Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 82 (2025) 1824–1838

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2023 Montreal 17-21 July 2023

Disaggregate travel demand analysis using big data sources: unsupervised learning methods for data-driven trip purpose estimation

Pierluigi Coppola[a], Fulvio Silvestri[a]*, Francesco De Fabiis[a], Luca Barbierato[a]

[a]*Politecnico di Milano, Department of Mechanical Engineering, Via G. La Masa 1, Milano 20156, Italy*

Abstract

Data is a paramount factor in the success of transport modeling. Smartphones can be employed to retrieve full individual trajectories, either locally through apps using the devices' integrated GPS sensors, or through the mobile network operator (MNO), by tracing the mobile antennas to which devices connect over time. Several studies have demonstrated of the utility of this data to infer users' door-to-door trips, and then to build door-to-door origin-destination matrices, which are a key feedstock for transport modeling and planning. Some MNOs already provide such services commercially, yielding notable time and cost savings with respect to matrices estimated through traditional surveys. However, these are often highly aggregated and lack supplementary relevant trip information, such as mode and purpose, and if any, these are commonly obtained by means of human-driven heuristic considerations and fixed rules. This study aims at exploring the suitability of machine learning techniques for data-driven mobility demand estimation and analysis. It identifies associated opportunities and challenges through a pilot experiment focused on trip purpose estimation via diverse clustering techniques. Despite the experiment's limitations due to a small sample size and altered mobility patterns resulting from the COVID-19 pandemic, clustering algorithms (both distance- and density-based) successfully yield meaningful outcomes. The results include the identification of travel purposes, such as trips to home with or without overnight stays, trips to occasional destinations, commutes to work, trips to holiday stays, and more. These preliminary yet promising findings suggest that machine learning holds significant potential in mobility analysis, and it could feasibly be employed to estimate big-data-driven demand matrices, offering a higher degree of disaggregation and consequently enhancing the quality of transport modeling practices.

© 2024 The Authors. Published by ELSEVIER B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 16th World Conference on Transport Research

Keywords: individual mobility; travel behavior; machine learning; data mining; clustering.

* Corresponding author. Tel.: +39-02-2399-8498.
E-mail address: fulvio.silvestri@polimi.it

2352-1465 © 2024 The Authors. Published by ELSEVIER B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 16th World Conference on Transport Research
10.1016/j.trpro.2024.12.158

1. Introduction

Traditionally, data for transport modeling used to be collected through surveys, supplemented with auxiliary validation techniques like traffic and passenger counts. The former, which remain the primary data source in most cases, suffer from the need for active involvement of interviewees, leading to a time- and cost-intensive process. Consequently, analyses are seldom updated and are often conducted on a limited sample of individuals over the whole population. Nevertheless, over the past years, transport modelers and data analysts have increasingly recognized the potential of smart wearable devices, which passively collect location data. These devices ultimately contribute to the pool of big data. In particular, smartphones have emerged as optimal tools for tracking door-to-door trips since they are typically carried by most individuals at all times (Willumsen, 2021) (Anda et al., 2017).

Hence, it is not surprising that smartphones are already widely utilized by both MNOs and app data aggregators to estimate Origin-Destination (OD) matrices. This process involves collecting a time series of coordinates, essentially creating an individual's trajectory. MNOs achieve this by tracking the mobile cell to which a device is connected, while apps and app aggregators use the smartphone's integrated location service. These trajectories are then cleaned and analyzed to extract origin-destination trips for each user, which are aggregated to generate the final matrix. The primary advantage over traditional demand estimation methods lies in the continuous collection of data for a huge sample size (sometimes up to one-third of the population) with automated and parallelized processes. This significantly reduce costs and time while increasing the representativity of the sample. However, a key drawback is that data from mobile phones mainly consist of geographical information, lacking details about the travelers' socioeconomic characteristics, trip modes, or purposes. Addressing this limitation often involves heuristic approaches, relying on fixed rules to distinguish, for instance, between systematic and non-systematic trips. Nonetheless, a significant research gap exists concerning alternative methods to tackle this issue and to maintain a more data-driven approach in treating and analyzing such new big data for transport modeling, although machine learning and deep learning paradigms are already widely used in mobility analysis (Wismans et al., 2018) (Luca et al., 2020).

This study aims at bridging the gap in this underexplored research area through a pilot experiment, specifically focusing on trip purpose identification using unsupervised learning algorithms. The objective is to test the capability of different clustering algorithms in segmenting trips made by a group of volunteers (due to the absence of openly accessible datasets) in a manner that can be meaningfully interpreted as their estimated purposes. The central research questions revolve around the value of integrating more machine learning into travel demand modeling using novel big data sources and the challenges that may arise, along with potential approaches for addressing them.

1.1. Structure of the paper

The paper is structured as follows. Section 2 describes the state of art regarding the innovative data sources for transport modeling mentioned earlier, their current utilization in constructing OD matrices, and introduces the associated ethical considerations. Section 3 provides a detailed description of the proposed methodological approach. Section 4 presents an overview of the collected data and the analysis of the results provided by unsupervised learning algorithms. Finally, Section 5 concludes the paper by emphasizing the key findings and offering recommendations for further research directions.

2. Literature review

2.1. Individual mobility data sources

For the last 50 years, transport modelers have relied on surveys as the primary data source. However, these present a notable set of limitations due to the need of traveler involvement of travelers (active techniques), which makes the process highly expensive and time-consuming. This leads to constraints on the available sample size and the reference analysis period (usually an average day of a neutral month), as well as on the flexibility and updateability of the collected data. Furthermore, the quality of the collected data can be biased as it relies essentially on the trustworthiness of the interviewed travelers, who often alter or omit information (either intentionally or unintentionally) (Willumsen, 2021).

In contrast, these limitations do not apply to the many new passive data collection tools and techniques that are already available and operational today (though they do present some challenges). These are sources of big data for mobility, since they can offer a wide amount and variety of data, with eventually a nearly continuous data acquisition, tracking travelers or vehicles throughout their actual trips. Examples include mobile network data, mobile apps, public transport smart cards, vehicles' GPS sensors, Bluetooth and Wi-Fi hotspots. While these tools are not primarily dedicated to mobility data acquisition, they can provide extensive datasets without incurring unsustainable costs and complexity (for instance, concerning smart cards, the whole population of card holders can be the sample). This allows for collecting more data over time, and extending the analysis beyond the traditional average day of a neutral month, offering at the same time a superior degree of agility and flexibility (Willumsen, 2021) (Torre-Bastida et al., 2018) (Iqbal et al., 2014) (Tolouei et al., 2017) (Isaacman et al., 2011).

It is important to note that big data sources vary significantly in nature, and their potential use differs depending on the specific source. While smart cards and navigation systems are mode-focused (e.g., providing stop-to-stop or parking-to-parking information), hotspots provide only very punctual data, similar to traditional traffic counts. In contrast, mobile network and smartphone app data have proven capable of tracking individuals in their actual door-to-door trips. They collect, although in a different way, a sequence of user-labeled locations, forming user trajectories, which are then mined to extract the actual trips (Willumsen, 2021) (Torre-Bastida et al., 2018).

Trajectory points from mobile network are part of the records generated by the MNO's antennas (BTS - Base Transceiver Stations) for various reasons, and then stored and analyzed both for diagnostics and selling purposes. The main types of records collected by MNOs include Call Detail Records (CDR), Internet Protocol Detail Records (IPDR), Location Update (LU) signals and Handover (HO) signals. The first two are generated namely when a device makes or receives a phone call over the mobile network or uses internet traffic. Thanks to the spread of mobile internet usage in the past two decades, IPDRs can have a much higher temporal coverage and granularity, especially if the user has many connected apps that also refresh in background, such as web messaging or email services. LU and HO are signals used to control the network's traffic: LUs are sent by the BTS to the mobile phone to probe its location, usually every one or two hours; HO signals are instead triggered by the shift of mobile cell to which the mobile phone is connected while still communicating (Pourmoradnasseri et al., 2019) (Bonnel et al., 2018) (Chen et al., 2019). Merging these records, after removing type-specific information (e.g. call duration in case of a CDR), results in data points with a timestamp, a device identifier, and a BTS identifier. The former is a random ID that is uniquely assigned to the SIM card for (usually) a month; after such period, SIM IDs are reassigned to protect subscribers' privacy. BTSs are instead identified by a pair of codes, the Location Area Code (LAC) and the Cell ID; the former indicates the roughly hexagonal spatial area served by the BTS (cell), whereas the latter indicates the neighborhood (Location Area) of cells to which it belongs (Chen et al., 2019). The location of the BTS, which is known by the MNO, represents an estimate of the user's location at the recorded timestamp; therefore it represents the spatial information of the trajectory point. The spatial granularity of this data depends on the mobile cell's radius, which can range from a few kilometers (2G/3G cells) in rural areas to around a hundred meters in urban environments (4G/5G cells) (Bonnel et al., 2018) (Huang et al., 2018) (Imai et al., 2021).

On the other hand, apps with permissions to access the device's location services use its (assisted) GPS sensor to geographically trace position with accuracy ranging from 10 to 50 meters. Since a single app rarely provides sufficient time coverage, most label their data with the device's unique marketing ID. App data aggregators gather data from different apps, grouping it by marketing ID (thus by smartphone user). In the final dataset, all data points provided by apps using the device's location service contribute to create its trajectory. The main advantage over mobile network data is that spatial information represents the actual measurement of the user's location, rather than a proxy (the antenna's location). However, despite the role of aggregators, datasets based on device location may still have long gaps in time coverage. Moreover, after the introduction of Apple's App Tracking Transparency, iOS users can choose to hide their marketing ID, preventing the sharing of data with aggregators.

2.2. Mobile data mining and Origin-Destination matrices

Mobile phone data, including both mobile network data and smartphone app data, can be mined to extract door-to-door trips and, consequently, OD matrices. Several examples and good practice recommendations are already available (Willumsen, 2021) (Iqbal et al., 2014) (Catapult, 2017) (Scholl et al., 2019).

First, we need to consider and compensate for all the biases and errors it may contain. For mobile network data, these biases include, for instance, cell jumps due to the MNO's traffic balancing policy – in these cases, the device alternates between two close antennas depending on how many users are connected to each one. For smartphone app data, instead, we must consider that GPS-retrieved locations are continuous and affected by noise, meaning an actual place corresponds to a range of locations, rather than a single point; otherwise, the user would always be recorded as moving (Iqbal et al., 2014) (Pourmoradnasseri et al., 2019) (Tolouei et al., 2017) (Jiang et al., 2017) (Toole et al., 2015) (Alexander et al., 2015).

Once this data cleansing phase is completed, mining typically relies on identifying each user's "dwells", which are times when their position is stationary. Consequently, trips are the events occurring between subsequent dwells. This requires quantitatively defining what constitutes a dwell, meaning how long a user should stay in a location to say that they are actually doing something there and not, for instance, just stuck in traffic, and which uncertainty radius should we consider, especially when dealing with GPS data. ML can be applied here if, instead of thresholds, we use spatial or frequency clustering, but this is not a common situation (Anda et al., 2017) (Wismans et al., 2018) (Bonnel et al., 2018) (Friso et al., 2018) (Calabrese et al., 2011).

If the study is conducted on a zoned study area, trips can be aggregated by origin-destination pairs to create an OD matrix, which can then be scaled up to the whole population. This final phase also includes the removal of non-moving devices, which may be IoT devices rather than actual smartphones. Many MNOs and app aggregators already offer this service commercially, sometimes providing matrices split by estimated mode or purpose. However, the latter splits are typically obtained through heuristic considerations (for modal split, instead we can find a literature about more advanced techniques), commonly distinguishing only systematic and non-systematic trips using fixed rules, such as how many times per week a user should undertake a trip to classify it as systematic or not (Willumsen, 2021) (Wismans et al., 2018) (Bassolas et al., 2019).

In this entire process, we generally observe that few artificial intelligence, if any, is employed, leaving the entire burden on the analysts' choices. Conversely, a substantial body of literature exists concerning machine learning, and even deep learning, in the analysis of mobility. These techniques have been applied, for example, for next-location prediction, trajectory generation, mode identification and traffic flow prediction (Luca et al., 2020) (Iqbal et al., 2014) (Chen et al., 2019) (Akhtar and Moridpour, 2021) (Hagenauer and Helbich, 2017) (Liang and Wang, 2017) (Zhou et al., 2018) (Assi et al., 2018). However, such techniques could offer significant advantages over manually crafted rules. Therefore, we believe it is worth investigating the possibility of extending machine learning to this aspect of mobility analysis, which has not yet received sufficient attention in research.

2.3. Privacy concerns and protection

Finally, we would like to address some privacy concerns related to the utilization of big data in individual mobility analysis. The right to privacy is a fundamental ethical and legal value that cannot be overlooked, as emphasized by the European Union's General Data Protection Regulation (GDPR). Therefore, safeguarding privacy is an essential requirement for any technological process or product and represents a significant source of added value. However, it also presents considerable challenges, as the necessity to analyze an individual's geo-referenced trips inevitably intersects with their privacy (Catapult, 2017) (Chen et al., 2019) (De Montjoye et al., 2018). This is why datasets are always anonymized (meaning the ID associated with a SIM card changes every month). Moreover, particularly concerning MNOs', these datasets are not shared with third parties, ensuring that the entire data mining process remains in house. Consequently, any commercialized outputs are consistently aggregated, eventually masking OD pairs with fewer than a minimum number of travelers (Willumsen, 2021) (Wismans et al., 2018) (Friso et al., 2018).

3. Methodological approach

Due to the privacy limitations mentioned earlier, we were unable to access real raw data collected by MNOs or app aggregators. Hence, we opted to conduct a pilot experiment to emulate such datasets with the assistance of some volunteers and a dedicated smartphone app. Once datasets were created, the data was processed as in the state-of-art applications before being input into the machine learning algorithms. Therefore, the methodology comprises the main phases illustrated in Figure 1.

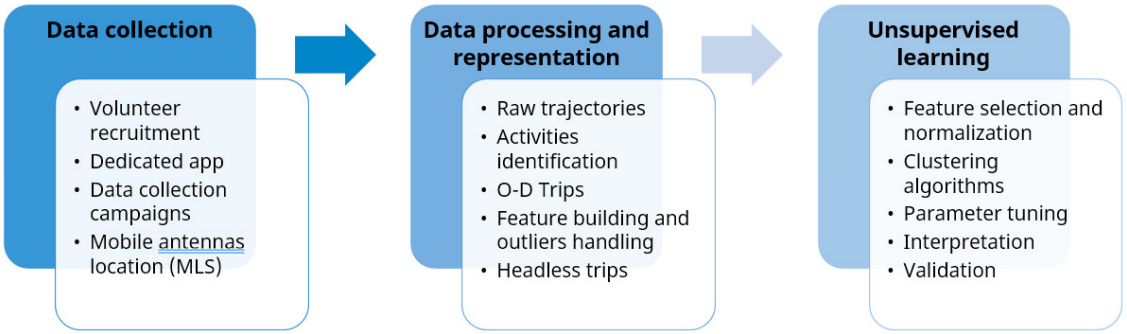

Figure 1 - Main phases of the proposed methodological approach.

3.1. Data collection

All volunteers who participated in the experiment were requested to install a data collection app on their Android smartphone and activate background location services. The app's interface primarily featured a start/stop button that volunteers would press after any device reboot to start data collection. It was then periodically querying Android APIs in the background to get a set of instantaneous information and generate a record. The key attributes included a unique user ID (given by the app itself) and a timestamp, along with GPS coordinates to create a smartphone app dataset and connected mobile cell identifiers (LAC and Cell ID) to replicate a mobile network data. During the experiment, the data acquisition frequency was set to record data every three to five minutes. These records were compiled into a dataset and then uploaded to the cloud every four hours or whenever the app was restarted; this constituted our raw data.

Data points, being spatially and temporally referenced, formed the user's trajectories. In particular, when using GPS coordinates, we obtained an actual app location-based dataset, whereas by utilizing the mobile cell's location, we had a MNO-like dataset. In the latter case, we first needed to replace mobile cell identifiers with the corresponding antenna's coordinates. To accomplish this, we relied on Mozilla Location Service (MLS), which is an open, community-driven dataset containing estimates regarding the location of mobile cells, albeit with limited precision. For cells not listed in the MLS, we inferred the antenna's location by averaging the GPS coordinates recorded while connected to that specific cell, which is essentially how the MLS dataset itself is updated.

3.2. Processing and representation

Any form of machine learning or data analytics, in general, necessitates the proper treatment and processing of raw data to attain a format with meaningful and extractable information content. Clearly, trajectories embed a lot of information, but they are hardly usable out-of-the-box when our goal is to estimate a trip's purpose. At the very least, we require a dataset of trips, and to achieve this, we rely on the well-consolidated techniques described in Section 2.2. Throughout this process, data exploration plays a crucial role, encompassing both univariate and multivariate analysis, including an examination of attribute correlations.

The very first step is however to select the relevant features mentioned earlier and merging them with the MLS dataset. This results in two distinct datasets: one emulating mobile network data (Mobile dataset), and the other resembling data typically aggregated by a smartphone app data aggregator (GPS dataset). Both datasets share identical

columns, namely user ID, timestamp, latitude, and longitude. The key distinction lies in the fact that the latter two attributes represent the coordinates of the connected mobile network antenna in one case and the actual GPS-retrieved location of the smartphone itself in the other. Subsequently, the two datasets undergo equivalent treatment but independently from each other.

In the second phase, each user's trips are extracted following the common approach of first identifying dwells (see Figure 2), fixing a minimum stationary time for users, above which we assume that they are performing an activity. Especially when dealing with GPS data, we must also establish an uncertainty radius for considering a set of points as stationary and for identifying recurrent locations in a user's mobility. In fact, for the experiment this was also applied while treating the Mobile dataset, due to the uncertainty in the antennas location. After conducting several tests and validation with volunteers, we set these two parameters to 15 minutes and 1 km, respectively. Additionally, we imposed a maximum threshold of 12 hours for the temporal gap between two neighboring trajectory points, above which we consider the user's dataset broken in two independent sub-datasets, resetting the user's location without identifying a trip. The reason for choosing such value is that volunteers may likely power off their device for the night; a longer time could most probably be caused by a volunteer who forgot to start the collection app, therefore they may have several undetected trips. Our algorithm searches for dwells and trips within each sub-dataset of each user by applying the 15 minutes and 1km thresholds. It is important to note that all points of a dwell must fall within the radius of 1km from their centroid; the search process employs a binary approach to reduce execution time.

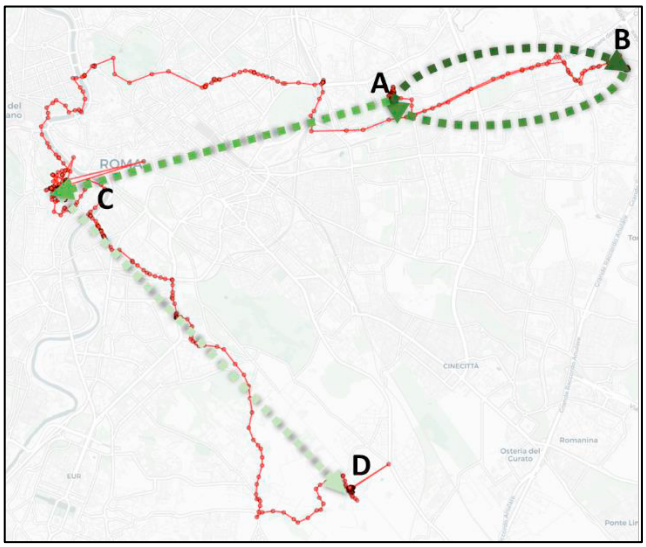

Figure 2 - Example of headless trips extraction from GPS trajectories (1st trip from A to B, 2nd trip from B to A, 3rd trip from A to C, 4th trip from C to D).

Thirdly, the trips mined from each sub-dataset of the same user are grouped, so to identify their common locations (within 1 km). Such passage is fundamental: since different individuals may have extremely different mobility habits, we are not directly interested in the origin-destination trips, but rather in headless trips, in which the endpoints are replaced by features representing the relationship between users and locations. These features include the user's trip frequency to destination D, expressed as trips to D per day of collection *(Freq-abs)*, per total trips collected by that user *(Freq-rel)*, or the range-normalization of this last one among the user's destinations *(Freq-norm)*,and the fraction of the user's non-travelling data collection time spent at location D *(%NTT@D)*, or its range normalization *(%NTT-norm)*. In the end, each headless trip is characterized by the following features: arrival time at destination, expressed as a 24-hour sine and cosine term *(Time-sin* and *Time-cos)* to account for its periodicity; travel time *(TT)*; trip length as the O-D direct distance *(Dist)*; duration of the following activity *(AT)*; and the aforementioned features related to its destination D. Consequently, the final outcome comprises two datasets of headless trips, allowing us to assess the effectiveness of clustering techniques on both types of extensive mobility data.

Unsupervised learning algorithms ought to be tested on each headless trips dataset with different combinations of attributes used as input. Specifically, trip length and duration are excluded from consideration as they hold relatively low significance in estimating the purpose of a trip. Users may cover varying distances and durations, particularly when commuting to work, and, in the collected data, these attributes exhibited almost zero correlation with other features. Likewise, Freq-abs among the frequency terms is removed from consideration. In all candidate attribute combinations, we retain time, activity duration, one attribute among Freq-rel and Freq-norm, and optionally, one attribute among %NTT@D and %NTT-norm. This results in six attribute combinations, which then expand to twelve because we test two different types of attribute normalization: range scaling and standard scaling. In all cases, attributes undergo winsorization at the 5th/95th percentile to address outliers. This operation does not apply to the sine and cosine representations of time.

3.3. Unsupervised learning

In our study, we tested two different clustering techniques: hierarchical clustering and DBSCAN. In addition to the combination of input attributes/normalization type, these algorithms also require the setup of specific input parameters. These choices represent the crux of the unsupervised learning phase.

Hierarchical clustering is a versatile technique that can be employed as an initial test before selecting other algorithms. It has the capability to provide a full breakdown of the data's similarities, allowing for the creation of any number of clusters ranging from one to the total number of data points. In its agglomerative version (see Table 1), it starts with each data point constituting an individual cluster. It then progressively selects pairs of clusters to merge into a single cluster until only one cluster, encompassing the entire dataset, remains. These merges are regulated by a distance metric for comparing two data points and a linkage criterion. In our case, we utilize Euclidian distance and Ward's method. The sole parameter in this case is the number of clusters, which essentially determines the outcome among the various results produced by hierarchical clustering. On the other hand, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based algorithm that clusters together points based on whether they are sufficiently close together in a densely populated region of the multidimensional feature space. The algorithm relies on two specific parameters, minPts (an integer) and epsilon (a decimal number) (see Table 1). A distance metric, in our case the Euclidian distance, is required as well. Points that do not belong to any cluster because they do not reside in a region of sufficient density are labeled as noise. It is important to note that setting epsilon too low or minPts too high may result in labeling the entire dataset as noise.

Table 1 – Pseudocode of the adopted clustering techniques: Hierarchical clustering (left side) and DBSCAN (right side).

```
HierarchicalClustering(Dataset):
    #Compute Euclidean distances
    clusters = initial_clusters(Dataset)
    #Compute Euclidean distances
    d = compute_pairwise_distances(Dataset)

    while len(clusters) > 1:
        min_merge_cost = infinity
        clusters_to_merge = ( )

        for i in range(len(clusters)):
            for j in range(i + 1, len(clusters)):
                merge_cost = ward_merge_cost(clusters[i], clusters[j], d)

                if merge_cost < min_merge_cost:
                    min_merge_cost = merge_cost
                    clusters_to_merge = (clusters[i], clusters[j])

        merged_cluster = merge_clusters(clusters_to_merge[0],
clusters_to_merge[1])
        clusters.remove(clusters_to_merge[0])
        clusters.remove(clusters_to_merge[1])
        clusters.append(merged_cluster)

    return clusters
```

```
DBSCAN(Eps, MinPts, DataSet):
    #List to store resulting clusters
    clusters = [ ]  # List to store resulting clusters
    #List to keep track of visited points
    visited_points = [ ]

    for each point P in DataSet:
        if P is already visited:
            continue

        visited_points.append(P)
        #Find points within Eps distance
        NeighborPts = RangeQuery(P, Eps, DataSet)

        if len(NeighborPts) < MinPts:
            P is noise
        else:
            C = NewCluster()
            ExpandCluster(P, NeighborPts, C, Eps, MinPts, DataSet)

            #Add the cluster to the list of clusters
            clusters.append(C)

return clusters
```

Since clustering falls within the machine learning's paradigm of unsupervised learning, the validation of the results involves assessing the quality and meaningfulness of the discovered patterns. Some common internal quality measures and techniques for validation includes:

- Silhouette score. This metric quantifies how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates well-defined clusters.
- Elbow method. In this technique, the within-cluster sum of squares is plotted against the number of clusters. The "elbow point" represents the optimal number of clusters, where adding more does not significantly reduce the sum of squares.
- Visual inspection. Tools like scatter plots, dendrogram trees, and t-SNE are utilized to visually assess the separation and coherence of clusters.

We thus preferred to run multiple clustering instances (see Figure 3), with different algorithm parameters and feature selections, rank them according to the internal measure, and then individually inspect the obtained clustering configurations, to determine if these are meaningfully interpretable for our score.

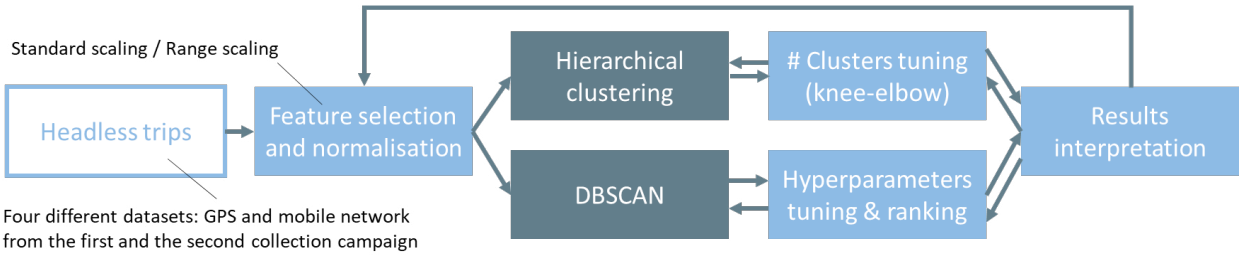

Figure 3 – Main steps of the iterative loop in the unsupervised learning phase.

As regards hierarchical clustering, we can test it on all the input features options, with 2 to 6 clusters, for a total of 60 output configurations per dataset; these can be individually inspected with the support of knee-elbow plots to find the most meaningful ones. On the other hand, selecting the input parameters for DBSCAN is more challenging. Therefore, we heavily rely on the ranking determined by the silhouette score. In particular, we launch an optimization framework for each input combination, aiming at maximizing the silhouette score. We then export all the results that have at least two actual clusters (in addition to a third cluster representing noise). Subsequently, we analyze these results individually, starting from the high-silhouette ones. Notably, we conduct 100 clustering attempts for each input combination, although in the experiment, approximately one-third of these attempts were discarded because they consisted solely of noise or noise and one cluster.

4. Results

We conducted data collection and analysis in two separate campaigns, both with the active participation of 10 volunteers. During these campaigns, we collected and processed both mobile and GPS data, resulting in a total of four headless trips datasets. Afterwards, volunteers were invited to validate the obtained trips. Additionally, they had the option to specify their home and work locations, and we are grateful that all of them chose to provide this information because it allowed us to validate the results of our cluster interpretations ex-post. This section presents an overview of the collected data and the results of unsupervised learning on the four datasets.

4.1. Collected data

Please note that in both campaigns, we did not focus our analysis on a specific study area, as our ten volunteers were geographically distributed in different parts of Italy. Anyway, this ought to be acceptable since we are primarily interested in their behavior, regardless of the geographical location of their trips origins and destinations.

The first campaign was held in July 2021, and we managed to collect 290 (GPS) and 254 (mobile cell) trips. There was a significant skewness in the users' actual collection days, with more than half of volunteers recording data for at

most 10 days. The widespread adoption of remote working caused the final dataset to present a bias, with very few trips to locations validated as workplaces by volunteers. Figure 4 provides an overview of the collected data.

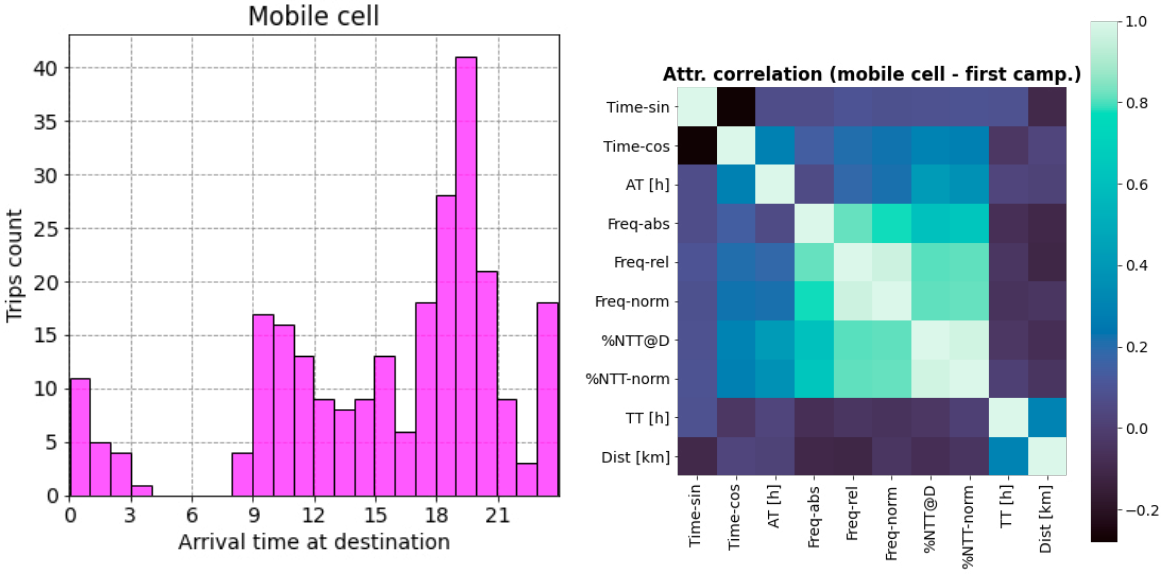

Figure 4 - Summary of the first collection campaign's mobile dataset, in particular the hourly trip distribution of recorded trips (on the left) and Spearman attributes correlation plot (on the right). Values for the GPS dataset are analogous.

The second campaign was took place between October and November 2021, during a more neutral period in terms of remote working, as many people had returned to working or studying in person. In this case, we were able to collect 404 (GPS) and 419 (mobile) trips. All users collected data for a period ranging from 19 to 25 days. Figure 5 provides an overview of the collected data.

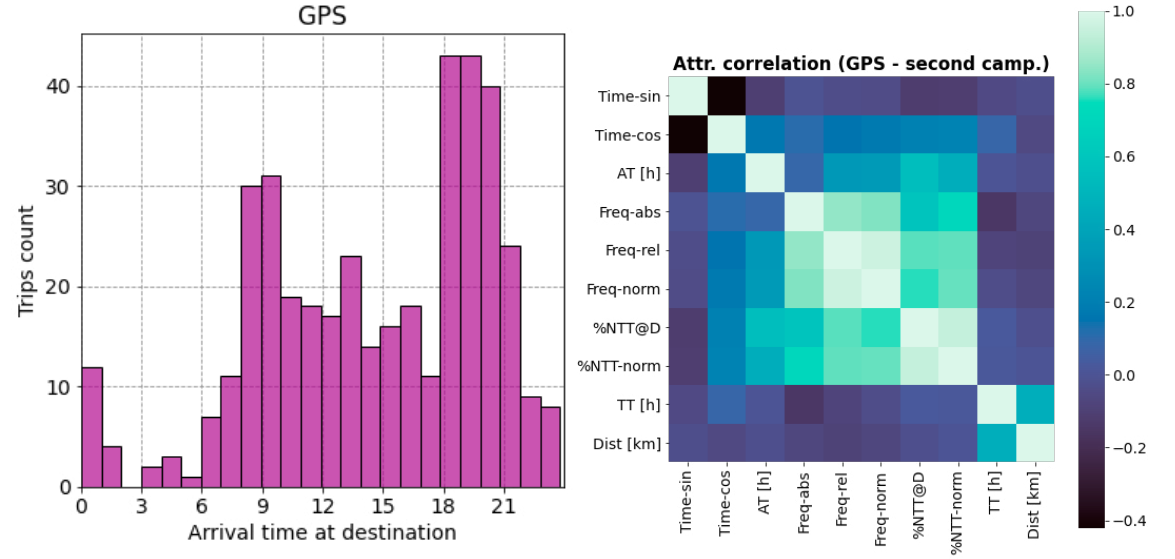

Figure 5 - Summary of the second collection campaign's GPS dataset, in particular the hourly trip distribution of recorded trips (on the left) and Spearman attributes correlation plot (on the right). Values for the mobile dataset are analogous.

4.2. Clustering results

Overall, we tested hierarchical clustering and DBSCAN on four datasets, that are those of GPS and mobile cell trips for each of the two campaigns. Additionally, in all those cases we tried to either include or exclude the term describing the percentage of time spent by the user at destination (%*NTT*, either range-normalized for the user or not), while we always kept time (*Time-sin* and *Time-cos*), activity duration (*AT*) and a relative trip frequency to destination (*Freq*, either range-normalized for the user or not). In general, not including %*NTT* tends to create more time-related splits and makes the results less relevant, therefore we discarded all configurations obtained without %*NTT*. For all the remaining options, we selected the best interpretable results, as discussed in the previous section; such configurations are summarized by Table 2. The final clustering results under these different scenarios capture various aspects of trips (data points). By comparing them, we can identify the strengths and limitations of the algorithms under examination, as well as those of the collected data, which, as mentioned earlier, has some limitations and biases.

Table 2 - Input options and cluster count for the best obtained clustering configurations.

Algorithm	Camp.	Data	Freq. norm.	%NTT norm.	Scaling	#Clusters
Hierarchical	**1st**	**GPS**	Yes	Yes	Std	5
		Mobile	No	No	Std	6
	2nd	**GPS**	No	Yes	Std	6
		Mobile	No	Yes	Std	6
DBSCAN	**1st**	**GPS**	No	Yes	Range	4
		Mobile	No	Yes	Std	5
	2nd	**GPS**	Yes	No	Range	6
		Mobile	No	Yes	Range	4

Although different datasets and algorithms yielded varying results, they generally follow a relatively similar pattern. This similarity arises from the fact that they are all computed on the same type of data, namely individual mobility, which was partially collected from volunteers with similar behaviors (across the two campaigns), or even the same set of volunteers (within a campaign, for GPS and mobile data). We can summarize the common patterns identified across different cases as follows:

- *Occasional trips*: These trips are characterized by a low frequency of visits to the same destination and a low percentage of time spent at the destination (%NTT). They represent non-systematic trips made by volunteers, typically during the daytime. Clustering algorithms consistently identify these trips (sometimes as sub-clusters) due to their high density.
- *Trips to home*: These are trips where home is the most frequently visited location and has the highest normalized %NTT, aligning with the volunteers' stated home location.
- *Trips to home with an overnight stay*: These trips, among those to home, are followed by extended periods of activity and typically occur between the afternoon and night. Identifying them requires distinguishing them from other home trips and avoiding labeling them as noise in the case of DBSCAN.
- *Trips to home with a multi-day stay*: Similar to the previous category, these trips involve even longer activity durations spanning multiple days. This is common among remote workers or students who spend several days at home or in the immediate vicinity without taking trips detectable by our processing algorithm. They may be clustered with other trips to home with overnight stays or treated as noise if none of these is present.
- *Trips to other places with an overnight stay*: These trips resemble trips to home in terms of activity duration and time distribution but are headed to locations with low frequency and %NTT. This might suggest users are spending the night at a second home, a holiday resort (especially during the first campaign held in July), or a friend's or partner's place. Due to their lower density relative to other trip types, DBSCAN, in particular, may categorize them as noise.

- *Trips to work*: These trips are characterized by the second-highest frequency and %NTT, a duration of about ten hours, and temporal concentration in the morning and early afternoon, with the latter associated with lower activity duration. Work location was also validated by users. However, in the first campaign, which took place when the majority of volunteers were not working or studying in person, this cluster is essentially unidentifiable.

For illustrative purposes, Figure 6 provides a visual representation of one of the clustering results, highlighting the isolated groups. Table 3, on the other hand, summarizes which of these groups are isolated in the different cases, considering the best parameters and input features configuration. Please, note that by *isolating* we mean that such trips are grouped in one or more clusters and separated from all the other data points; indeed, there is frequently a further split by time for trips with the same interpretation (e.g. trips to home with a short stay in the morning and those in the afternoon are often in two different clusters).

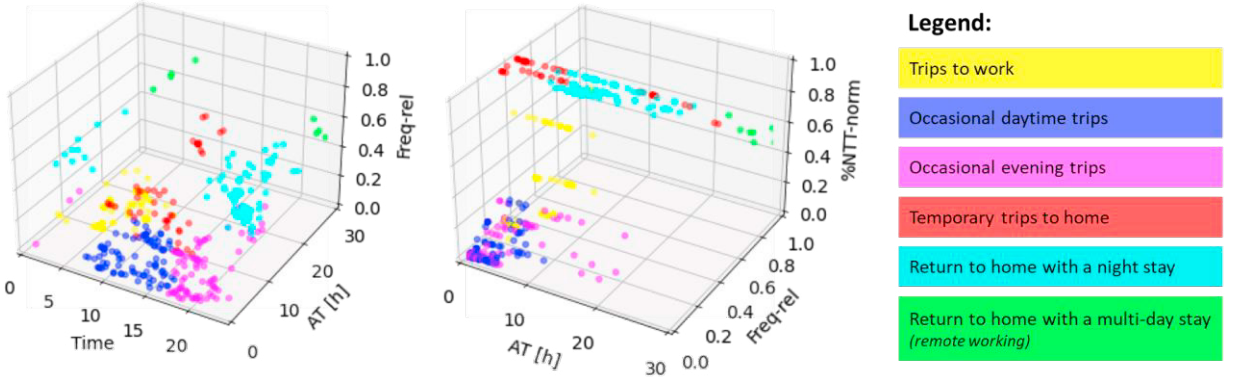

Figure 6 - Example of scatterplots displaying hierarchical clustering configuration (6 clusters) for the second campaign's GPS data.

Table 3 - Summary of clustering results. The integer number indicates how many clusters cover the corresponding group; empty values mean that the algorithm is not able to isolate such group.

Algorithm	Camp.	Data	Occasional trips ground	Trips to home	Home with night stay	Home/multi-day stay	Other places/ night stay	Work trips
Hierarchical	1st	GPS	2	1	1		1	
		Mobile	2	2	1		1	
	2nd	GPS	2	1	1	1		1
		Mobile	3	2		1		
DBSCAN	1st	GPS	1	1	1			
		Mobile	1	1	1	1		
	2nd	GPS	2		1			1
		Mobile	1	1	1			

In all four datasets, both hierarchical clustering and DBSCAN successfully distinguish trips to home from those to occasional destinations. Hierarchical clustering also manages to identify multi-day stays at home, separating them from the trips mentioned earlier in the second campaign's data. Similarly, it accomplishes this for trips to other locations followed by an overnight stay in the first campaign (i.e. to a hotel, as this data was collected during the holyday season). DBSCAN, on the other hand, tends to create fewer splits within the same group of trips.

Both algorithms are generally effective in identifying trips to work in the GPS dataset from the second campaign. However, they fail on mobile data, primarily due to specific circumstances related to some volunteers for whom some mobile cells returned phantom trips, because of both the inaccuracy of the MLS dataset and partial unsuitability of

our trip mining algorithm for mobile network data; i.e. the parameters of 1 km and 15 minutes are too binding for cells, although they are good for GPS data. Nevertheless, this last limitation is specific of the pilot experiment and could be easily overcome by MNOs while treating their actual data.

5. Discussion and conclusions

As described in the previous section, almost all clustering options allow to separate trips to high-frequency destinations (i.e. to home) and the others, that compose the ground of occasional trips; in some cases also other purpose trips (e.g. to work) are identified. Results are similar if we use either GPS data or mobile data; though this is partly due to the way mobile antennas' locations are estimated if not available in MLS dataset. Nevertheless, in general, we can expect trip purpose to be more linked to the time domain rather than the space domain, therefore the important point is to identify the common places visited by a user, the time spent there and the trip frequency. This is confirmed by the relevance that these features have in this experiment's results, since they are the key features of headless trips (e.g. omitting %*NTT* is detrimental for the final clustering configuration).

In fact, the tested algorithms managed to find meaningful results, despite the limitation and biases of the data that we have collected. This suggests us that, if wider and more consistent datasets was given, such as those already employed to build OD matrices, trip purpose identification could be actually accomplished with machine learning techniques. Of course, this is a pilot study, and thus deeper and wider research is still needed, to find the best way to address all the related criticalities, and to look for data integration to provide better attributes and more competitive solution.

Overall, we are led to believe that machine learning in transport modeling has impressive potential, which is definitely worthy of further studies and developments. In particular, disaggregate travel demand analysis could provide invaluable data for transportation planners and policymakers, especially when developing Activity-Based Models (ABMs). ABMs are advanced tools that simulate individual travel behavior by modeling each person's activities, trips, and mode choices. The outcomes of disaggregate travel demand analysis can serve as a rich source of calibration data for ABMs, especially with respect to activity scheduling (i.e. providing information on how individuals schedule activities throughout their day), trip generation and distribution (i.e. capturing the complexity of spatial travel patterns), and mode choice modeling (i.e. accounting for factors such as trip purpose, allow for a comprehensive understanding of mode choice decisions).

As a source of calibration data, machine learning can enhance the ABMs predictive capabilities, making them more reliable for simulating travel behaviors and predict the impacts of transportation policies accurately.

5.1. Encountered criticalities

During this study, we had to address several challenges related to the necessity of collecting our own data. This was because there were no real, raw, open smartphone app data or mobile network data available due to privacy restrictions. Hence, we developed a dedicated app in order to emulate the datasets, with volunteers recording their trips for several days. Moreover, it is worth noting that this data collection occurred during a period characterized by altered mobility habits due to the COVID-19 pandemic. The low number of volunteers, along with the resulting limited representativeness of the population, introduced biases and incompleteness in our dataset, as it does not capture a wide range of behaviors. Furthermore, in the case of mobile cell data, our processing was constrained due to the unavailability of precise antennas' locations. This limitation would not be an issue if a similar analysis were conducted by a mobile network operator.

However, estimating trip purposes through unsupervised learning faces challenges that extend beyond the scope of this particular experiment and are likely to arise with any provided dataset. Importantly, we lack a known ground truth against which we can compare clustering results when tuning parameters and selecting input combinations. In our study, *Freq-rel* and %*NTT-norm* appeared to be the most effective attributes in most cases, with standard scaling working better for hierarchical clustering and range scaling for DBSCAN. However, such selections should always be validated through rigorous checks on the meaningfulness of the returned results. Moreover, the transferability of parameters is limited, as the performance of clustering algorithms is highly dependent on the specific dataset. For example, the choice of DBSCAN's *epsilon* and *minPts* depends on the density of data points, which means that

datasets with different sizes require different settings to achieve similar results. This highlights that, despite the use of unsupervised clustering algorithms, trip purpose estimation still requires active oversight and control from analysts to ensure the reliability of the model's outputs.

5.2. Applicability and scalability

Unsupervised learning for analyzing individual mobility holds significant potential. Even within the limitations and biases of our datasets, clustering algorithms effectively segmented trips into meaningful categories. These categories included trips to home, with distinctions between temporary and longer (overnight) stays, and, in the case of the least biased dataset (GPS trips from the second campaign), trips to work. With a larger and more diverse dataset, possibly enriched with additional data related to trip origins and destinations, these algorithms are expected to perform even better in estimating trip purposes.

In the context of transport modeling, clustering for trip purpose estimation could play a pivotal role, ultimately providing OD matrices categorized by trip purpose. This task could be carried out by mobile network operators, app aggregators or affiliated data analytics companies, as they typically estimate OD matrices for commercial purposes. The clustering techniques we presented operate on individual trips derived from precise data processing, which is often inaccessible due to privacy concerns (as it can be overly disaggregated). However, OD matrices and trip purpose estimation would allow mobile network operators and smartphone app data aggregators to generate valuable mobility statistics for a study area and, with advanced analytics, segment travelers.

The advantage of employing machine learning, as opposed to simplistic threshold-based and rule-based techniques currently in use, lies in the ability to learn trip partitioning directly from the data. This reduces the reliance on analyst-driven decisions, although complete elimination of human intervention is unlikely. As demonstrated in this experiment, analysts are still required to inspect and interpret clustering results in order to choose the best input features, normalization methods, eventual parameters, and the clustering algorithm itself.

It is important to note that algorithm parameters must be tailored to the specific dataset. For instance, the choice of DBSCAN's *epsilon* and *minPts* parameters depends on the data point density, requiring different settings for datasets of varying sizes to achieve similar results. Consequently, while clustering algorithms are employed for trip purpose estimation, analysts remain central in ensuring the reliability of model outputs. Nevertheless, this produces the need of manual inspection and interpretation of the clustering configurations' meaning, that however collides with the need to protect users' privacy by aggregating and anonymizing data. The latter is a crucial element, in ethical terms for any kind of data, and in legal terms mainly for MNO data.

In terms of privacy protection, several anonymization approaches can be considered, particularly for datasets with a substantial number of users. One approach involves removing user ID labels from headless trips, adding noise to certain features susceptible to backtracking, and examining the resulting data. Another option is to put the whole analysis behind a firewall, that does only allow to get a summary of the obtained features, such as a kernel density estimate of its distribution on the selected attributes, or an actual partitioning of the features space obtained by means of a classifier which assigns points to the nearest cluster; the latter could be, for instance, a k-NN classifier. Then, the interpretation would be made on the space partitioning or the summary of clusters, so to attribute each one a potential purpose. In either case, trips are aggregated within clusters, enabling the creation of OD matrices for each cluster with separate interpretations. Mobile network operators or app data aggregators can then sell these matrices, provided they meet the minimum user count requirement and include cluster interpretations.

5.3. Further research perspectives

As ours was a preliminary pilot experiment, we acknowledge that several steps ought still to be taken. Firstly, additional experiments on larger and more diverse datasets are needed to validate the scalability of the modeling framework. This is crucial due to the high dependence of clustering results on the input dataset. Secondly, testing the strategies discussed for ensuring anonymization is essential, as their applicability is vital for the practical and commercial success of clustering for trip purpose estimation, especially concerning OD matrices commercialized by mobile network operators. Thirdly, the development of a unified framework to identify the best clustering results, although challenging, would be highly beneficial.

In addition to these aspects directly linked to the technique's applicability, several other ancillary points and ideas merit exploration to enhance the quality of results for transport modeling:

- Testing other clustering algorithms, such as k-Means, Expectation Maximization algorithm, etc.
- Combination of clustering and classification and cross-test on different datasets.
- Locations clustering, still based on the time spent there by users, trip frequency and so on.
- Traveler clustering and user segmentation, e.g. to identify workers, remote workers, students, etc.
- Fusion with external data, such as information about the trip's origin and destination, for instance the nearby points of interest.

Finally, this research has successfully identified not only work-related and return to home trips but also occasional trips. However, it has been challenging to pinpoint the precise reasons behind the occasional journeys, such as those for accompanying, shopping, healthcare, and leisure. This level of granularity would require incorporating additional data sources, such as points of interest, which provide information about activities in the vicinity of a destination, including stores, hospitals, gyms, etc.

The inability to precisely determine the purpose of occasional trips underscores the importance of integrating supplementary data to enrich our understanding of travel behavior. Further research involving points of interest data can significantly contribute to this endeavor by shedding light on the activities individuals engage in during their journeys. This, in turn, can provide valuable insights for transportation planning, urban development, and policymaking, as it allows for a more comprehensive analysis of travel demand and the factors influencing it.

References

Akhtar M. and Moridpour S., A Review of Traffic Congestion Prediction Using Artificial Intelligence, J. Adv. Transp., vol. 2021, pp. 1–18, doi: 10.1155/2021/8878011, Jan. 2021

Alexander L., Jiang S., Murga M., and González M. C., Origin–destination trips by purpose and time of day inferred from mobile phone data, Transp. Res. Part C Emerg. Technol., vol. 58, pp. 240–250, doi: 10.1016/j.trc.2015.02.018, Sep. 2015

Anda C., Erath A., and Fourie P. J., Transport modelling in the age of big data, Int. J. Urban Sci., vol. 21, no. sup1, pp. 19–42, doi: 10.1080/12265934.2017.1281150, Aug. 2017

Assi K. J., Shafiullah M., Nahiduzzaman K. M., and Mansoor U., Travel-To-School Mode Choice Modelling Employing Artificial Intelligence Techniques: A Comparative Study, Sustainability, vol. 11, no. 16, p. 4484, doi: 10.3390/su11164484, Dec. 2018

Bassolas A., Ramasco J. J., Herranz R., and Cantú-Ros O. G., Mobile phone records to feed activity-based travel demand models: MATSim for studying a cordon toll policy in Barcelona, Transp. Res. Part Policy Pract., vol. 121, pp. 56–74, doi: 10.1016/j.tra.2018.12.024, Mar. 2019

Bonnel P., Fekih M., and Smoreda Z., Origin-Destination estimation using mobile network probe data, Transp. Res. Procedia, vol. 32, pp. 69–81, doi: 10.1016/j.trpro.2018.10.013, 2018

Calabrese F., Di Lorenzo G., Liu L., and Ratti C., Estimating Origin-Destination Flows Using Mobile Phone Location Data, IEEE Pervasive Comput., vol. 10, no. 4, pp. 36–44, doi: 10.1109/MPRV.2011.41, Apr. 2011

Catapult, Mobile phone data in transport modelling - Recommendations Paper. UK Department for Transport. Available online at: https://www.gov.uk/government/publications/mobile-phone-data-in-transport-modelling , 2017

Chen X. et al., Trip-Chain-Based Travel-Mode-Shares-Driven Framework using Cellular Signaling Data and Web-Based Mapping Service Data, Transp. Res. Rec. J. Transp. Res. Board, vol. 2673, no. 3, pp. 51–64, doi: 10.1177/0361198119834006, Mar. 2019

De Montjoye Y. A. et al., On the privacy-conscientious use of mobile phone data, Sci. Data, vol. 5, no. 1, p. 180286, doi: 10.1038/sdata.2018.286, Dec. 2018

European Parliament, General Data Protection Regulation. Available online at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679 , 2016

Fekih M., Bellemans T., Smoreda Z., Bonnel P., Furno A., and Galland S., A data-driven approach for origin–destination matrix construction from cellular network signalling data: a case study of Lyon region (France), Transportation, vol. 48, no. 4, pp. 1671–1702, doi: 10.1007/s11116-020-10108-w, Aug. 2021

Friso K. et al., The Use Of Mobile Phone Data In Transport Planning In The Netherlands: Experiences And Vision, 2018

Hagenauer J. and Helbich M., A comparative study of machine learning classifiers for modeling travel mode choice, Expert Syst. Appl., vol. 78, pp. 273–282, doi: 10.1016/j.eswa.2017.01.057, Jul. 2017

Horn C., Gursch H., Kern R., and Cik M., QZTool—Automatically Generated Origin-Destination Matrices from Cell Phone Trajectories, in Advances in Human Aspects of Transportation, vol. 484, N. A. Stanton, S. Landry, G. Di Bucchianico, and A. Vallicelli, Eds. Cham: Springer International Publishing, pp. 823–833. doi: 10.1007/978-3-319-41682-3_68, 2017

Huang Z. et al., Modeling real-time human mobility based on mobile phone and transportation data fusion, Transp. Res. Part C Emerg. Technol., vol. 96, pp. 251–269, doi: 10.1016/j.trc.2018.09.016, Nov. 2018

Imai R., Ikeda D., Shingai H., Nagata T., and Shigetaka K., Origin-Destination Trips Generated from Operational Data of a Mobile Network for Urban Transportation Planning, J. Urban Plan. Dev., vol. 147, no. 1, p. 04020049, doi: 10.1061/(ASCE)UP.1943-5444.0000635, Mar. 2021

Iqbal M. S., Choudhury C. F., Wang P., and González M. C., Development of origin–destination matrices using mobile phone call data, Transp. Res. Part C Emerg. Technol., vol. 40, pp. 63–74, doi: 10.1016/j.trc.2014.01.002, Mar. 2014

Isaacman S. et al., Identifying Important Places in People's Lives from Cellular Network Data, in Pervasive Computing, vol. 6696, K. Lyons, J. Hightower, and E. M. Huang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 133–151. doi: 10.1007/978-3-642-21726-5_9, 2011

Jiang S., Ferreira J., and Gonzalez M. C., Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore, IEEE Trans. Big Data, vol. 3, no. 2, pp. 208–219, doi: 10.1109/TBDATA.2016.2631141, Jun. 2017

Liang X. and Wang G., A Convolutional Neural Network for Transportation Mode Detection Based on Smartphone Platform, in 2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), Orlando, FL, pp. 338–342. doi: 10.1109/MASS.2017.81, Oct. 2017

Luca M., Barlacchi G., Lepri B., and Pappalardo L., A Survey on Deep Learning for Human Mobility, 2020

Pourmoradnasseri M., Khoshkhah K., Lind A., and Hadachi A., OD-Matrix Extraction based on Trajectory Reconstruction from Mobile Data, in 2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), Barcelona, Spain, pp. 1–8. doi: 10.1109/WiMOB.2019.8923358, Oct. 2019

Scholl L. et al., Como aplicar big data en la planificación del transporte urbano: El uso de datos de telefonía móvil en el análisis de la movilidad, Inter-American Development Bank. doi: 10.18235/0002009, Dec. 2019

Tolouei R., Psarras S., and Prince R., Origin-Destination Trip Matrix Development: Conventional Methods versus Mobile Phone Data, Transp. Res. Procedia, vol. 26, pp. 39–52, doi: 10.1016/j.trpro.2017.07.007, 2017

Toole J. L., Colak S., Sturt B., Alexander L. P., Evsukoff A., and González M. C., The path most traveled: Travel demand estimation using big data resources, Transp. Res. Part C Emerg. Technol., vol. 58, pp. 162–177, doi: 10.1016/j.trc.2015.04.022, Sep. 2015

Torre-Bastida A. I., Del Ser J., Laña I., Ilardia M., Bilbao M. N., and Campos-Cordobés S., Big Data for transportation and mobility: recent advances, trends and challenges, IET Intell. Transp. Syst., vol. 12, no. 8, pp. 742–755, doi: 10.1049/iet-its.2018.5188, Oct. 2018

Willumsen L., Use of Big Data in Transport Modelling, vol. No. 2021/05. Paris: OECD Publishing, 2021

Wismans L. J. J., Friso K., Rijsdijk J., De Graaf S. W., and Keij J., Improving A Priori Demand Estimates Transport Models using Mobile Phone Data: A Rotterdam-Region Case, J. Urban Technol., vol. 25, no. 2, pp. 63–83, doi: 10.1080/10630732.2018.1442075, Apr. 2018

Zhou Z., Yang J., Qi Y., and Cai Y., Support vector machine and back propagation neutral network approaches for trip mode prediction using mobile phone data, IET Intell. Transp. Syst., vol. 12, no. 10, pp. 1220–1226, doi: 10.1049/iet-its.2018.5203, 2018