population with respect to geodemographic characteristics, making these findings less conclusive.

The uneven geodemographic sampling in this dataset highlights a tradeoff between the strengths and limitations of in-app mobile phone location data. On one hand, these data contain a variety of documented limitations including uncertainty regarding the correspondence between individuals and devices, intermittent observation windows across individuals, data collection driven by individuals' interaction with a specific application, the potential for 'opt-in' bias for those willing to consent to location sharing, or biassed sampling caused by applications marketed towards specific consumer segments (34,35). On the other hand, this type of large scale mobility data is well suited to capturing observations in the 'tails' of the behavioural distribution, which would otherwise be costly or impossible to detect using traditional survey methods. In-app location data is thereby well suited to detecting uncommon travel behaviours that can have an outsized effect on population-level dynamics, such as those in the 'long-tailed' clusters identified in this study (36).

In the analysis of large-scaled mobility data, preserving individual privacy is of the utmost concern. Our research, however, shows how some questions about the quality of mobility datasets and the underlying human dynamics that they represent cannot be answered merely through treating a dataset as an undifferentiated whole. In this analysis, we

present a ‘middle path’ in which we illuminate some of these individual differences that drive important structural changes in a travel network, while avoiding risks of identifying personal information by grouping activities into characteristic clusters. To do this, we performed individual-level analysis within a secure research facility, and then described the collective characteristics of identified behavioural clusters (37).

4. Conclusion

Our analysis highlights the need for scientific understanding of the biases present in sources of mobility data and an understanding of how these biases shape observed dynamics. This understanding cannot be achieved merely through the analysis of aggregated mobility data, and highlights the tension between the imperative to preserve privacy, and the need to understand limitations within mobility data in order for researchers to draw robust conclusions. Publicly available mobility datasets should be accompanied by granular analysis which highlights the characteristics of the data generation process and illuminates how a dataset is influenced by individual variations in modes of travel behaviour, temporal and spatial sampling heterogeneity, and how these variations influence the human dynamics reported in the dataset.

5. Methods

5.1 Data description

We used GPS location data collected from a panel of mobile applications between January 1st 2018 and December 31st 2019 collected by a commercial location data aggregator in the United Kingdom. Unique device identifiers, and identifiers pertaining to the application that collected location data were anonymized prior to data sharing. These anonymized individual location histories were then analysed within a secure data environment, with research outputs checked by data scientists prior to release to prevent the disclosure of personally identifiable information (37).

5.2 GPS data preprocessing

We transformed sequences of GPS locations (recording the time, x and y coordinates of a device) into a diary of ‘stop points’ describing locations where a device remained within a defined spatial diameter d_1 for a minimum time threshold t according to the method proposed in (30). We then used agglomerative clustering to group nearby visits (defined by a distance radius

d_2) into recurrent 'visits' to a set of locations for each device (Figure 1c). We defined an 'activity' as a collection of GPS positions remaining within a spatial radius of 200m for a duration of at least 5 minutes. These parameters are based on established definitions of the minimum duration of activities in travel surveys, and average walking speed (38,39).

5.3 Sample selection

Due to turnover in the panel of users in the location dataset, individuals' activity was recorded for varying length periods. This is a well-documented limitation of mobile phone location data including in-app data of the type used in this study (40). We took three steps to account for this bias in the dataset: (1) we based our clustering of travel modes around 'days' of travel, rather than individual devices by splitting visits crossing midnight (00:00 GMT) into distinct visits (ending 23:59:59 and beginning 00:00:01); (2) We filtered the location dataset for days of high quality location sampling, defined as days with at least 300 minutes of recorded activity, based on previous work using similar mobile-phone in-app location data (41). We performed a further sensitivity analysis to understand the effect that this threshold had on the size of the sample used in subsequent analysis (see Supplemental Figure 1.2); (3) We normalised measures of daily activity that depended on the sampling window of a particular device for comparison across devices with different sampling intervals. We estimated the home location for devices with sufficient daily observation windows based on a simple heuristic which identifies the most frequently visited nighttime location (defined between 22:00 and 06:00) per month.

5.4 Discrimination of modes of daily travel

To discriminate between different modes of daily travel, we computed eleven features based on the travel diaries of individual mobile devices (Table 4). We chose metrics which highlighted the distance an individual travels; in total and in displacements from their identified home location, and their frequency of visitation to the same locations; including frequency of visits to a location on a given day and entropy (the probability of visiting a given location out of all possible locations visited by an individual). We also calculated the total number of visits to distinct locations per day and the amount of time spent at individuals' identified home locations.

Some daily metrics, like the total distance travelled by a device, are directly comparable between devices, while frequency measures are related to the total amount of time in which a device is included in the dataset. Similarly, entropy is related to the total number of locations

ever visited by an individual. To account for this dependence, we normalised frequency and entropy measures by the observation window of a device (defined as the number of days the device was observed), and the total number of locations visited by a device, respectively.

Measure	**Measures**	**Units**	**Description**
Total distance travelled	Daily total	Kilometre	Total daily distance travelled
Radius of gyration (r_g)	Daily	Kilometre	Distance travelled around a daily "centre of mass"
Distance from home	Daily maximum, minimum, mean	Kilometre	Distance travelled away from identified home location
Frequency of visitation ($p(i)_{norm}$)	Maximum, minimum, mean		Likelihood to visit previously-visited locations on a given day (normalised by days observed)
Normalised entropy (H_{norm})	Daily		Predictability of visitation with respect to all possible locations visited by a device (normalised by days observed)
Number of visits	Daily	Count	Daily number of locations visited
Time spent at home	Daily	Seconds	Amount of time spent at identified home location

Table 4. Measures of daily travel activity. Measures of travel activity computed for individual days of device activity.

Descriptions of individual behaviour are based on a set of N identified stop points $V_{1...N}$ for each day of travel activity for each device. The sequence of stop points for a specific day are defined by stop point identifiers $Stop_{1..N}$, spatial coordinates $(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)$, and time intervals $(t_{start_1}, t_{end_1}), (t_{start_2}, t_{end_2}), \ldots (t_{start_N}, t_{end_N})$.

We calculate distances (total distance travelled and distance from home measures as the Euclidean distance between the locations of relevant visits.

Radius of gyration r_g is based on the centre of mass (x_{cm}, y_{cm}) of a daily trajectory where:

$$x_{cm} = \frac{\sum_{i=1}^{N} x_i}{N} \qquad \text{(Equation. 1)}$$

$$y_{cm} = \frac{\sum_{i=1}^{N} y_i}{N} \quad \text{(Equation. 2)}$$

r_g is derived from squared displacements ${r_i}^2$ from the centre of mass for each location i in the trajectory:

$${r_i}^2 = (x_i - x_{cm}) + (y_i - y_{cm}) \quad \text{(Equation. 3)}$$

Thus:

$$r_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N} {r_i}^2} \quad \text{(Equation. 4)}$$

We calculate Normalised Frequency of Visitation $p(i)_{norm}$ for each location i in the trajectory relative to the number of observation days for each device D based on the number of times the location was visited by an individual v_i .

$$p(i)_{norm} = \frac{v_i}{D} \quad \text{(Equation. 5)}$$

We calculate Normalised Entropy H_{norm} based on the probability of visiting individual location $p(i)$ based on the number of visits to each location v_i and the total number of visits to all locations V:

$$p(i) = \frac{v_i}{V} \quad \text{(Equation. 6)}$$

Entropy H is then defined as:

$$H = -\sum_{i=i}^{N} p(i) \cdot log_2(p(i)) \quad \text{(Equation. 7)}$$

And Normalised Entropy H_{norm} is defined as:

$$H_{norm} = \frac{H}{D} \quad \text{(Equation. 8)}$$

Number of Visits is the count of distinct visits recorded for an individual per-day, and Time Spent Home measures the total number of seconds an individual was observed in at their identified home location.

5.5 Principal component analysis

We used principal component analysis (PCA) to account for collinearity among the daily travel measures which can influence the results of clustering analysis, and to reduce the dimensionality of the daily travel measures, increasing the computational efficiency of the clustering. We assessed the correlation between input features prior to performing the PCA to understand possible axes of differentiation in the original data and to identify highly correlated input features (Supplemental Figure 2.2). After performing PCA, we assessed the correspondence between principal components and the original features by computing the Pearson Correlation Coefficient between the components and the original features. This gives an indication of the axes along which subsequent clustering identified groupings of similar activity patterns. We used the first 6 principal components (a 45% dimensionality reduction) which captured >95% of the variance of the original dataset (Supplemental Figure 2.3).

5.6 Clustering analysis

We clustered the principal components of daily travel patterns using the k-means algorithm (42). K-means is a widely used clustering algorithm which partitions a series of data points into user defined k groups by repeated random selection of cluster centroids to identify the partition which minimises within-cluster variance. We chose to partition daily activity patterns using the k-means algorithm due to its efficient implementation and ability to handle the clustering of large datasets (43). Our choice of clustering algorithm, combined with dimensionality reduction from the PCA allows for the efficient clustering of >10 million daily activity patterns.

To gain an intuitive understanding of the effect of different numbers of clusters on our dataset, we created a clustergram (44,45) for values of k between 2 and 9. The clustergram displays the size, and PCA-weighted similarity of clusters as well as the transition of data between clusters at different values of k. This visualisation can be a useful tool to spot regularities in the dataset across different numbers of clusters (for example, “branches” of similar clusters at higher values

of k). To choose an appropriate value of k we then assessed the quality of the identified clusters using three measures: Silhouette Score (46), Davies-Bouldin Index (47), and Calinski-Harabasz Index (Variance Criterion Ration) (48). Each metric highlights different aspects of the clusters: Silhouette Score compares the within-cluster distance to the distance between the next-most-similar cluster, Davies-Bouldin Index measures the average of the maximum ratio of within-cluster distance to between-cluster distance for each cluster, and Calinski-Harabasz Index measures the ratio of the sum of within-cluster variance to between-cluster variance. We selected 4 clusters based on the value of *k* producing high Silhouette Score, low Davies-Bouldin Index, and high Calinski-Harabasz Index (Supplemental Figure 3.1).

Calculation of clustering quality metrics requires the computation of distances between all observations which is not practical for the size of our dataset, as the number of required distance calculations scales as $O(n^2)$ to the number of observations (quadratic complexity). To reduce the number of required distance calculations, we performed bootstrapping for 1000 samples of 0.1% of the original dataset to produce a density estimate of each clustering quality metric.

5.7 Effect on travel network connectivity

To assess the influence that identified clusters have on the topology of the overall travel network, we aggregated daily movement patterns to a regular spatial grid called H3 (31). We then performed a series of experiments, removing identified clusters of travel from the aggregated network one-at-a-time and assessed how this removal affected different measures of network connectivity. In an effort to capture topological changes to the network, we computed five measures: the total number of edges in the network, the average shortest path distance (both spatially-explicit and topological distance), the diameter (maximum distance between any nodes), transitivity (also known as clustering coefficient), describing the number of triangles or cycles of length three in the network.

5.8 Spatiotemporal characteristics of long-tailed travel

To understand temporal dynamics of travel in different clusters, we aggregated cluster labels, which apply to unique days of travel, into weekly periods and calculated the proportion of travel days represented by each cluster relative to the total number of travel days each week. The overall number of travel days varies through time in the dataset due to the dynamic nature of

data collection, where devices may appear or drop out of the dataset over time. We found that the time series of cluster membership was influenced by changes in the overall volume of data in the dataset (Supplemental Figure 5.2, Supplemental Figure 5.3). To relate periods of high and low data volume, we calculated a scaled index of change in cluster membership (Δp_t) over time periods $t \subset 0 \ldots n$ based on the proportion of travel days represented by each cluster per week (p_t) and the number of travel days per week (n_t):

$$\Delta_t = p_t \cdot \frac{n_t}{\sum n_t}$$

(Equation. 9)

We also explored geodemographic patterns in the assignment of cluster labels to days of travel from individuals in different geodemographic groups. To do this, we attributed devices to the geodemographic group of their identified home location based on the 2011 Output Area Classification (from the most recent UK census preceding the study period). The Output Area Classification is an official geodemographic classification released by the UK Office for National Statistics based on the characteristics of Output Areas, small statistical geographies which represent 110 and 139 households (32,49). The principle of geodemographic analysis is to provide a meaningful description of groupings of individuals with similar characteristics. These groupings are identified through a clustering analysis of census characteristics, and are provided with descriptive titles and ‘pen-portraits’ (short descriptions of key cluster characteristics). Geodemographic categories are classified hierarchically into supergroups, groups, and subgroups which provide increasingly detailed summaries of common geodemographic characteristic. In this study, we use geodemographic supergroups, the most general geodemographic category. We calculate the sampling bias of our mobile phone location dataset based on an index of over- and under-sampling which identifies bias with respect to the distribution of geodemographic categories of the overall UK population (Supplemental Figure 6.2).

Competing Interests

The authors declare no competing interests.

Ethics

This research has been approved by the University College London Research Ethics Service (ref: 21813/001).

Data Availability

This analysis is based on GPS location data collected by the UK-based data aggregator Huq. Data is available for academic research by application to Huq Industries (https://huq.io/). Analysis of individual-level GPS trajectories was performed in a secure computing environment (https://www.ucl.ac.uk/isd/services/file-storage-sharing/data-safe-haven-dsh). Data and code supporting this manuscript are openly available via GitHub (https://github.com/hamishgibbs/huq_behaviour_clustering_paper).

Acknowledgements

The authors acknowledge the following funding sources: ESRC UBEL Doctoral Training Partnership (ESRC: ES/P000592/1) (HG). Retail Business Datasafe (ESRC: ES/L011840/1) (JC). RME was also supported by the National Institute for Health Research (NIHR) Health Protection Research Unit (HPRU) in Modelling and Health Economics, a partnership between UK HSA, Imperial College London, and LSHTM (grant number NIHR200908). The views expressed are those of the authors and not necessarily those of the UK Department of Health and Social Care (DHSC), NIHR, or UKHSA.

References

1. Jacobsen GD, Jacobsen KH. Statewide COVID-19 Stay-at-Home Orders and Population Mobility in the United States. World Med Health Policy. 2020;12(4):347–56.
2. Ballantyne P, Singleton A, Dolega L. Using unstable data from mobile phone applications to examine recent trajectories of retail centre recovery. Urban Inform. 2022 Dec 20;1(1):21.
3. Wesolowski A, Buckee CO, Engø-Monsen K, Metcalf CJE. Connecting Mobility to Infectious Diseases: The Promise and Limits of Mobile Phone Data. J Infect Dis. 2016 Dec 1;214(suppl_4):S414–20.
4. Tizzoni M, Bajardi P, Decuyper A, King GKK, Schneider CM, Blondel V, et al. On the Use of Human Mobility Proxies for Modeling Epidemics. PLOS Comput Biol. 2014 Jul 10;10(7):e1003716.
5. Perrotta D, Frias-Martinez E, Piontti AP y, Zhang Q, Luengo-Oroz M, Paolotti D, et al. Comparing sources of mobility for modelling the epidemic spread of Zika virus in Colombia. PLoS Negl Trop Dis. 2022 Jul 20;16(7):e0010565.
6. Panigutti C, Tizzoni M, Bajardi P, Smoreda Z, Colizza V. Assessing the use of mobile phone data to describe recurrent mobility patterns in spatial epidemic models. R Soc Open Sci. 2017 May 17;4(5):160950.
7. Zhang J, Zheng Y, Sun J, Qi D. Flow Prediction in Spatio-Temporal Networks Based on

Multitask Deep Learning. IEEE Trans Knowl Data Eng. 2020 Mar;32(3):468–78.
8. Bachir D, Khodabandelou G, Gauthier V, El Yacoubi M, Puchinger J. Inferring dynamic origin-destination flows by transport mode using mobile phone data. Transp Res Part C Emerg Technol. 2019 Apr 1;101:254–75.
9. Lu X, Bengtsson L, Holme P. Predictability of population displacement after the 2010 Haiti earthquake. Proc Natl Acad Sci. 2012 Jul 17;109(29):11576–81.
10. Bagrow JP, Wang D, Barabási AL. Collective Response of Human Populations to Large-Scale Emergencies. PLOS ONE. 2011 Mar 30;6(3):e17680.
11. Quadri C, Zignani M, Capra L, Gaito S, Rossi GP. Multidimensional Human Dynamics in Mobile Phone Communications. PLOS ONE. 2014 Jul 28;9(7):e103183.
12. Phithakkitnukoon S, Smoreda Z, Olivier P. Socio-Geography of Human Mobility: A Study Using Longitudinal Mobile Phone Data. PLOS ONE. 2012 Jun 28;7(6):e39253.
13. Kang Y, Gao S, Liang Y, Li M, Rao J, Kruse J. Multiscale dynamic human mobility flow dataset in the U.S. during the COVID-19 epidemic. Sci Data. 2020 Nov 12;7(1):390.
14. Apple. COVID□19 - Mobility Trends Reports [Internet]. 2021 [cited 2021 May 11]. Available from: https://www.apple.com/covid19/mobility
15. Google. Overview - Community Mobility Reports Help [Internet]. 2021 [cited 2023 Feb 10]. Available from: https://support.google.com/covid19-mobility/answer/9824897?hl=en&ref_topic=9822927
16. Maas P. Facebook Disaster Maps: Aggregate Insights for Crisis Response & Recovery. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining [Internet]. Anchorage AK USA: ACM; 2019 [cited 2022 Aug 1]. p. 3173–3173. Available from: https://dl.acm.org/doi/10.1145/3292500.3340412
17. Jiang S, Fiore GA, Yang Y, Ferreira J, Frazzoli E, González MC. A review of urban computing for mobile phone traces. Univ Web Domain [Internet]. 2013 Aug [cited 2023 Mar 25]; Available from: https://dspace.mit.edu/handle/1721.1/81784
18. Bonnel P, Fekih M, Smoreda Z. Origin-Destination estimation using mobile network probe data. Transp Res Procedia. 2018 Jan 1;32:69–81.
19. Buckee CO, Balsari S, Chan J, Crosas M, Dominici F, Gasser U, et al. Aggregated mobility data could help fight COVID-19. Science. 2020 Apr 10;368(6487):145–6.
20. Zipf GK. The P1 P2/D Hypothesis: On the Intercity Movement of Persons. Am Sociol Rev. 1946;11(6):677–86.
21. Brockmann D, Hufnagel L, Geisel T. The scaling laws of human travel. Nature. 2006 Jan;439(7075):462–5.
22. Song C, Koren T, Wang P, Barabási AL. Modelling the scaling properties of human mobility. Nat Phys. 2010 Oct;6(10):818–23.
23. Simini F, González MC, Maritan A, Barabási AL. A universal model for mobility and migration patterns. Nature. 2012 Apr;484(7392):96–100.
24. Alessandretti L, Aslak U, Lehmann S. The scales of human mobility. Nature. 2020 Nov;587(7834):402–7.
25. Alessandretti L, Lehmann S. Trip frequency is key ingredient in new law of human travel. Nature. 2021 May;593(7860):515–6.
26. Perc M. The Matthew effect in empirical data. J R Soc Interface. 2014 Sep 6;11(98):20140378.
27. Barabási AL. The origin of bursts and heavy tails in human dynamics. Nature. 2005 May;435(7039):207–11.
28. Barabási AL, Albert R. Emergence of Scaling in Random Networks. Science. 1999 Oct 15;286(5439):509–12.
29. Barabási AL, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T. Evolution of the social network of scientific collaborations. Phys Stat Mech Its Appl. 2002 Aug 15;311(3):590–614.
30. Hariharan R, Toyama K. Project Lachesis: Parsing and Modeling Location Histories. In:

Egenhofer MJ, Freksa C, Miller HJ, editors. Geographic Information Science. Berlin, Heidelberg: Springer; 2004. p. 106–24. (Lecture Notes in Computer Science).
31. Uber Technologies, Inc. H3 Hexagonal hierarchical geospatial indexing system [Internet]. 2022 [cited 2022 May 2]. Available from: https://h3geo.org/
32. Gale CG, Singleton AD, Bates AG, Longley PA. Creating the 2011 area classification for output areas (2011 OAC). J Spat Inf Sci. 2016;2016(12):1–27.
33. Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F, Barabási AL. Returners and explorers dichotomy in human mobility. Nat Commun. 2015 Sep 8;6(1):8166.
34. Yabe T, Moro E. Social Sciences Community. 2023 [cited 2023 Oct 18]. Behind the paper: COVID-19 reduced diverse urban interactions. Available from: http://socialsciences.nature.com/posts/covid-19-has-reduced-diverse-urban-interactions
35. Wesolowski A, Eagle N, Noor AM, Snow RW, Buckee CO. The impact of biases in mobile phone ownership on estimates of human mobility. J R Soc Interface. 2013 Apr 6;10(81):20120986.
36. Salganik M. Bit By Bit: Social Research in the Digital Age. Princeton University Press; 2017.
37. Consumer Data Research Center. Consumer Data Research Center Data Service User Guide [Internet]. 2023 [cited 2023 Oct 18]. Available from: https://data.cdrc.ac.uk/sites/default/files/D11%20CDRC%20Data%20User%20Guide.pdf
38. Wang F, Wang J, Cao J, Chen C, Ban X (Jeff). Extracting trips from multi-sourced data for mobility pattern analysis: An app-based data example. Transp Res Part C Emerg Technol. 2019 Aug 1;105:183–202.
39. Transportation Research Board, Institute of Medicine. Does the Built Environment Influence Physical Activity?: Examining the Evidence -- Special Report 282 [Internet]. Washington, D.C.: Transportation Research Board; 2005 [cited 2023 Jan 18]. Available from: https://www.nap.edu/catalog/11203
40. Garber MD, Labgold K, Kramer MR. On selection bias in comparison measures of smartphone-generated population mobility: an illustration of no-bias conditions with a commercial data source. Ann Epidemiol. 2022 Jun 1;70:16–22.
41. Yabe T, Bueno BGB, Dong X, Pentland A, Moro E. Behavioral changes during the COVID-19 pandemic decreased income diversity of urban encounters. Nat Commun. 2023 Apr 21;14(1):2310.
42. scikit-learn [Internet]. [cited 2023 Oct 20]. sklearn.cluster.KMeans. Available from: https://scikit-learn/stable/modules/generated/sklearn.cluster.KMeans.html
43. Arthur D, Vassilvitskii S. How slow is the k-means method? In: Proceedings of the twenty-second annual symposium on Computational geometry [Internet]. New York, NY, USA: Association for Computing Machinery; 2006 [cited 2023 Oct 20]. p. 144–53. (SCG '06). Available from: https://dl.acm.org/doi/10.1145/1137856.1137880
44. Fleischmann M. Clustergram: Visualization and diagnostics for cluster analysis. J Open Source Softw. 2023 Sep 2;8(89):5240.
45. Schonlau M. The Clustergram: A Graph for Visualizing Hierarchical and Nonhierarchical Cluster Analyses. Stata J. 2002 Dec 1;2(4):391–402.
46. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987 Nov 1;20:53–65.
47. Davies DL, Bouldin DW. A Cluster Separation Measure. IEEE Trans Pattern Anal Mach Intell. 1979 Apr;PAMI-1(2):224–7.
48. Caliński T, Harabasz J. A dendrite method for cluster analysis. Commun Stat. 1974 Jan 1;3(1):1–27.
49. Office for National Statistics. Output areas [Internet]. 2016 [cited 2022 Jul 27]. Available from: https://www.ons.gov.uk/census/2001censusandearlier/dataandproducts/outputgeography/outputareas