Data-Driven Location Annotation for Fleet Mobility Modeling

Riccardo Guidotti
University of Pisa
Pisa, Italy
riccardo.guidotti@di.unipi.it

Mirco Nanni
ISTI-CNR, Pisa
Pisa, Italy
mirco.nanni@isti.cnr.it

Francesca Sbolgi
University of Pisa
Pisa, Italy
f.sbolgi@unipi.it

ABSTRACT

The large availability of mobility data allows studying human behavior and human activities. However, this massive and raw amount of data generally lacks any detailed semantics or useful categorization. Annotations of the locations where the users stop may be helpful in a number of contexts, including user modeling and profiling, urban planning, activity recommendations, and can even lead to a deeper understanding of the mobility evolution of an urban area. In this paper, we foster the expressive power of individual mobility networks, a data model describing users' behavior, by defining a data-driven procedure for locations annotation. The procedure considers individual, collective, and contextual features for turning locations into annotated ones. The annotated locations own a high expressiveness that allows generalizing individual mobility networks, and that makes them comparable across different users. The results of our study on a dataset of trucks moving in Greece show that the annotated individual mobility networks can enable detailed analysis of urban areas and the planning of advanced mobility applications.

1 INTRODUCTION

The large availability of digital traces of individuals is offering novel possibilities for understanding the patterns characterizing human mobility [7]. However, the personal data collected by smartphones or devices installed by car telematics companies for business and insurance purposes is generally limited to the positions of the vehicle, with no vision of what happens around it. On the other hand, planning individual and collective advanced mobility applications as well as providing detailed analysis of urban and suburban areas require additional and more complex information [6, 10]. Raw mobility data like GPS positioning describes elementary events (position, acceleration, etc.), while any proper modeling requires a higher-level vision of what is happening to the user, to other users living in the surroundings, and to the environment in which the user is moving. Such higher-level modeling should provide some clear categorization, annotation or semantics for locations and/or movements. Recognizing different individual mobility behaviors abstracting at a higher comparable level and making them explicit is mandatory for enabling novel applications or empowering existing ones.

Many studies semantically enrich mobility data with annotations about human activities and build individual mobility models on top of that. For instance, these approaches estimate home/work locations of an individual by analyzing the frequency of visits in a particular place [14], observing the sequence of movements to derive the sequence of activities [15], their semantic [19] and possible different aspects [3], or extract network-based personal data-driven models named Individual Mobility Network (IMN) capturing the structured patterns of visits to locations [21]. The existing works either focus on building mobility models or on adding semantics from external data sources. On the other hand, in this work we exploit the mobility data used to build the individual model also to annotate the model itself.

In particular, we advance IMNs proposing a data-driven procedure for extracting Annotated Individual Mobility Networks (AIMNs). A limitation of actual IMNs is that they are indeed "individual" and not easily comparable. Following [1], our goal is to add semantics to the locations modeling the nodes of the IMNs in order to make them comparable among different users. We accomplish this task by designing a procedure considering individual, collective, and contextual features for turning locations into annotated locations. The location annotation procedure, besides differentiating the locations with respect to the different features, also provides a hierarchy showing the reasons for the different annotation categories. The procedure is basically a two-step-clustering. The first step applies a distance-based clustering to group the different locations. The second step exploits a hierarchical approach for further summarizing the locations, and for better describing them according to features characterizing the annotation categories. This provides to the annotated locations a higher expressiveness and allows to generalize IMNs by making AIMNs comparable. Therefore, the analysis on AIMNs allows to segment the users moving on different areas by means of the data-driven semantics provided by the area itself.

We employ the proposed methodology on a dataset of trucks moving in Greece, Albania, and other EU countries. We focused our analysis on two areas with a different size finding in both cases six types of locations for annotating the IMNs of the vehicles. The first main differences between the location types are the number of stops in a location, the number of links with other individual locations, and the average arrival/leaving times. Then, the locations differ on their centrality with respect to individual locations of other users and with respect to existing points of interest like gas stations, parking areas, shops, supermarkets, etc. A preliminary analysis of the AIMNs highlights that in each area there are distinct types of users (trucks, in our study), that visit the various types of location with different frequencies. Low entropy trucks visit frequently the same type of annotated location, while high entropy trucks have a central node from which they reach all the differently annotated locations.

The rest of the paper is organized as follows. Section 2 summarizes related work on locations annotation. In Section 3 we recall individual mobility networks and further concepts for understanding the procedures for locations annotation and annotated individual mobility networks extraction described in Section 4. Section 5 presents experiments in the form of a case study in which we employ the proposed methodology. Finally, Section 6 concludes the paper and illustrates future research directions.

2 RELATED WORK

In the literature, various works address the problem of describing and characterizing visited locations for modeling mobility behavior and for describing land use. In [4, 5] it is presented a technique to determine land uses in specific urban areas based on tweeting patterns, and to identify points of interest as places with high activity of tweets. Human activities and geographical areas are modeled in [17] by means of Foursquare place categories. A spectral clustering algorithm is used on areas and users for identifying user communities that visit similar categories of places, and for comparing urban neighborhoods within and across cities. Semantic information attached to places could be used for location-based applications. Indeed, in [18], both Foursquare and cellular data are exploited to infer user activities in urban environments. The authors employ user communication patterns to predict the activity of Foursquare users who check-in at nearby venues adopting a machine learning approach. In [2] it is presented a location-based and preference-aware recommender system considering user preferences learned from location history with a predefined weighted category hierarchy (from Foursquare), and social opinions mined from location histories. Similarly, in [29], Foursquare check-in category information is exploited to model user's movement patterns and to predict the category of user activity at the next step by means of a mixed hidden Markov model. In order to describe and characterize the locations, the aforementioned works adopt social network data like Twitter and Foursquare. None of them use GPS positioning as it is not sufficiently informative for their purpose. On the opposite, our approach extracts the location categorization through data mining and location modeling applied to GPS trajectories.

In the following, we illustrate works modeling human mobility from GPS data but not explicitly characterizing locations with respect to activities at locations or neighborhood description. In [25] it is introduced the *mobility profile* of a user as the set of her routine trips, i.e., a set of very systematic and repetitive movements. On the other hand, in [9] the authors focus on the locations and points of interest by developing a parameter-free method for extracting individual locations with a data-driven approach. Combining these approaches, in [12, 21] the authors define individual mobility networks (IMNs) for modeling all the salient aspects of individual mobility. An IMN describes the mobility of an individual through a graph representation of her locations and movements, grasping the relevant properties and removing unnecessary details. In [1] the authors define a general approach to use state-transition graphs (STGs) in movement analysis. A STG is an aggregate representation of a behaviour by a directed weighted graph, where the nodes stand for the possible states and the edges for the transitions between the states. Information from the states and activities is gathered from external sources like land allocations or presence of points of interest. The main difference between IMNs and STGs lies in the fact that the first one models the mobility and it is more attached to geographic components, while the second one models the events, states or activities that can occur one after another.

Modeling individual behavior is a precious task as, besides providing a succinct and understandable representation of the mobility patterns of the users, it enables the development of applications like carpooling [11], or trajectory prediction [26]. As a consequence, an individual model with enhanced location descriptions can be undoubtedly beneficial and informative.

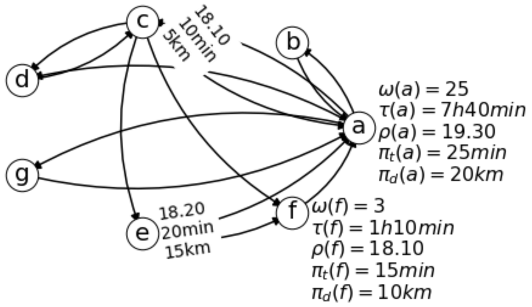

Figure 1: IMN extracted from the mobility of an individual. Edges represent the existence of a route between the locations. The functions characterize each location.

3 SETTING THE STAGE

In the following, we introduce the definitions of *trajectory* [25] and *individual mobility network* [12, 21], useful for understanding the rest of the paper. We adapt them to the needs of the problem we are facing and the approach designed to solve it.

Definition 3.1 (Trajectory). A *trajectory* is a sequence $t = \langle p_1, \ldots, p_n \rangle$ of spatio-temporal points, each being a tuple $p_i = (x_i, y_i, z_i)$ that contains latitude x_i, longitude y_i and timestamp z_i of the point. The points of a trajectory are chronologically ordered, i.e., $\forall 1 \le i < n : z_i < z_{i+1}$.

Given a trajectory t we refer to its i-th point p_i with the notation $t[i]$, and to its number of points with $t.n$. Also, we indicate the longitude, latitude and timestamp components of point $t[i]$ respectively with the notation $t[i].x$, $t[i].y$, and $t[i].z$.

Definition 3.2 (Individual History). Given a user u, we indicate with $H_u = \langle t_1, \ldots, t_n \rangle$ the *individual history* of user u as the set of trajectories traveled by u in a certain time period.

Given the individual history H_u of user u, we can extract from it the *individual mobility network* (IMN) G_u. An IMN describes the individual mobility of a user through a graph representation of her locations and movements, grasping the relevant properties of individual mobility and removing unnecessary details.

Definition 3.3 (Individual Mobility Network). Given a user u, we indicate with $G_u = (L_u, M_u)$ the *individual mobility network* of user u, where L_u is the set of nodes and M_u is the set of edges. On the nodes and edges we define the following functions:

- $\omega : L \longrightarrow \mathbb{N}$ returns the number of stops in location $l \in L_u$;
- $\tau : L \longrightarrow \mathbb{R}$ returns the *typical* time spent in location $l \in L_u$;
- $\rho : L \longrightarrow \mathit{Time}$ returns the *typical* time of arrival of u along all the movements $m \in M_u$ reaching location $l \in L_u$.
- $\pi_t : L \longrightarrow \mathbb{R}$ returns the *typical* time travelled by u along all the movements $m \in M_u$ reaching location $l \in L_u$.
- $\pi_d : L \longrightarrow \mathbb{R}$ returns the *typical* distance travelled by u along all the movements $m \in M_u$ reaching location $l \in L_u$.

Nodes represent locations L_u and edges represent movements M_u between locations. With the term *typical* we refer to using aggregating function like mean and median, also including the associated dispersion indexes like standard deviation. To clarify the concept of IMN, let us consider Figure 1. It describes the IMN extracted from the mobility of an individual who visited seven distinct locations. Location (a) has been visited by user u for a total of $\omega(a) = 25$ times, i.e., 25 stops, with a typical stay of $\tau(a) = 7h40min$. On the other hand, user u stopped in location

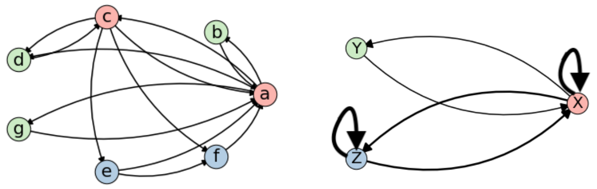

Figure 2: AIMN extracted from the IMN of Figure 1. Annotations are represented with colors. The first graph reports the annotated nodes, while the second graph is the AIMN with the novel annotated nodes and edges.

(f) for $\omega(f) = 3$ times with a typical stay of $\tau(f) = 1h10min$. Edges (c, f) and (e, f) lead to location f, typically arriving at time $\rho(f) = 18.10$, traveling $\pi_t(f) = 15min$, and $\pi_d(f) = 10km$.

The computation of an IMN G_u starts from the ordered sequence history H_u of user u. Locations are obtained by aggregating the origin and destination points of the trajectories using the TOSCA location clustering algorithm [9, 12]. A location identifies a set of points, and a *location prototype* is the point minimizing the distance with the other observations part of the location.

4 PROPOSED APPROACH

In this section, we describe the location annotation procedure and how it is applied to extract annotated individual mobility networks (AIMNs) from individual mobility networks (IMNs).

4.1 Annotated Individual Mobility Network

An *annotated individual mobility network* (AIMN) extracted from an IMN describes the individual mobility of a user through a very simple graph representation of *annotated* locations and movements, summarizing the relevant properties of the IMN and compressing the information contained in the mobility model.

The extraction of an AIMN $\overline{G}_u = (\overline{L}_u, \overline{M}_u)$ of a user u starts from her IMN $G_u = (L_u, M_u)$, and an annotation function λ. Given an individual location $l \in L_u$, the annotation function $\lambda : L \longrightarrow \mathbb{N}$ returns its annotation as a consequence of a learning process. Two locations l_1, l_2 have the same annotations, i.e., $\lambda(l_1) = \lambda(l_2)$, if they are "similar". Details of the location annotation procedure that returns λ and the meaning of *similar locations* are presented in the next section.

The AIMN $\overline{G}_u$ is built by mapping each node $l \in L_u$ to a node $\bar{l} = \lambda(l)$ corresponding to its annotation, and by merging the corresponding edges. Thus, locations with the same annotation are collapsed into the same annotated location. More in detail, given an edge between l_1, l_2 if they have a different annotation $\lambda(l_1) \neq \lambda(l_2)$, then two annotated locations $\bar{l}_1, \bar{l}_2$ are created together with the movement-edge connecting them. On the other hand, if both l_1, l_2 have the same annotation $\lambda(l_1) = \lambda(l_2)$, then a unique annotated location $\bar{l}$ is created together with a self-loop.

We exemplify the concept of AIMN through Figure 2. Given a certain λ function, the first graph reports the nodes/locations annotation of the graph in Figure 1, where the annotations are represented through colors. Nodes having the same colors are described by similar features, i.e., are similar according to some aspects. The second graph illustrates the AIMN obtained by collapsing nodes with the same annotation into an annotated node, merging edges and creating self-loops.

Name	Description
lat, lon	location prototype coordinates
next locs	number of different subsequent locations
radius	radius of gyration
entropy	entropy of ingoing/outgoing movements
stops	total and categorized number of stops*
staytime	typical† stay time*
arrival times	typical† arrival times*
duration/length	typical† duration/length of movements*
is regular	if a location is frequently visited

Table 1: Individual Features. ($*$) indicates that the features are calculated with respect to weekdays vs. weekend, and daytime (from 7 am to 8 pm) vs. nighttime. ($\dagger$) indicates features for both mean and standard deviation.

Name	Description
r-exclusivity	number of stops within r km w.r.t. other vehicles
r-centrality	number of locations in a radius of r km
k-distance	the distance of the k^{th} nearest locations

Table 2: Collective Features. They are calculated comparing an individual location with other individual locations.

Name	Description
POI r-centrality	number of different POIs within a radius of r km
POI k-centrality	number of different POIs within the k^{th} nearest POI
POI k-distance	distance to closest POIs within the k^{th} nearest POI

Table 3: Contextual Features. All measures are also computed separately over nine different categories of POIs: gas stations, parking areas, piers, hotels, food shops, leisure activities, (no-food) shops, services, supermarkets.

4.2 Location Annotation Procedure

In this section, we describe the location annotation procedure we designed to obtain a non-trivial annotation function $\lambda : L \longrightarrow \mathbb{N}$. As previously discussed, the aim of λ is to provide the same annotation for similar locations. Hence, a definition of *similarity* must be provided in order to design λ. Given an IMN G_u of user u, in order to compare two locations with a distance function, every location $l \in L_u$ must be described with a set of attributes (see [12, 21] for more details). Individual locations represented as a vector of features can be grouped using clustering algorithms [24]. As a consequence, λ is the function which annotates each location with the cluster label to which the location belongs to.

Simple λ functions can be obtained by modeling an individual location through a basic set of attributes, for instance using features like latitude and longitude, i.e., spatial features, or the average stay time and arrival time, i.e., temporal features. However, such basic representations of a location can be insufficient to capture various aspects related to *(i)* the mobility behavior of the individual user, *(ii)* the user behavior in relationship with the mobility behavior of other users, *(iii)* the user behavior in relationship with the geography and context in which she moves.

To overcome these weaknesses, we design the following location annotation procedure that takes as input a set of IMNs $\mathcal{G} = \{G_1, \ldots, G_n\}$, a context C where the users move and returns an annotation function λ. The procedure works as follows. For each IMN $G_u \in \mathcal{G}$, for each individual location $l \in L_u$, we describe l with a vector of features capturing three distinct aspects.

- *Individual features.* These features characterize an individual location l only with the mobility of the vehicle u stopping in l. The list attributes is shown in Table 1. Some features are calculated with respect to weekdays vs.

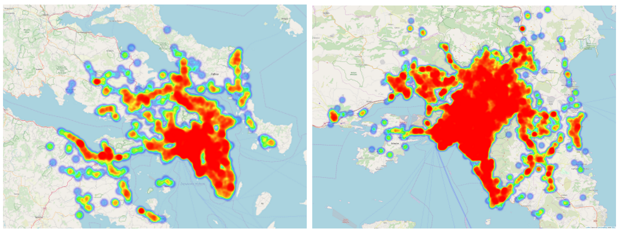

Figure 3: Heatmaps of the stop points in the areas analyzed in Greece: Inter-regional (left), and Urban (right).

weekend, and daytime (from 7 am to 8 pm) vs. nighttime. For features describing aggregates we report mean and standard deviation. A location is *regular* if it is frequently visited more than the others (see [12] for details).

- *Collective features.* These features characterize an individual location l with respect to all the other locations both of user u and of the other users in $\mathcal{G}$. Details in Table 2.
- *Contextual features.* These features characterize an individual location l with respect to facilities, i.e., different types of points of interests (POIs) in the surrounding areas given by the context C. Table 3 reports a detailed description.

Given the locations described by these features, we implement the annotation function by means of a clustering algorithm. Inspired by [8, 9], instead of using a simple clustering approach, we design a two steps clustering allowing to simultaneously summarize the different locations and obtain a hierarchy for better describing them. In particular, we use K-Means clustering algorithm [16] for the first clustering phase. As illustrated in the next section, we observe that we need a high k (between 100 and 200) in order to have a good clustering with respect to internal evaluation measures like the Sum of Squared Error (SSE). The clustering with K-Means allows to consistently reduce the number of different location prototypes. The second phase is aimed at further reducing the number of clusters for keeping simple the annotation computed by λ, and simultaneously obtaining a hierarchy for the different location prototypes. To this purpose we adopt a hierarchical clustering approach with the Ward's criterion [28] to determine the clusters to be merged. A visual and interactive inspection of the resulting dendogram and centroids describing the clusters allows to understand the reasons [20] for having different types of annotated locations and which is a reasonable number of clusters representing the different types of annotated locations in an AIMN.

5 EXPERIMENTS

In this section, we present a case study on a dataset of trucks in which we employ the proposed methodology[1]. First, we briefly analyze the dataset. Then, we provide details for the location annotations and the clustering results. Finally, we show some preliminary analysis enabled by the extracted AIMNs.

5.1 Dataset Description

We analyze a dataset of about 15 million of trajectories of trucks moving in Greece, Albania, Cyprus, and other few EU countries from July 2017 to June 2018. There are different kinds of trucks, depending on the size, number of carts, etc. We focus on the following types of trucks which are the most frequent in the dataset: Van, Truck 3, Truck 3 ax., Flatbed Truck, Truck[2].

Truck Type	Inter-regional Area	Urban Area
Van	214,129 (71.03%)	105,065 (75.24%)
Truck 3	82,697 (27.43%)	33,353 (23.88%)
Truck 3 ax.	3,478 (1.15%)	1,223 (0.88%)
Flatbed Truck	978 (0.32%)	-
Truck	179 (0.05%)	-

Table 4: Distribution of different type of vehicles for the two areas analyzed (percentages in brackets).

Measure	Inter-regional Area	Urban Area
Traj per Vehicle	419.86 ± 255.75	498.72 ± 278.72
Avg Length	10.40 ± 14.17	6.64 ± 9.65
Avg Duration	23.85 ± 23.82	19.25 ± 19.85

Table 5: Mean and standard deviation of some descriptive statistics for the trajectories in the two areas analyzed.

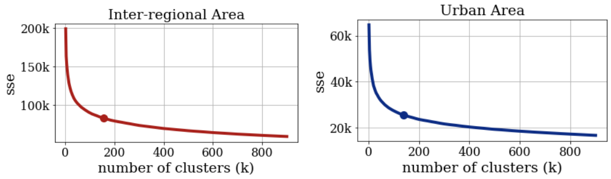

Figure 4: K-Means Sum of Squared Error (SSE) varying the number of clusters k. We selected $k = 155$ for the inter-regional area, and $k = 140$ for the urban area.

Analyzing all the vehicles together would not lead to a fair comparison due to the very different types of trajectories followed by the trucks. For instance, we cannot compare the IMN of a truck performing daily deliveries in a radius of 20 km with the IMN of a truck moving across regions or countries along very long trajectories. Therefore, we partition the trajectories of the dataset through a simple, K-means-like iterative procedure. First, each vehicle u is associated to the geographical bounding box r_u of its trajectories in H_u, and the set of areas A is initialized to $\emptyset$. Then, we iteratively consider each vehicle u and compare its r_u with all the existing areas $a_i \in A$. If, for at least 75% of the vehicles $v \in a_i$, $r_u \cap r_v \neq \emptyset$ and $1/4 \leq area(r_u)/area(r_v) \leq 4$, then u is added to a_i; otherwise, if $\nexists a_i \in A$ satisfying this condition, a new area a_j is created and u is added to a_j. Finally, we check that each vehicle belongs to the area with the highest overlap, and in case we move a vehicle between two areas until convergence. The above procedure recognizes twelve different areas.

In the following, we focus on two areas depicted in Figure 3. We name the first *inter-regional area* since it contains the trajectories of vehicles moving in various regions of Greece. On the other hand, the second is an *urban area* containing the movements of trucks in Athens. Details about the number of different vehicles can be found in Table 4. In addition, in Table 5 we report basic statistics of the trajectories for the two different areas. We notice how in the urban area a vehicle performs on average much more trajectories than a vehicle in the inter-regional area. As expected, the trajectories in the inter-regional area are on average longer than those in the urban area. Moreover, the high standard deviation means that the trajectories in the inter-regional area are also more variegate than those in the urban area. The similar average duration might be due to the higher traffic in the urban area, that makes the speed lower than in the inter-regional area.

[1] The code for performing the analysis is available at: https://github.com/fsbolgi/AnnotatedIndividualMobilityNetwork. The dataset is not publicly available.

[2] More details are available at https://trackandknowproject.eu/about/deliverables/.

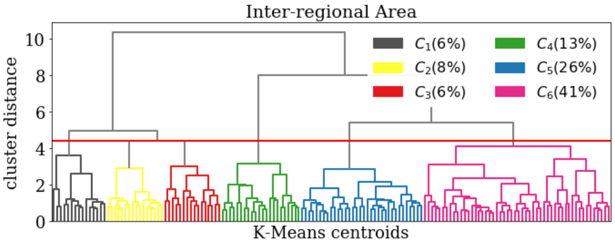

Figure 5: Inter-regional Area: hierarchical clustering dendogram with a cut at six clusters (clusters size in brackets).

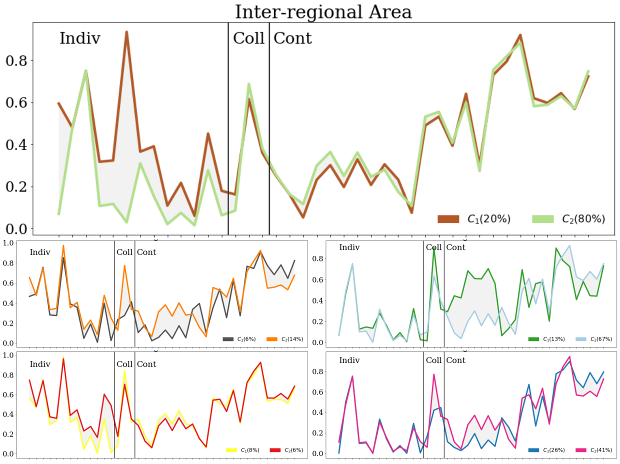

Figure 6: Inter-regional Area: centroids parallel plots showing discriminant features for dendogram in Fig. 5.

5.2 Annotated Locations Clustering Analysis

Using the procedure described in [12], we extract the IMNs for the vehicles in both areas, and we use them as input for the proposed methodology. In particular, in our analysis, we focus on September and October 2017, obtaining 883 IMNs in the inter-regional area and 373 IMNs in the urban area. Statistics describing the IMNs are reported in Table 7. As context C, we exploit a dataset containing the POIs of the whole Europe[3]. We extracted only the POIs relative to the areas we are interested in and we restricted to some more general and relevant categories, namely: *gas*, *parking*, *pier*, *hotel*, *food*, *leisure*, *(no-food) shop*, *service*, *supermarket*.

Given the IMNs for the two areas, the location annotation procedure received as input about 110k locations for the inter-regional area and about 39k locations for the urban one. We described each location with the individual, collective, and contextual characteristics illustrated in Section 4.2, ending in a vector of 72 features. For r-exclusivity we adopt $r = 0.2$ km, for r-centrality $r \in \{1, 5, 15\}$ km, for k-distance $k \in \{1, 3, 5, 8, 10, 20\}$. For POI r-centrality, POI k-centrality and POI k-distance we adopt $r = 0.5$ km and $k = 30$, respectively. Before running the location annotation procedure, we performed a correlation analysis using the Pearson correlation coefficient. This step allowed us to reduce the number of features to 55. In particular, *(i)* the total number of stops is removed in favor of next loc, entropy and number of daytime weekday stops; *(ii)* among the temporal aggregations of stay times, arrival and leaving times, only the weekday vs weekend is considered; *(iii)* movement duration features are removed since they are strongly correlated with movement length features; *(iv)* we remove r_5-centrality and k_1-centrality, k_5-centrality, k_{20}-centrality as they are highly correlated with the remaining ones.

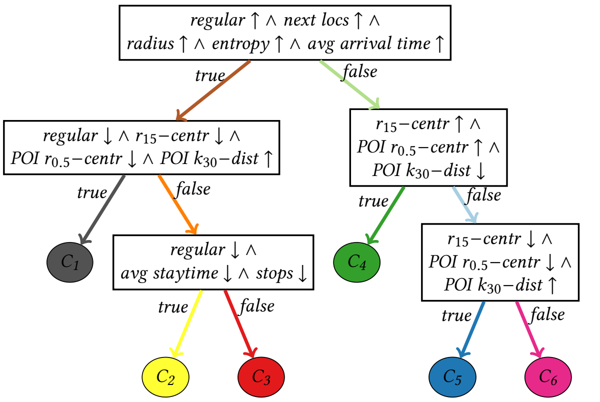
regular ↑ ∧ *next locs* ↑ ∧
radius ↑ ∧ *entropy* ↑ ∧ *avg arrival time* ↑
true
false
regular ↓ ∧ r_{15}−*centr* ↓ ∧
POI $r_{0.5}$−*centr* ↓ ∧ *POI* k_{30}−*dist* ↑
r_{15}−*centr* ↑ ∧
POI $r_{0.5}$−*centr* ↑ ∧
POI k_{30}−*dist* ↓
true
false
true
false
C_1
regular ↓ ∧
avg staytime ↓ ∧ *stops* ↓
C_4
r_{15}−*centr* ↓ ∧
POI $r_{0.5}$−*centr* ↓ ∧
POI k_{30}−*dist* ↑
true
false
true
false
C_2
C_3
C_5
C_6

Figure 7: Inter-regional Area: Tree showing the most discriminant features for the dendogram in Figure 5.

We ran K-Means with $k = 2, \ldots, 900$. The visual inspection of the Sum of Squared Error (SEE) in Figure 4 suggested to select $k = 155$ for the inter-regional area, and $k = 140$ for the urban area. The subsequent run of the hierarchical clustering with the Ward's criterion yields the dendograms in Figures 5 and 8. In both cases, we cut the dendograms to obtain *six* clusters that characterize the description of the different annotated locations with a good trade-off between a sufficiently high level of abstraction and a detailed specification. We observe that in both cases, more than 50% of the locations end in the rightmost part of the dendogram, making it slightly imbalanced. That produces some small clusters, yet none of them is negligible in terms of size. In the following, we combine the hierarchy of the dendogram with the information returned by the parallel plots of the centroids of the clusters for each split. This visualization, due to the interpretable features [13] describing the locations, allows to explain [20] the hierarchy and consequently the annotations of the various clusters $C_1, \ldots, C_6$.

In order to better understand the reasons that led to differentiate the locations into the six clusters – and therefore understand what each cluster contains –, we show in Figures 6 and 9 the corresponding parallel plots of the most significant features, respectively for the inter-regional and urban areas; we also summarize in Figures 7 and 10 the insights we obtained through inspection of the features at different levels of the dendogram, by means of a decision tree representation.

In both dendograms the first split is a consequence of differences relative to individual features. The first split (Figures 6 and 9 top), using the individual features, separates on the left branch regularly visited locations (↑ *regular*) from which is possible to reach various destinations (↑ *next locs* and ↑ *entropy*) with early arrivals and leavings (↑ *avg arrival times*, i.e., the vehicle leaves before 8 am and is back before 7 pm) in a location defined by a not very specific area (↑ *radius*), from the other locations (right branch). Thus, with respect to our case study, the locations on the left-hand side can be matched with storage points and/or deposits of the trucks analyzed, while those on the right are all the others. This consideration can also justify the fact that the majority of locations lie in the rightmost part of the dendogram. Moving forward on the left branch, we have a different split for the inter-regional area and for the urban area.

For the inter-regional area (Figure 6 center left) the second split, making use of collective and contextual features, separates not regular locations (*regular* ↓) not close to individual locations

[3] The POIs are points collected from OpenStreetMap filtered based on Geofabrik's taxonomy of OpenStreetMap features, i.e., points with the label "POI" are kept.

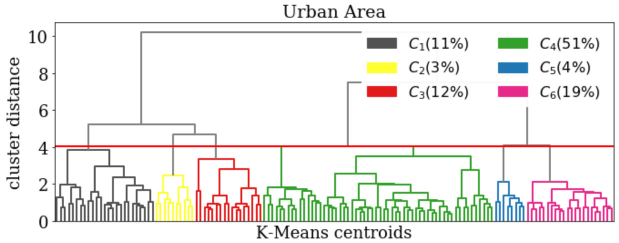

Figure 8: Urban area: hierarchical clustering dendogram with a cut yielding six clusters (clusters size in brackets).

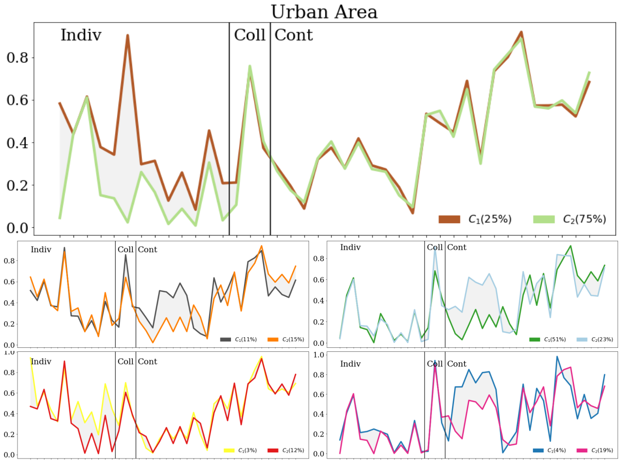

Figure 9: Urban Area: centroids parallel plots showing discriminant features for dendogram in Figure 8.

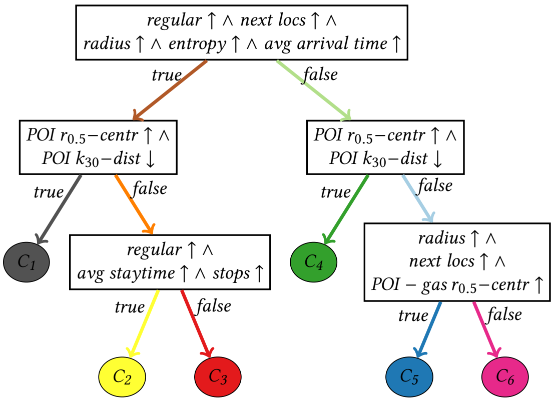

Figure 10: Urban Area: Tree showing the most discriminant features for the dendogram in Figure 8.

of other vehicles (r_{15}−*centrality* ↓, i.e., not central w.r.t. the others) nor to POIs (*POI* $r_{0.5}$−*centrality* ↓ and *POI* k_{30}−*distance* ↑) from the rest. Cluster C_1 models suburban and peripheral storage points. On the other hand, the subsequent split (Figure 6 bottom left) identifies less frequent locations with shorter stops (cluster C_2) and more frequent locations with longer stops (cluster C_3). The right branch of the inter-regional tree in Figure 7 performs a symmetric split with respect to the left branch. Thus, in cluster C_4 we find not regularly visited locations surrounded by many other personal locations and POIs (r_{15}−*centr* ↑, *POI* $r_{0.5}$−*centr* ↑, *POI* k_{30}−*dist* ↓). Due to the high density, we can infer that these locations are central regions of the inter-regional area. Finally, clusters C_5 and C_6 containing most of the locations are placed in suburbans regions far away from many POIs.

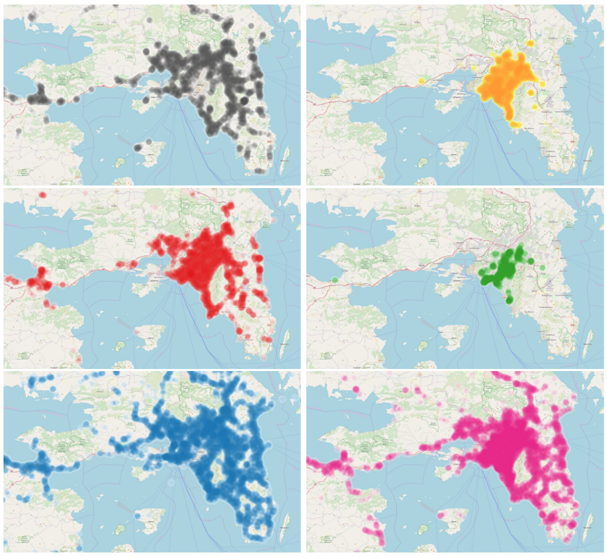

Figure 11: Heatmaps of the different clusters for the annotated locations of the Inter-regional area. From left to right, top to bottom, clusters $C_1, C_2, \ldots, C_6$.

Following the same logic used to describe the splits of the inter-regional area we can understand the splits for the annotations of the urban area from Figures 9 and 10. The second split (Figure 9 center left) discriminates between locations close to various existing POIs (*POI* $r_{0.5}$−*centr* ↑ and *POI* k_{30}−*dist* ↓). Since the closest categories for the black lines are *food*, *leisure*, *shop*, *service* and *supermarket* we can assume that cluster C_1 mainly identifies "regular" locations in the city center. A further split on individual features (Figure 9 bottom left) identifies more regularly visited locations (↑ *regular*) with a high stay time and number of stops during both weekend and weekday (↑ *avg staytime*, ↑ *stops*). The other branch separates the locations based on the distance to existing POIs (*POIr*$_{0.5}$−*centr* ↑ and *POIk*$_{30}$−*dist* ↓). This time there is a clear separation for all the contextual features also considering *gas*, *parking* and *hotel*. Hence, the cluster C_4 captures central locations (probably even more central than those on the left branch of the tree) visited sporadically by the vehicles analyzed. This cluster is the biggest in the urban area. Finally, the last split relative to suburban locations distinguish cluster C_5, containing suburban locations with a large radius from which can be reached other many individual locations but far away from POIs except gas stations (probably located into an industrial area), from cluster C_6, which is formed by suburban locations close to facilities but reached sporadically.

In Figure 11 we show the heatmaps of the locations for the various clusters for the Inter-regional area. We can notice how the position of the different annotated locations on the maps is coherent with the descriptions reported above. From a very high level we can summarize the distinction between the different annotated locations as delivery "origins" (C_1, C_2, C_3) and "destinations "(C_4, C_5, C_6), i.e., recipients. Then, the clusters distinguish according to the closeness with respect to other individual locations and POIs, to usage in terms of stay times and arrival times. For instance, we have locations in the city center in C_4, isolated suburban locations C_5, and suburban locations surrounded by POIs in C_6. Moreover, we highlight that latitude and longitude are not crucial for distinguishing the various clusters at the final level of the dendogram. Indeed the points are not entirely separated from a spatial point of view. This implies that the other

Measure	Inter-regional Area	Urban Area
IMN vs TMP	0.1580	0.1465
IMN vs SPT	0.1186	0.0723
TMP vs SPT	0.0036	0.0136

Table 6: NMI comparing the proposed location annotation procedure based on IMNs against a temporal annotation (TMP) and a spatial annotation (SPT) procedure.

Measure	Inter-regional Area	Urban Area
Nodes	96.84 ± 74.98	138.96 ± 112.04
Edges	270.47 ± 190.32	312.21 ± 224.90
Density	0.07 ± 0.15	0.05 ± 0.10
Degree	5.78 ± 1.46	4.85 ± 2.39
Clus. Coef.	0.19 ± 0.13	0.17 ± 0.13

Table 7: IMNs characteristics (mean ± std dev).

Measure	Inter-regional Area	Urban Area
Nodes	5.38 ± 0.89	5.06 ± 1.02
Edges	15.12 ± 4.69	13.31 ± 4.56
Density	1.28 ± 0.28	1.28 ± 0.32
Degree	5.48 ± 1.16	5.10 ± 1.10
Clus. Coef.	0.82 ± 0.22	0.85 ± 0.20

Table 8: AIMNs characteristics (mean ± std dev).

features are much more relevant than the geographical aspects for capturing different characteristics.

Finally, we adopt a clustering evaluation measure to show that the proposed location annotation procedure instantiated with the truck case study is meaningful and not trivial. We implement two simple annotation functions λ. The first procedure (SPT) describes the locations only using spatial features, i.e., latitude and longitude. The second procedure (TMP) describes the locations only using the average stay time and arrival time. In both procedures, the locations are grouped and annotated using K-Means with $k = 6$, i.e., the same number of annotated locations discovered with the proposed procedure. We compare the annotations, i.e., the clustering results, returned by the three procedures using the Normalized Mutual Information score [27]. The more similar are the annotations between two location annotation procedures, the higher is the NMI score. The very low NMI scores reported in Table 6 confirms that it is not possible to design an annotation procedure similar to the proposed one with a trivial approach.

5.3 Inspection of Annotated IMNs

Given the locations annotation described in the previous section, we know the behavior of the annotation function λ and the meaning of the different annotations. Thus, given a location l with the vector of features describing it, λ assigns l to the most similar cluster with respect to the six prototypes represented by the centroids reported in Figures 6 and 9, i.e., $\lambda : L \longrightarrow [1, \dots, 6]$.

Using the annotation functions λ for the IMNs of the two areas we obtain the corresponding AIMNs. We report in Table 8 statistics describing the AIMNs. By comparing Table 8 with Table 7 we notice that the number of nodes and edges obviously drops. On the other hand, we observe that AIMNs are much denser than IMNs and with a higher average degree about 5 with a standard deviation of 1.1, meaning that typically each vehicle from an annotated location can reach at least 3 other annotated locations. Consequently, AIMNs show a much higher clustering coefficient.

Moreover, we exploit the AIMNs for a preliminary analysis aimed at comparing vehicles. From the AIMNs we model each truck with a vector $v_u = \langle f_1, f_2, \dots, f_6 \rangle$ of six elements containing the relative number of stops in each type $C_1, C_2, \dots, C_6$ of annotated locations. Using these vectors of relative frequencies f_i, we calculate for each user the *purity* and (normalized) *entropy* [22] as $purity(v) = max_{i \in [1,6]}(f_i)$, and as $entropy(v) = -\sum_{i \in [1,6]} f_i \log_2(f_i) / \log_2(6)$. We highlight that purity and entropy capture two different aspects. We report the distributions of purity and entropy for the inter-regional and urban areas in Figure 12. For the purity (top row), we observe two normal-like distributions with most of the users having a purity of 0.5 in both areas meaning that half of the stops of the vehicle refer to the same annotated location. On the other hand, there are also trucks with a purity close to 1.0, meaning that almost all the stops belong to the same annotated locations. There are no vehicles having a purity lower than 0.2. With respect to entropy (bottom row), we observe right-skewed distributions with a mean of about 0.7, showing that, despite there is an average high purity, most of the trucks stop in different annotated locations. In addition, there are very few vehicles with an entropy close to 0.0 signaling that a truck generally visits a minimum number of different annotated locations. Finally, we report that we found no relationships between the type of truck and entropy or purity.

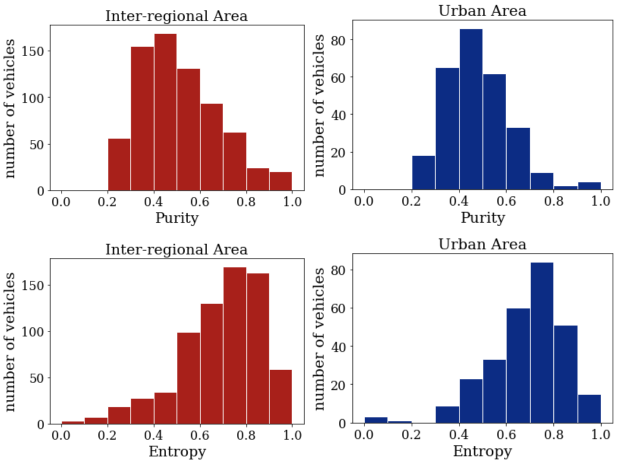

Figure 12: Purity and entropy distributions for number of stops in different annotated locations.

We report in Figure 13 the AIMNs of three trucks moving in the urban area with different levels of purity and entropy. The leftmost AIMNs show a high entropy and low purity, visiting all the types of annotated locations with a balanced number of stops, i.e., similar sizes. We notice on the map how it covers a larger area compared to the other AIMNs. The big yellow node C_2 models the parking/storage while the others model different types of destinations. On the other hand, the rightmost AIMN shows a low entropy and high purity: the majority of the stops are on annotated locations of type C_3 (red), which are on the north-west and south-east in the map. Finally, the central AIMN shows a typical vehicle with a medium level of entropy and purity.

6 CONCLUSION

We have proposed a data-driven procedure for annotating individual locations, and we have employed it to extract annotated individual mobility networks (AIMNs) from individual mobility networks (IMNs). A case study experimentation on a real dataset of fleet moving in Greek areas has shown the effectiveness of the proposed approach. We have found a hierarchical structure

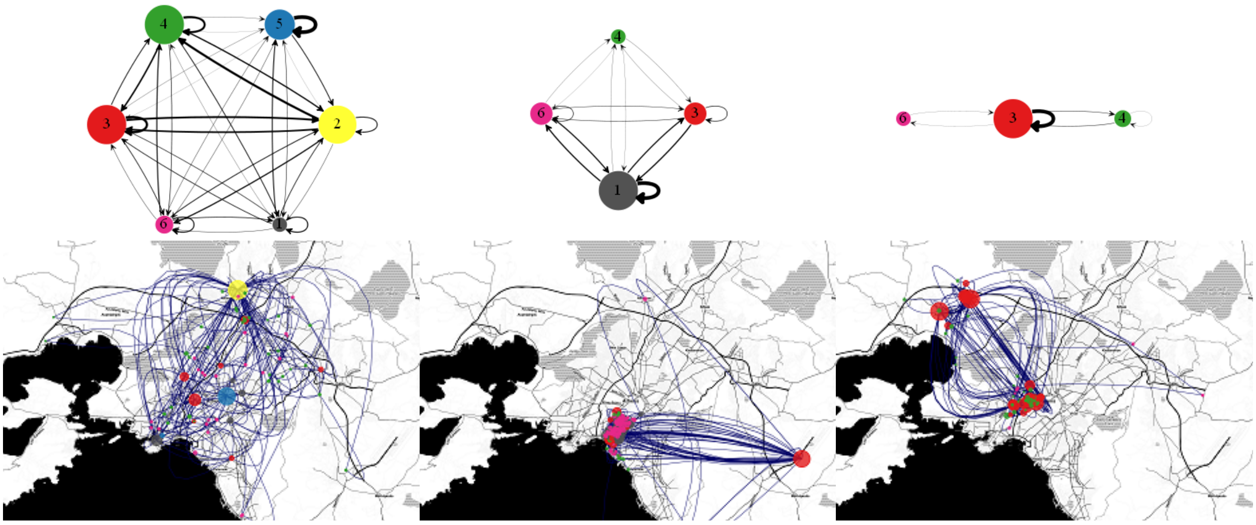

Figure 13: AIMNs for three vehicles moving in the urban area with different purity and entropy levels. Left low purity and high entropy. Center medium purity and entropy. Right high purity and low entropy. The top row reports the AIMNs while the bottom row contains the corresponding AIMNs with a spatial disaggregation of the annotated locations on a real map. In both cases the higher is the size of the node/location, the higher the number of stops in the locations.

describing the annotated locations through individual, collective, and contextual features. The principal discrimination is based on the frequency of visits, length of stay, and centrality with respect to other locations and existing points of interest. As a consequence of the annotation, we can observed different vehicles studying the corresponding AIMNs. Such information can be applied for a detailed analysis of inter-regional and urban areas and for planning ad-hoc and personalized mobility applications.

Future research directions can purse various goals. First, we would apply the proposed methodology to different data sources. For instance, observing the personal mobility of individual car drivers, or of users adopting different types of means of transportations like bicycle, foot, public services, would probably lead to identify other types of annotations for the locations. Second, we would like to perform a deeper analysis of the users by clustering the AIMNs with respect to the annotated locations for discovering relevant users' segmentation and geographical characterization of the territory. Third, we would like to use the annotated locations and the mobility history for building sequences of "states" [1]. Exploiting sequential pattern mining algorithms would allow us to search for mobility patterns represented as sequences of different annotated locations. Fourth, we would like to analyze the benefits of integrating our concepts of annotated locations and AIMNs into visualization platforms [1, 23] for semantic annotation of individual mobility.

ACKNOWLEDGMENTS

This work is partially supported by the E.C. H2020 programme under the funding scheme *Track &Know*, G.A. 780754, https://trackandknowproject.eu/. We thank Haosheng Huang and Cheng Fu from University of Zurich for sharing the POIs dataset.

REFERENCES

[1] Natalia Andrienko and Gennady Andrienko. 2018. State transition graphs for semantic analysis of movement behaviours. *IV* 17, 1 (2018), 41–65.
[2] Jie Bao et al. 2012. Location-based and preference-aware recommendation using sparse geo-social networking data. In *SIGSPATIAL*. ACM, 199–208.
[3] Ronaldo Dos Santos Mello et al. 2019. MASTER: A Multiple Aspect View on Trajectories. *TGIS* 12526 (2019).
[4] Vanessa Frias-Martinez et al. 2012. Characterizing urban landscapes using geolocated tweets. In *PASSAT*. IEEE, 239–248.
[5] Vanessa Frias-Martinez and Enrique Frias-Martinez. 2014. Spectral clustering for sensing urban land use using Twitter activity. *EAAI* 35 (2014), 237–245.
[6] Fosca Giannotti et al. 2011. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *VLDB* 20, 5 (2011), 695–719.
[7] Marta C Gonzalez et al. 2008. Understanding individual human mobility patterns. *Nature* 453, 7196 (2008), 779.
[8] Riccardo Guidotti et al. 2014. Retrieving points of interest from human systematic movements. In *SEFM*. Springer, 294–308.
[9] Riccardo Guidotti et al. 2015. TOSCA: two-steps clustering algorithm for personal locations detection. In *SIGSPATIAL*. ACM, 38.
[10] Riccardo Guidotti et al. 2016. Unveiling mobility complexity through complex network analysis. *SNAM* 6, 1 (2016), 59.
[11] Riccardo Guidotti et al. 2017. Never drive alone: Boosting carpooling with network analysis. *IS* 64 (2017), 237–257.
[12] Riccardo Guidotti et al. 2017. There's a path for everyone: A data-driven personal model reproducing mobility agendas. In *DSAA*. IEEE, 303–312.
[13] Riccardo Guidotti et al. 2018. Discovering temporal regularities in retail customers' shopping behavior. *EPJ Data Science* 7, 1 (2018), 6.
[14] Shan Jiang et al. 2012. Clustering daily patterns of human activities in the city. *DAMI* 25, 3 (2012), 478–510.
[15] John Lafferty et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).
[16] James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *BSMSP*, Vol. 1. Oakland, CA, USA, 281–297.
[17] Anastasios Noulas et al. 2011. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *AAAI*.
[18] Anastasios Noulas et al. 2013. Exploiting foursquare and cellular data to infer user activity in urban environments. In *ICMDM*, Vol. 1. IEEE, 167–176.
[19] Christine Parent et al. 2013. Semantic trajectories modeling and analysis. *CSUR* 45, 4 (2013), 42.
[20] Dino Pedreschi et al. 2019. Meaningful explanations of Black Box AI decision systems. In *AAAI*, Vol. 33. 9780–9784.
[21] Salvatore Rinzivillo et al. 2014. The purpose of motion: Learning activities from individual mobility networks. In *DSAA*. IEEE, 312–318.
[22] Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell system technical journal* 27, 3 (1948), 379–423.
[23] Amílcar Soares et al. 2019. VISTA: A visual analytics platform for semantic annotation of trajectories.. In *EDBT*. 570–573.
[24] Pang-Ning Tan. 2018. *Introduction to data mining*. Pearson Education India.
[25] Roberto Trasarti et al. 2011. Mining mobility user profiles for car pooling. In *KDD*. ACM, 1190–1198.
[26] Roberto Trasarti et al. 2017. Myway: Location prediction via mobility profiling. *IS* 64 (2017), 350–367.
[27] Nguyen Xuan Vinh et al. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *JMLR* 11, Oct (2010), 2837–2854.
[28] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *JASA* 58, 301 (1963), 236–244.
[29] Jihang Ye, Zhe Zhu, and Hong Cheng. 2013. What's your next move: User activity prediction in location-based social networks. In *ICDM*. SIAM, 171–179.