



HAL
open science

Combining Bayesian inference and clustering for transport mode detection from sparse and noisy geolocation data

Danya Bachir, Ghazaleh Khodabandelou, Vincent Gauthier, Mounim El Yacoubi, Eric Vachon

► **To cite this version:**

Danya Bachir, Ghazaleh Khodabandelou, Vincent Gauthier, Mounim El Yacoubi, Eric Vachon. Combining Bayesian inference and clustering for transport mode detection from sparse and noisy geolocation data. ECML PKDD 2018: Machine Learning and Knowledge Discovery in Databases, Sep 2018, Dublin, Ireland. pp.569-584, 10.1007/978-3-030-10997-4_35 . hal-01939608

HAL Id: hal-01939608

<https://hal.science/hal-01939608v1>

Submitted on 21 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Combining Bayesian Inference and Clustering for Transport Mode Detection from Sparse and Noisy Geolocation Data

Danya Bachir¹²³ [✉], Ghazaleh Khodabandelou², Vincent Gauthier², Mounim El Yacoubi², and Eric Vachon³

¹ IRT SystemX, Palaiseau, France
danya.bachir@gmail.com,

² SAMOVAR, Telecom SudParis, CNRS, Université Paris Saclay, France

³ Bouygues Telecom Big Data Lab, Meudon, France

Abstract. Large-scale and real-time transport mode detection is an open challenge for smart transport research. Although massive mobility data is collected from smartphones, mining mobile network geolocation is non-trivial as it is a sparse, coarse and noisy data for which real transport labels are unknown. In this study, we process billions of Call Detail Records from the Greater Paris and present the first method for transport mode detection of any traveling device. Cellphones trajectories, which are anonymized and aggregated, are constructed as sequences of visited locations, called sectors. Clustering and Bayesian inference are combined to estimate transport probabilities for each trajectory. First, we apply clustering on sectors. Features are constructed using spatial information from mobile networks and transport networks. Then, we extract a subset of 15% sectors, having road and rail labels (e.g., train stations), while remaining sectors are multi-modal. The proportion of labels per cluster is used to calculate transport probabilities given each visited sector. Thus, with Bayesian inference, each record updates the transport probability of the trajectory, without requiring the exact itinerary. For validation, we use the travel survey to compare daily average trips per user. With Pearson correlations reaching 0.96 for road and rail trips, the model appears performant and robust to noise and sparsity.

Keywords: Mobile Phone Geolocation, Call Detail Records, Trajectory Mining, Transport Mode, Clustering, Bayesian Inference, Big Data

1 Introduction

The growing use of smartphones generates massive ubiquitous mobility data. With unprecedented penetration rates, mobile networks are supplying the largest geolocation databases. Mobile phone providers collect real-time Call Detail Records (CDR) from calls, text messages or data at no extra-cost for billing purposes. Still, traditional transport planning models have so far relied on expensive travel surveys, conducted once a decade. Consequently,

surveys are rapidly outdated, while suffering from sampling bias and biased users' responses. Past research used CDR to estimate travel demand [21], optimal locations for new transport infrastructures [7], weekly travel patterns [9], activity-based patterns [12], urban land-use [19], impact of major events or incidents [6] and population dynamics [5,14]. A few studies used triangulation, based on signal strength e.g., in Boston U.S. [4,20]. In Europe, privacy policies restrict triangulation usage to police demands. CDR and GPS data both respect privacy compliance for geolocation. Still GPS data collection requires users to install tracking applications and activate GPS, which has greedy battery consumption. Consequently, GPS samples represent subsets of users' trips while CDR generate locations from larger populations over longer time periods. However CDR geolocation is coarse, noisy and affected by the usage frequency of devices. Raw CDR provide approximate and partial knowledge of true users' paths, hence requiring careful pre-processing. Past methods on transport mode detection mainly involved GPS data and are hardly transposable to CDR. In addition, these studies applied supervised learning [10,18,22] requiring a training dataset of trajectories with transport mode labels. Transport modes were either collected via applications where users consent to enter their travel details, or manually identified using expert knowledge, which is a costly task. In real world scenarios, transport modes of traveling populations are unavailable. Therefore we need new unsupervised approaches to tackle this issue.

This paper presents the first unsupervised learning method for transport mode detection from any CDR trajectory. As this is a first study, we focus on a bimodal separation between road and rail trips. In collaboration with a mobile phone provider, we process one month trajectories from the Greater Paris, which are anonymized and aggregated for privacy. Trajectories are represented as sequences of visited mobile network areas, called sectors. Our model combines clustering with Bayesian inference to determine the probability that cellphones traveled by road or rail knowing their trajectories on the mobile network. The transport probability of a trajectory is initialized with a prior obtained from the travel survey and updated with each new visited sector. Transport probabilities for sectors are derived after clustering sectors by transport type. Sectors features are constructed using both mobile networks and transport networks spatial properties. Then, for a subset of 15% sectors, we extract transport labels, being road or rail, (e.g., equipments inside train stations, on highways etc.) while the remaining sectors are multimodal. For each cluster, we use the binary labels to calculate continuous transport probabilities as the proportion of labeled sectors among total sectors. Trajectories are thus attributed the most probable mode among road, rail or mixed (i.e., when probabilities are close). For validation, we calculate daily average rail and road trip counts per user and obtain Pearson correlations with the travel survey above 0.96, for the 8 departments of the region. In the next sections, we review the literature in Sec. 2 and describe data engineering in Sec. 3. The methodology steps are presented in Sec. 4. Eventually, we discuss main results in Sec. 5 and provide conclusion.

2 Related Work

Common applications for geolocation data mining are the identification of travel patterns for personal travel recommendation [23,24], anomalous behavior detection[17] and transport planning [12]. Several works used supervised transport mode learning from GPS trajectories. A multilayer perceptron was used to identify car, bus and walkers modes for 114 GPS trajectories in [10]. Features were the average and maximum speed and acceleration, the total and average travel distance, the number of locations divided by travel distance and the number of locations divided by travel time. The best accuracy was 91% using a 10-folds cross validation. In [18], speed and acceleration features were collected from 16 GPS trajectories. Several classification models (Decision Tree, Kmeans, Naïve Bayes, NNeighbor, SVM, Discrete and Continuous HMM) were compared. The Decision tree with Discrete Hidden Markov Model obtained the highest accuracy (74 %). Still, supervised approaches with GPS are constrained by the small size of the training data. Moreover, although transport labels can be collected for small GPS datasets, they are unavailable for CDR.

Meanwhile, few studies tackled unsupervised transport mode detection. In [8] fuzzy logic was used as a scoring function calculated between consecutive GPS traces. The transport score was calculated with boolean conditions on speed, distances to transport network and previous mode. Still, this work lacked a performance evaluation. In [15], base stations located inside Paris underground were used to identify underground mode from CDR trips. A record detected by an underground antenna was labeled accordingly. This approach is limited as it relies exclusively on indoor equipment inside the underground. No additional modes were identified. To our knowledge, only one work addressed unsupervised transport mode learning for two modes, road and public transport, using triangulated CDR [20]. The approach applies travel times clustering followed by a comparison with Google travel times. Still, CDR low frequency induces important incertitude and delay on start and end travel times of CDR trips. Consequently a device may not be detected as traveling when the real trip begins and ends. Moreover the presented approach was demonstrated on one unique Origin and Destination (OD) pair which is not sufficient to validate the method. In dense urban areas, travel times can be affected by traffic states (e.g., rush hours) and can be identical for several modes, depending on the OD.

Our work presents a novel method for transport mode detection by combining two unsupervised techniques, namely clustering and Bayesian inference. This model classifies millions of CDR trajectories into road and rail trips. Instead of clustering trajectories with features such as speed or travel time, highly impacted by the imprecision, sparsity and noise of CDR geolocation, we apply clustering on sectors and build spatial features using transport networks. A small subset of road and rail labels is collected for sectors in order to calculate sectors transport probabilities. After the Bayesian inference step, we conduct a large-scale validation for the complete region, using the travel survey. The high Pearson correlations, obtained on daily average trips per user, proves the method is generalizable, performant and robust to noise and sparsity.

3 Data Engineering

For this study, we collect anonymized CDR trajectories from the Greater Paris region, over one month. Sectors features are constructed using the base stations referential jointly with transport networks infrastructures. For data normalization, we introduce a specific procedure accounting for heterogeneous urban density. Label extraction is realized to gather transport labels for a small subset of sectors. For model validation we use the household travel survey from 2010 conducted by Île de France Mobilités-OMNIL-DRIEA [1].

3.1 Mobile Network

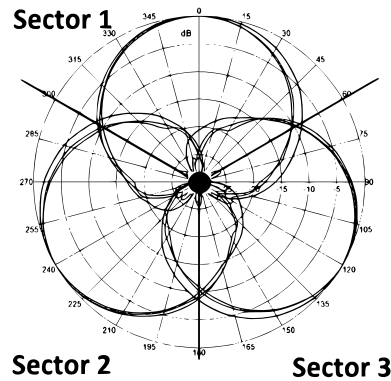


Fig. 1: *schema of a tri-sector antenna. The antenna is represented by the black dot. Circular areas are cells for 2G, 3G and 4G signals.*

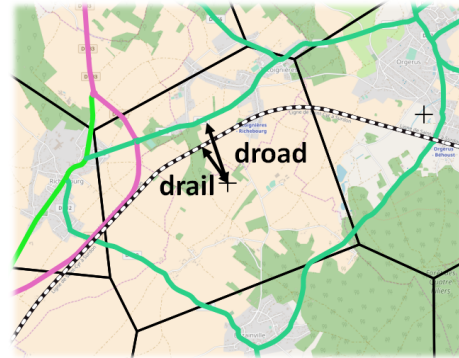


Fig. 2: *Example of a voronoi sector and its associated shortest distance to transports axes. Five roads (colored lines) and one rail line (dashed line) intersect the sector.*

Mobile providers do not have access to GPS coordinates of mobile phones. Although we know which base station is connected to a device, it is unlikely to encounter mobile users positioned exactly at the base station. Devices are located inside mobile network areas covered by base stations signal range. For this study, we use the mobile network referential of the Greater Paris region. This region has a 12000 km^2 area with more than 1200 cities and 12 millions inhabitants. It is covered by thousands of mobile network antennas. Each base station emits 2G, 3G or 4G radio signals. Cells are circular areas covered by signals (see Fig. 1). Each cell equipment is oriented toward one direction. The partitions of cells directions are called sectors. The average sector number per antenna is 3 where one sector covers 120° around the base station. A cellular tessellation is composed of a multitude of overlapping areas. We use the sector tessellation to get rid of overlaps and create the voronoi partitions using sectors centroids (see Fig. 2). We associate each mobile phone record to a sector location.

3.2 Transport Networks

Transport networks are used to construct sectors features. We retrieve rails infrastructures for underground, overground, tramway and train stations shared by Île-de-France Mobilité on their platform Open Data STIF [2]. In addition we use OpenStreetMap (OSM) [3] to retrieve highspeed rails and road infrastructures. Roads are categorized by traffic importance. We filter residential roads which have highest road count and lowest traffic.

3.3 Raw Features Construction

We construct our dataset $D = \{d_{rail}, d_{road}, d_{station}, n_{road}, n_{rail}, w_{station}\}$ where features stand for:

- d_{road} : shortest distance between sector centroid and road (see Fig. 2).
- d_{rail} : shortest distance between sector centroid and rail network (see Fig. 2).
- $d_{station}$: shortest distance between sector centroid and train station centroid.
- n_{road} : number of roads intersecting the voronoi.
- n_{rail} : number of rail lines intersecting the voronoi.
- $w_{station}$: weight of train stations calculated as the sum of stations area intersecting the sector voronoi area.

3.4 Data Normalization

We aim to find transport mode usage in sectors. As our raw features are built with spatial information they are impacted by urban density. In the city center the density is higher than in the suburb. Consequently sector areas and distances to transport networks are smaller while there are more transport hubs. We normalize our features to reduce the bias induced by urban density over transport usage. We introduce a normalization specific to our problem:

$$d_{norm,m} = \frac{d_m}{\sum_i d_i} \in [0, 1] \quad (1)$$

$$n_{norm,m} = \frac{n_m}{\sum_i n_i} \in [0, 1] \quad (2)$$

$$w_{norm,station} = \frac{w_{station}}{A_v} \in [0, 1] \quad (3)$$

where $d_m \in \{d_{road}, d_{rail}, d_{station}\}$, $n_m \in \{n_{road}, n_{rail}\}$ and $d_{norm,m}$, resp. $n_{norm,m}$, is the normalized vector for feature d_m , resp. n_m . Feature $w_{norm,station}$ is the normalization of $w_{station}$ by voronoi area A_v .

3.5 Sector Label Extraction

A few base stations are located on transport hubs, such as rail lines, train stations, highways or tunnels. We process this information to construct labels for a small subset of antennas. We assume that each sector inherits from its

base station label. We attribute rail labels to indoor equipments located inside the underground and train stations, which represent 4% sectors. We assign road mode to indoor antennas in tunnels, constituting less than 1% sectors. We add outdoor antennas on highways (11% sectors) to increase the size of the road subset. In total we obtain 15% transport labels. In what follows, we use our subset of sectors with categorical transport labels $\{road, rail\}$, as prior knowledge. Still, categorical transport labels are not appropriate for most sectors, including outdoor equipments. In urban areas, such as the Greater Paris, the classic scenario is to encounter several transport modes inside an outdoor sector because of mobile networks' coarse granularity. Thus, we aim to find continuous transport probabilities $P \in [0, 1]$ for all sectors, where indoor labeled equipments have maximal probabilities $P \in \{0, 1\}$.

3.6 Trajectories Pre-processing

For this study, the mobile provider pre-processed raw anonymized users' positions using noise reduction and segmentation (see Fig. 3). For segmentation, users' locations were separated into stay points i.e., when users remain in the same area, and moving points i.e., when users are assumed traveling. We define a trajectory as a sequence of moving points $T_j^u = \{(X_0, t_0), \dots, (X_l, t_l)\}$, j being the j^{th} trajectory of the user u . The i^{th} position recorded at timestamp t_i is $X_i = (x_i, y_i)$, where (x_i, y_i) are the centroid coordinates of the visited sector. One trajectory corresponds to one user trip. We construct 95 millions CDR trajectories from 2 millions anonymized users during one month. Similar trajectories are aggregated to respect privacy policies. In order to compare our results with household travel survey, which was conducted for residents of the Greater Paris region, the mobile provider filters users by home department (first two digits of billing address postcode) and exclude visitors.

4 Model

This section presents the unsupervised learning scheme combining clustering and Bayesian inference to estimate transport modes of CDR trajectories. First, the prior transport probability is obtained from the travel survey. Second, the transport likelihood is calculated from the observed records, such as each new visited sector updates the probability. In this perspective, we apply a clustering on sectors. Then, our subset of sectors labels is used to calculate transport probabilities within each cluster. Each sector is assigned a continuous score in $[0, 1]$ reflecting the real transport usage inside i.e., the probability to detect more users on the roads or on the rails. For each trajectory, we assign the mode with highest probability. Eventually, results are validated against the survey.

4.1 Clustering

We aim to find transport clusters for mobile network sectors with an underlying hierarchical structure. Thus we use an agglomerative hierarchical clustering. The

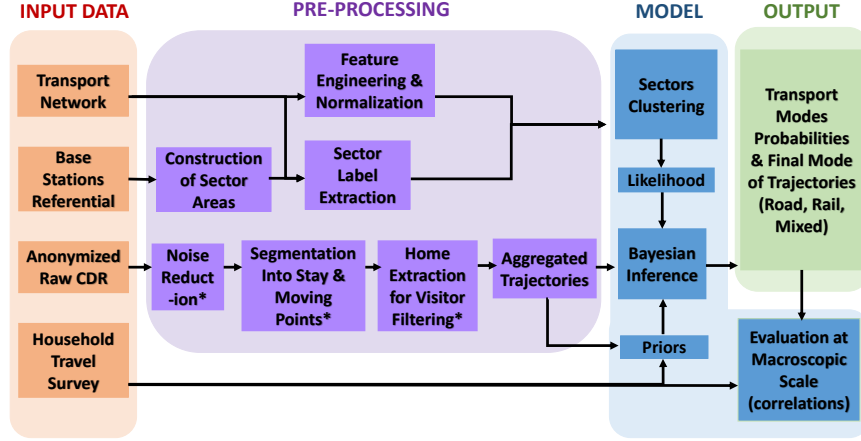


Fig. 3: Transport mode detection workflow applied for this case study. Pre-processing steps annotated with (*) were performed by the mobile operator.

clustering starts with N clusters of size 1, N being the number of sectors. Each sector is recursively merged with its closest neighbor according to a linkage criterion and a distance function. We test three linkage types with three distance functions (euclidean, Manhattan and cosine). Complete linkage minimizes the maximal distance between two points from two clusters. Average linkage minimizes the average distance between clusters points. Ward linkage, with euclidean distance, minimizes the sum of squared error $ESS = \sum_{k,i,j} |X_{ijk} - \bar{x}_{kj}|^2$, where X_{ijk} is the sample value for sector i , feature j and cluster k ; \bar{x}_{kj} is the mean value of feature j for cluster k . The agglomerative clustering applies until all data points are merged into a single cluster of size N . A good clustering solution should divide rail transport sectors from road sectors.

4.2 Evaluation Metrics

We use internal evaluation metrics to assess the clustering performance and to identify the optimal cluster number. We used the Silhouette (S) to evaluate clusters separability [13] (see Eq. 4).

$$s_{ik} = \frac{b(i) - a(i)}{\max(b(i), a(i))} \quad (4)$$

$$S_k = \frac{1}{N_k} \sum_{i=1}^{N_k} s_{ik} \quad (5)$$

$$S = \frac{1}{N} \sum_k S_k \quad (6)$$

where $a(i)$ is the average intra cluster distances for sector i and $b(i)$ is the lowest value among average inter cluster distances. Here N_k stands for the size of cluster k . The number of samples equals N . The optimal number of clusters K maximizes the silhouette [16]. In addition we used the S_{dbw} validity index.

$$S_{dbw}(k) = Scat(k) + Dens_{db}(k) \quad (7)$$

$$\text{where } Scat(k) = \frac{1}{k} \sum_{i=1}^k \frac{\sigma(v_i)}{\sigma(D)} \quad (8)$$

$$\text{and } Dens_{db}(k) = \frac{1}{k(k-1)} \sum_{i,j=1}^k \frac{dens(u_{ij})}{\max(dens(v_i), dens(v_j))} \quad (9)$$

where v_i denotes centroid of cluster i and u_{ij} is the middle point between clusters i and j i.e., at mid distance from the two centroids (v_i, v_j) . The scattering index $Scat$ is used to estimate the intra cluster compactness based on standard deviations σ of clusters over total dataset D . The term $Dens_{db}$ represents clusters densities. It calculates the average ratio of clusters middle point densities over clusters centers densities. The underlying assumption is that well defined clusters are denser around their centroids than at their mid distance. This index is a trade-off between clusters densities and variances. It has been depicted as the most performing among internal clustering evaluation metrics in [11, 16]. The optimal cluster number is found when the index reaches its minimum.

4.3 Probability Scores of Sectors Transport Mode

For each cluster k we calculate the score $p_{k,m}$ for transport mode $m \in \{rail, road\}$.

$$p_{k,m} = \frac{N_{k,m}}{N_m} \quad (10)$$

where $N_{k,m}$ is the number of labeled sectors belonging to class m in cluster k and N_m is the total number of sectors from class m in the dataset. We normalize $p_{k,m}$ to obtain the probability $P(m|S_i) \in [0, 1]$ of using mode m given a visited sector S_i , belonging to a cluster k .

$$P(m|S_i) = \frac{p_{k,m}}{\sum_j p_{k,j}} \quad (11)$$

Unlabeled sectors obtain transport probabilities according to their cluster. In addition we update the probabilities of outdoor labeled sectors (i.e., highways) using Eq. 10 and 11. Indoor labeled sectors have binary probabilities in $\{0, 1\}$.

4.4 Bayesian Inference of Trajectories Transport Mode

Bayesian inference is used to determine the main transport mode associated to mobile phone trajectories. In this perspective, we calculate the probability

$P(m|T_j^u)$ to take a mode $m \in \{rail, road\}$ knowing the trajectory T_j^u , using Bayes theorem:

$$P(m|T_j^u) = \frac{P(T_j^u|m) * P(m)}{P(T_j^u)} \quad (12)$$

Trajectories are sequences of sectors $\{S_0, \dots, S_l\}$ visited by mobile phone holders. Thus we have $P(T_j^u|m) = P(S_0, \dots, S_l|m)$. We assume independence between sectors probabilities such as $P(S_i, S_{i+1}|m) = P(S_i|m)P(S_{i+1}|m)$. This assumption is motivated by the need to reduce the computational cost of the calculation. Thus we can rewrite $P(T_j^u|m) = \prod_{i=0}^l P(S_i|m)$. Eq. 12 becomes:

$$P(m|T_j^u) = \frac{P(m)}{P(T_j^u)} \prod_{i=0}^l P(S_i|m) \quad (13)$$

The term $P(m|S_i)$, previously calculated with Eq.11, is introduced by applying Bayes theorem a second time, to Eq. 12:

$$P(m|T_j^u) = \frac{\prod_{i=0}^l P(S_i)}{P(T_j^u)} P(m)^{1-l} \prod_{i=0}^l P(m|S_i) \quad (14)$$

The term $\frac{\prod_{i=0}^l P(S_i)}{P(T_j^u)}$ does not influence the mode choice. The prior transport probability $P(m)$ can be seen as the initial guess, before observing records. The prior probability is obtained from the travel survey and is calculated as the average trip counts per user given the home location of cellphone holders, here at the department scale. For rail mode we have $p_{rail,dep} = \frac{AVG_{dep}(c_{rail})}{AVG_{dep}(c_{rail}) + AVG_{dep}(c_{road})} \in [0, 1]$ and $p_{rail,dep} = 1 - p_{road,dep}$, where c_{rail} and c_{road} are the rail and road trip counts, for the day of survey, per user living in the department dep . At last we normalize the posterior transport probability to be in range $[0, 1]$.

$$P(m|T_j^u) \leftarrow \frac{P(m|T_j^u)}{P(rail|T_j^u) + P(road|T_j^u)} \quad (15)$$

Finally we affect the mode obtaining the higher probability to each trajectory. When probabilities are in $[0.4, 0.6]$ the mode is considered mixed.

5 Results

This section summarizes our main results. For the clustering we demonstrate how we determine the number of clusters. We describe clusters according to transport probabilities. From the Bayesian inference of trajectories' transport modes, we visualize transport flows per week day and observe the travel patterns. We provide detailed results comparison with survey, at department scale, using Pearson correlations as evaluation metric.

5.1 Clustering Evaluation

We first compare the three linkage types. Average and complete linkage fail to separate sectors in the city center, with any distance metric. One huge centered cluster is produced with tiny clusters located at the region borders. We retain ward linkage with euclidean distance which produce clusters of comparable size, evenly present across the region. In order to find the optimal number of cluster we draw the dendrogram of the ward agglomerative clustering (see Fig. 4). The latter shows $k = 2$ is a good cluster number as it corresponds to the highest distance gap between merges. A small k leads to a macroscopic partitioning. We look for a higher k to detect finer transport modes tendencies. A clear cut was possible for $k \in \{3, 4, 5, 9\}$, which were therefore also good candidates. We decide to bound the cluster number between 2 and 10. We

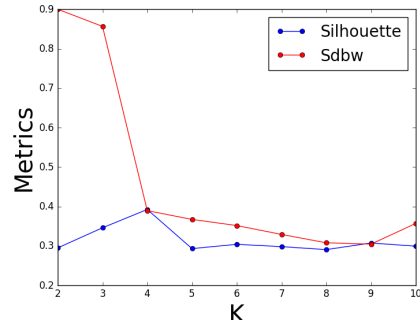
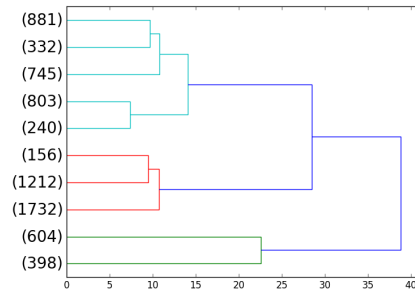


Fig. 4: Dendrogram for $k \in [2, 10]$. The xaxis is the height i.e., distances between clusters leaves and nodes. The yaxis shows the number of leaves per cluster. **Fig. 5:** Silhouette (blue) and S_{dbw} validity index (red) plotted in function of the number of cluster k

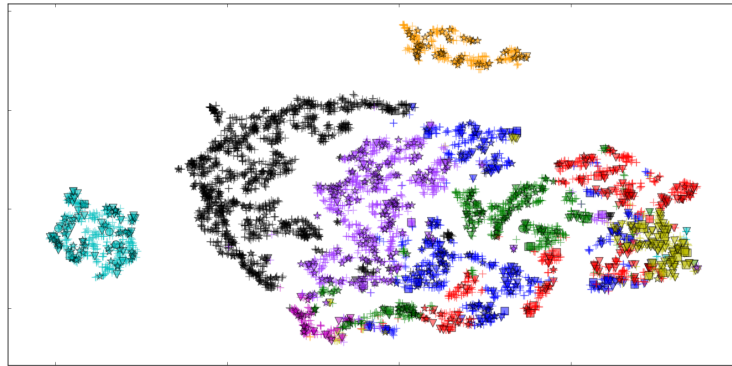


Fig. 6: t -sne projection for dataset D after normalization and z-score transformation. Colors represent clusters for k varying from 1 to 9. The parameters are $n_{component} = 2$, perplexity = 30, learningrate = 200, $n_{iteration} = 1000$. Stars correspond to road labels, Triangle to rails and crosses to unlabeled sectors.

use additional intra-cluster metrics. We calculate S and S_{dbw} with several k values (see Fig. 5). The silhouette reaches a maximum for $k = 4$, for which separability is the highest. According to the S_{dbw} minimization criterion, the optimal number of clusters is $k = 9$, for which clusters are the most compact and dense. For $k \in [5, 10]$ the silhouette reaches a local maximum for $k = 9$. For our problem we favor the larger k hence we select $k = 9$. We visualize the 9 clusters with t-sne (see Fig.6) and project them on the sectors map (see Fig.7).

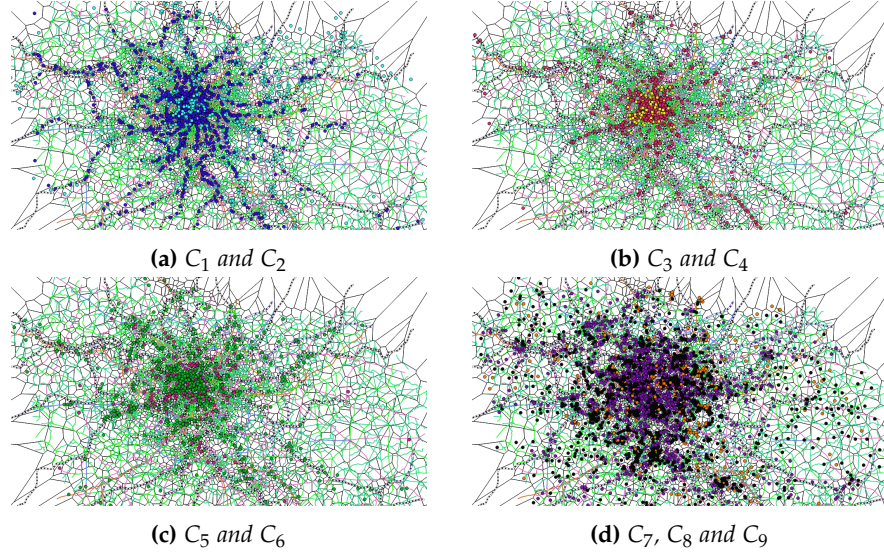


Fig. 7: QGIS Clusters projection

5.2 Sectors Probabilities and Visualization

We calculate the transport probabilities per cluster (see Tab. 1). We describe clusters regarding transport usage. Each cluster is displayed in Fig.6 and Fig.7.

Table 1: Transport Mode probabilities and cluster size for $k = 9$

Cluster	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
Size (%)	14.7	8.50	12.4	4.67	2.20	10.5	24.4	5.60	17.1
P_{RAIL}	0.651	0.567	0.824	0.949	0.421	0.387	0.095	0.071	0.199
P_{ROAD}	0.348	0.432	0.176	0.051	0.579	0.613	0.905	0.929	0.801

- C_1, C_2 : mixed-rail clusters with a higher probability for rails, depicted in blue and cyan on Fig. 7a.
- C_3, C_4 : rail dominated clusters with many underground sectors located in the city center. It corresponds to the red and yellow cluster on Fig. 7b.
- C_5, C_6 : mixed road clusters, shown in magenta and green on Fig. 7c.
- C_7, C_8, C_9 : road clusters represented in black, orange and purple on Fig. 7d.

5.3 Trajectories

We infer transport probabilities for one month trajectories, filtering bank holidays. We count the number of rail and road trips (see Fig. 8). Only 3% trips have probabilities in range $[0.4, 0.6]$. We consider such trips have mixed (or uncertain) mode. In Fig. 8 we observe hourly travel patterns for a typical week. For business days, peak hours occur in the morning and early evening, with a smaller midday peak at lunch time. Morning and evening peaks appear unbalanced. One reason is that mobile phone usage tends to be more important in the evening thus we detect more users and more trips. A second reason could be that users travel more at the end of the day. This phenomenon is more pronounced for road trips, the highest gap being on friday evening.

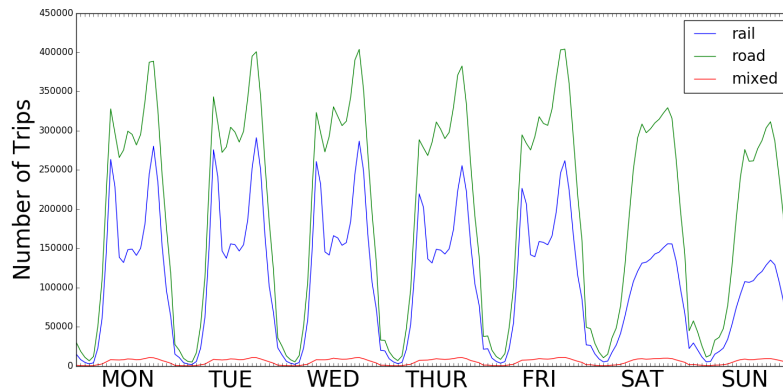


Fig. 8: Estimated trip counts are averaged per week day, per hour and per transport mode. Results are given for 1 month data from the Greater Paris.

5.4 Comparison with Survey

We compare our results with the latest household travel survey, from 2010, for the Greater Paris. About 43000 residents were asked about their travels during their past day, outside holidays. We calculate mobility statistics from survey and MP results (see Tab. 2). We average survey trip counts per resident: $C^S = \frac{\sum_{i=1}^k N_i * w_i}{\sum_{i=1}^k w_i}$ where an individual i of weight w_i reported N_i trips for the day he was questioned. The weight w_i was calculated during survey with socio-demographic information to rescale the individual to the entire population. Similarly we average CDR trip counts per day and per device: $C^{MP} = \frac{\sum_{i=1}^U \sum_{t=1}^T \frac{1}{U} \frac{1}{T} n_{u,i}}{\sum_{i=1}^U \sum_{t=1}^T \frac{1}{U} \frac{1}{T}}$ where U is the number of phones, T is the number of days and $n_{i,t}$ is the number of trips detected for phone i for day t . In the survey, transport modes are separated in two categories, motorized modes including public transport, cars and motorbikes, and unmotorized modes i.e., walk and bike. Our model outputs the majority mode of a given CDR trajectory, between

rail and road. We first examine results for all residents (see Tab. 2). The survey indicates the average trip number per user during a business day is 4.16 for all modes and 2.45 for motorized trips. We found an average of 2.10 daily trips per person. It seems we were able to detect 86% motorized modes. Because of the coarseness of the mobile network, walkers might be considered as non moving as their movement occurs at a too microscopic scale. In addition, the detection of travels is affected by CDR frequency. When a device is turned-off or unused for a long period of time, users are undetected. Compared to the survey, 14% daily motorized trips are undetected in average. We further analyze results for residents aggregated by home given for the city center, first ring, second ring and department scale (first two digits of postcode). We calculate Pearson correlations between survey and CDR estimates for all trips, motorized, road and rail trips. In addition we calculate the ratio between road and rail trips: $C_{ratio} = \frac{C_{road}}{C_{rail}}$. There is a negative correlation between total survey trips and CDR trips, due to the possible undetection of unmotorized modes. Correlations for rail, road and ratio are all above 0.96 for the three rings scale and the department scale. Still we have smaller ratio than the survey. The department

Table 2: Mobility statistics for average trip number per user during a business day (Monday-Friday). Results are given per home location (i.e., 2nd ring, 1st ring and the 8 departments including city center). Left: results for survey (source: EGT 2010-Île de France Mobilités-OMNIL-DRIEA) Right: results with CDR

Home Scale	Survey (S)					Mobile Phone (MP)			
	C_{All}^S	C_{Motor}^S	C_{Rail}^S	C_{Road}^S	C_{Ratio}^S	C_{All}^{MP}	C_{Rail}^{MP}	C_{Road}^{MP}	C_{Ratio}^{MP}
All population	4.16	2.45	0.61	1.85	3.03	2.10	0.80	1.30	1.62
City Center (CC)	4.37	1.93	1.11	0.83	0.75	1.94	1.22	0.72	0.59
1st Ring (R1)	4.03	2.25	0.61	1.64	2.69	2.07	0.80	1.27	1.60
2nd Ring (R2)	4.18	2.86	0.38	2.49	6.55	2.24	0.50	1.74	3.45
Dep 77 (D2)	4.12	2.90	0.30	2.60	8.79	2.37	0.49	1.88	3.83
Dep 78 (D3)	4.23	2.88	0.41	2.47	6.03	2.21	0.52	1.69	3.28
Dep 91 (D4)	4.30	3.07	0.34	2.73	7.91	2.15	0.44	1.71	3.92
Dep 92 (D5)	4.18	2.22	0.62	1.60	2.56	1.98	0.83	1.15	1.38
Dep 93 (D6)	3.84	2.20	0.62	1.58	2.57	2.15	0.80	1.35	1.69
Dep 94 (D7)	4.05	2.34	0.60	1.74	2.91	2.11	0.75	1.35	1.79
Dep 95 (D8)	4.06	2.57	0.45	2.13	4.76	2.21	0.57	1.65	2.90

Table 3: Pearson correlation coefficients between survey and results. We calculate correlations across the 3 rings (city center, rings 1 and 2) and across the 8 departments.

Home Scale	$(C_{All}^S, C_{All}^{MP})$	$(C_{Motor}^S, C_{All}^{MP})$	$(C_{Road}^S, C_{Road}^{MP})$	$(C_{Rail}^S, C_{Rail}^{MP})$	$(C_{Ratio}^S, C_{Ratio}^{MP})$
Rings (CC, R1-2)	-0.496	0.993	0.995	0.990	0.999
Deps (CC, D2-8)	-0.348	0.751	0.960	0.986	0.978

obtaining results most similar with the survey is the city center (Paris). For the latter we detect the same number of motorized trips. This means that all users' trips were detected, suggesting that mobile phone activity of travelers is more important in the city center. From these observations we emit several hypothesis to explain remaining differences. First, because of their cost, surveys are performed on small population samples. Despite the use of weights to scale the sample to the total population, results can still contain sampling bias in addition with users' responses bias. Second, travel surveys are performed every 10 years because of their high cost. The latest complete survey is anterior to our study (seven years difference) which can lead to differences in results. In particular, transport policies over the past years were oriented to favor public transport in the Greater Paris (e.g., introduction of a unique price for transport pass that reduced the price for suburbs). This could have influenced users to take public transports, especially in the suburb. In our opinion trips segmentation might impact results. Indeed our trajectories are segmented based on stay times. Public transport users sometimes experiment waiting times in stations e.g., when users change lines, and signals loss when entering the underground. This could cause higher trip segmentation for CDR rail trips. At last we detect 100% trips in the city center versus 80% in the suburb. In parallel the city center has the highest rail transport usage. This could indicate a bias in mobile phone usage i.e., public transport users are more likely to call, text or navigate on the web than drivers. Therefore some road trips could possibly be undetected.

6 Conclusion

From mobile phone data mining we can capture travel behavior of urban populations on multimodal transport networks. Compared to traditional travel surveys, Call Detail Records are a low-cost and up-to-date knowledge base for smart transport research. In this paper, we have introduced a novel transport mode detection method using CDR trajectories from the Greater Paris. Our model uses three data sources: mobile network data, transport networks and household travel survey. After significant data pre-processing, we combine clustering on mobile network areas, called sectors, with Bayesian inference for trajectories. From the clustering we find 9 clusters best described transport usage in the region. Three clusters exhibit high road probabilities, two had high rail probabilities while four had mixed usage. We compare our final results on trajectories with the household travel survey. Trips are aggregated by users' home location, at the department scale. We calculate the average number of trips per day for each user, averaged over all users. We obtain Pearson correlations above 0.96 for motorized, rail and road modes. It seems we detect exclusively motorized trips, as walkers movements are too microscopic regarding the mobile network scale. To our knowledge this is the first method separating road from rail trips considering all CDR trajectories from all users, with substantial comparison with survey data. Still it is hard to obtain exact same results as the survey. First we might have a different trip segmentation.

When users travel, their path on the network are likely to be segmented into subtrips because CDR are affected by waiting times and signals loss. This phenomenon could be more pronounced for public transport travels, as users often change lines and wait in stations. In addition, the detection of travels is impacted by usage frequency of phones. We observe that trips are most likely to be undetected when road usage is predominant. At last, surveys might contain bias, be outdated and miss particular events. This makes validation a difficult task as no available data source is a perfect ground truth. Our work shows encouraging results yet we have several pending issues we want to address in future works. First, although our model proved to be robust to noisy locations, oscillations filtering could be enhanced during CDR pre-processing. Second, as our model outputs one dominant mode, we need to address multi-modal and uncertain behaviors. For future work, we will extend model evaluation with finer scale Origin-Destination trips. We look forward to adding a fourth data source (e.g., travel cards data) for validation. We aim to enrich our model with additional transport modes. Our final model will be implemented by the mobile phone provider for B-2-B with transport operators and urban planners.

Acknowledgments

This research work has been carried out in the framework of IRT SystemX, Paris-Saclay, France, and therefore granted with public funds within the scope of the French Program “Investissements d’Avenir”. This work has been conducted in collaboration with Bouygues Telecom Big Data Lab.

References

1. OMNIL. <http://www.omnil.fr>.
2. Open Data STIF. <http://opendata.stif.info>.
3. OpenStreetMap. <http://openstreetmap.org>.
4. Lauren Alexander, Shan Jiang, Mikel Murga, and Marta C González. Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 58:240–250, 2015.
5. Danya Bachir, Vincent Gauthier, Mounim El Yacoubi, and Ghazaleh Khodabandelou. Using mobile phone data analysis for the estimation of daily urban dynamics. In *Intelligent Transportation Systems (ITSC), 2017 IEEE 20th International Conference on*, pages 626–632. IEEE, 2017.
6. James P Bagrow, Dashun Wang, and Albert-Laszlo Barabasi. Collective response of human populations to large-scale emergencies. *PloS one*, 6(3):e17680, 2011.
7. M. Berlingerio, F. Calabrese, G. Di Lorenzo, R. Nair, F. Pinelli, M.L. Sbodio, H. Blockeel, K. Kersting, S. Nijssen, and F. Zelezny. Allaboard: a system for exploring urban mobility and optimizing public transport using cellphone data. volume pt.III, IBM Research, Dublin, Ireland, 2013.
8. Filip Biljecki, Hugo Ledoux, and Peter Van Oosterom. Transportation mode-based segmentation and classification of movement trajectories. *International Journal of Geographical Information Science*, 27(2):385–407, 2013.

9. Francesco Calabrese, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):0036–44, 2011.
10. P Gonzalez, J Weinstein, S Barbeau, M Labrador, P Winters, Nevine Labib Georggi, and Rafael Perez. Automating mode detection using neural networks and assisted gps data collected using gps-enabled mobile phones. In *15th World congress on intelligent transportation systems*, 2008.
11. Maria Halkidi and Michalis Vazirgiannis. Clustering validity assessment: Finding the optimal partitioning of a data set. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 187–194. IEEE, 2001.
12. Shan Jiang, Joseph Ferreira, and Marta C Gonzalez. Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore. *IEEE Transactions on Big Data*, 3(2):208–219, 2017.
13. Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.
14. Ghazaleh Khodabandelou, Vincent Gauthier, Mounim El-Yacoubi, and Marco Fiore. Population estimation from mobile network traffic metadata. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2016 IEEE 17th International Symposium on A*, pages 1–9. IEEE, 2016.
15. Anahid Nabavi Larijani, Ana-Maria Olteanu-Raimond, Julien Perret, Mathieu Brédif, and Cezary Ziemlicki. Investigating the mobile phone data to estimate the origin destination flow and analysis; case study: Paris region. *Transportation Research Procedia*, 6:64–78, 2015.
16. Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. Understanding of internal clustering validation measures. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 911–916. IEEE, 2010.
17. Linsey Xiaolin Pang, Sanjay Chawla, Wei Liu, and Yu Zheng. On detection of emerging anomalous traffic patterns using gps data. *Data & Knowledge Engineering*, 87:357–373, 2013.
18. Sasank Reddy, Min Mun, Jeff Burke, Deborah Estrin, Mark Hansen, and Mani Srivastava. Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN)*, 6(2):13, 2010.
19. Jameson L Toole, Michael Ulm, Marta C González, and Dietmar Bauer. Inferring land use from mobile phone activity. In *Proceedings of the ACM SIGKDD international workshop on urban computing*, pages 1–8. ACM, 2012.
20. Huayong Wang, Francesco Calabrese, Giusy Di Lorenzo, and Carlo Ratti. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 318–323. IEEE, 2010.
21. Ming-Heng Wang, Steven D Schrock, Nate Vander Broek, and Thomas Mulinazzi. Estimating dynamic origin-destination data and travel demand using cell phone network data. *International Journal of Intelligent Transportation Systems Research*, 11(2):76–86, 2013.
22. Yu Zheng, Yukun Chen, Quannan Li, Xing Xie, and Wei-Ying Ma. Understanding transportation modes based on gps data for web applications. *ACM Transactions on the Web (TWEB)*, 4(1):1, 2010.
23. Yu Zheng, Like Liu, Longhao Wang, and Xing Xie. Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of the 17th international conference on World Wide Web*, pages 247–256. ACM, 2008.
24. Yu Zheng and Xing Xie. Learning travel recommendations from user-generated gps traces. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1):2, 2011.