

# Identifying Atypical Travel Patterns for Improved Medium-Term Mobility Prediction

Roland Herberth\*, Leonhard Menz\*, Sidney Körper, Chunbo Luo, Frank Gauterin,  
Ansgar Gerlicher and Qi Wang

**Abstract**—During the last decades, concepts of Intelligent Transportation Systems (ITS) were continuously adapted and improved based on new insights into human travel behavior. Drivers for improvements are the quantity and quality of available mobility data, which increased significantly in recent years. Based on travel behavior, literature proposes a large number of different solutions for next step or future location prediction. However a holistic spatio-temporal prediction, which could further improve the quality of ITS, creates a more complex task. The prediction of medium-term mobility for one to seven days is challenging in particular for atypical travel behavior, since the weekdays’ order delivers no reliable indication for the next day’s travel behavior. With our contribution, we explore the benefits of various prediction approaches for medium-term mobility prediction and combine them dynamically to predict individual mobility behavior for a period of one week. The derived framework utilizes an exhaustive search approach to benefit from a machine learning based clustering method on location data. In conjunction with an Artificial Neural Network, the prediction framework is robust against prediction errors created by atypical behavior. With two data sets consisting of smartphone and vehicle data, we demonstrate the framework’s real-world applicability. We show that clustering an individual’s historical movement data can improve the prediction accuracy of different prediction methods that will be explained in detail and illustrate the interrelation of entropy and prediction accuracy.

**Index Terms**—Spatio-temporal, medium-term mobility prediction, pattern mining, atypical travel pattern, Intelligent Transportation System.

## I. INTRODUCTION

While mobility prediction has been an active field of research for a long time, the range of applications has grown significantly due to the availability of high resolution mobility data, mainly driven by the vast amount of smartphone data

This work was supported in part by the Dr. Ing. h.c. F. Porsche AG, Karlsruhe Institute of Technology, University of Exeter, Stuttgart Media University and European Commission under Grant Agreement H2020-ICT-2016-2/761913 SliceNet.

Roland Herberth, Leonhard Menz and Sidney Körper are with Dr. Ing. h.c. F. Porsche AG, 71287 Weissach, Germany. E-mail: {roland.herberth, leonhard.menz, sidney.koerper1}@porsche.de

Leonhard Menz and Chunbo Luo are with the College of Engineering, Mathematics and Physical Science, University of Exeter, Exeter, United Kingdom. E-Mail: {lm514, c.luo}@exeter.ac.uk

Frank Gauterin is with the Institute of Vehicle Systems Technology, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany. E-Mail: frank.gauterin@kit.edu

Ansgar Gerlicher is with the Institute for Applied Research, Stuttgart Media University, 70569 Stuttgart, Germany. E-Mail: gerlicher@hdm-stuttgart.de

Qi Wang is with the school of Engineering and Computing, University of the West of Scotland, Paisley, Scotland. E-Mail: qi.wang@uws.ac.uk

\*These authors contributed equally to this project and should be considered co-first authors.

[1]. A major challenge in predicting mobility for different applications is determining the right format and representation of the data, i.e. the prediction horizon and the required spatial and temporal resolution of the prediction outcome. A correct classification is challenging in particular, as the variety of application domains has been constantly growing since early studies in the 1940’s about a theory, relating mobility and distance [2]. This sets a starting point for mobility prediction.

A significant number of applications refer to mobility prediction in context of “next place” prediction in which a user’s next relevant location is predicted [3]. We refer to this kind of prediction as *short-term mobility prediction*, which is not only limited to the next-step, but also to the next upcoming hours. Different clustering methods have been proposed to improve short-term mobility prediction. Lv et al. [4] clustered users according to their living habits based on hourly entropy characteristics and identified four different user types. Based on their findings, the authors were able to demonstrate that, according to the user type, an adjustment of the applied prediction method can improve the short-term prediction accuracy.

On the other side of the spectrum there are *long-term mobility predictions*, aiming for mobility prediction in the scale of months to years. Sadilek et al. [5] point out that techniques, such as *Markov Models* (MM) and random walk-based formalism, which work well for short-term mobility prediction, are of little help in the context of long term prediction. This supports our theory that not only the scope, but also the available data and prediction range are essential for the study of an adequate mobility prediction algorithm.

In this paper we want to focus on what we call *medium-term mobility prediction*, which typically covers a time span from one day to one week. Focusing on this specific prediction horizon, our proposed framework identifies atypical travel patterns. This is a particular challenge for medium-term mobility prediction. The phenomenon of atypical travel behavior, which is characterized by movement that is not repeated on a weekly periodicity, has also been reported in [6]. We refine the definition of atypical travel behavior given in [6] and define travel events that cannot be associated to a specific weekday in an interval of one week as atypical. Examples for atypical travel behavior would be a periodic location visit every fortnight or a location visit every other day.

To illustrate this aspect, we refer to our previously introduced prediction model [7], in which day-specific mobility was clustered according to the weekday. Hence a prediction for Monday was always based and limited on data of all

historic Mondays. This, as well as a strict distinction between working days and weekends, which has also been applied in other works [8] [9], is a simplification that does not necessarily catch an individual’s atypical mobility behavior. For instance, for a shift worker with three-shift operation outside of a weekly cycle, weekday-specific clustering may lead to wrong results. Hence, in some cases it makes sense to cluster days according to the individual’s movement characteristics instead of the weekday’s name. Clustering based on day-specific travel characteristics also helps to identify day-specific routines and cross-day patterns. Thus, we propose a new clustering method that works in conjunction with an *Artificial Neural Network* (ANN) based sequence prediction and evaluate its impact on the performance of location dependent and location independent prediction on medium-term mobility. Investigations of hybrid models that combine different prediction methods and ANN have also been used in other researches in context of intelligent transportation systems [10]. With the proposed ANN we investigate a clustering solution that is able to take factors such as holidays for predicting the mobility of an individual into account.

By experiments on two different mobility data sets, consisting of *vehicle data* (DS1) and *smartphone data* (DS2), we illustrate the correlation between data sources, prediction method and prediction accuracy. We show that based on a user’s mobility profile entropy, the proposed method tends to reach the maximum predictability and that vehicle movement is more predictable than a person’s movement data collected with a smartphone. The contributions of our approach are:

- An adaptable medium-term prediction framework for practical ITS applications, which enables new applications, e.g. ride sharing, electric vehicles charging, vehicle software update planing and traffic analysis.
- Atypical movement pattern recognition and prediction, which contributes to a more reliable prediction scheme for individuals that do not follow typical travel patterns.
- Validation on two real-world GPS data sets, which demonstrates the advantages and applicability of the proposed prediction framework.

This work continues as follows: Section II provides an overview of the current state of research and technology; Section III introduces our prediction framework and describes the individual methods in detail; Section IV analyzes and compares the achieved prediction accuracy of our proposed framework for different real-world data profiles; Finally, Section V concludes this paper.

## II. RELATED WORK

Inherent in the historical development of human societies, human mobility behavior covers different domains. A comprehensive overview of the different domains of human mobility models and applications is given in [11]. For example, in [8] mobility prediction is used to pre-fetch relevant smartphone data. Gambs et al. [12] lists geo-privacy mechanisms and location based services as possible use cases for next place prediction. Using taxi movement data, Jenelius et al. [13] proposes a network travel time prediction method for ITS,

covering horizons from several minutes up to around one hour. Do et al. [14] uses GPS data collected on smartphones and a probabilistic kernel method to predict a user’s location for one and three hour intervals. Independent from the domain, human mobility patterns can be broken down to three key indicators: the trip distance distribution, the radius of gyration and the number of visited locations [15].

A number of researchers dedicated their work to find statistical models which can be used to describe these three key indicators. Brockmann et al. [16] found out that when  $P(d)$  is interpreted as the probability of finding a displacement of length  $d$ , the trip distance distribution can be approximated with  $P(d) \sim d^{-\beta}$  where  $\beta \approx 1.59$ . Gonzales et al. [17] concludes that the radius of a gyration distribution  $P(r_g)$  can be approximated with  $P(r_g) = (r_g + r_g^0)^{-\beta_r} \exp(-r_g/\kappa)$  with  $r_g^0 = 5.8$  km,  $\beta_r = 1.65 \pm 0.15$  and  $\kappa = 350$  km. In [18] it is shown that the number of distinct locations  $S(t)$  visited by a randomly moving human can be approximated with  $S(t) \sim t^\mu$  with  $\mu = 0.6 \pm 0.02$ .

We address two critical aspects of the previously mentioned findings. One is that the majority of data sets that were used for the formulation of the aforementioned equations do not consist of ground truth data. Instead, the majority of location data is derived from mobile phone records that were either captured when a person received a text message/call or periodically once per hour. The individual’s location was then assumed to be in the vicinity of the closest telephone pole. The second aspect is that the proposed methods were tested on long-term mobility behavior. As already mentioned in [15] and [18],  $S(t)$  does not show a robust scaling exponent  $\mu$  for  $t < 24$ h.

The rather theoretical description of human travel behavior provides limited value for the design of a medium-term prediction framework. This is because the description of movement patterns of a large quantity of individuals will not help to identify an individual’s atypical movement characteristics which creates a direct impact on this individual’s predictability for periods between one and seven days.

However, insights into individual movement characteristics [6], as well as the continuous progress in machine learning applications, help to capture movement on a more individual level and to adapt existing prediction methods to take atypical behavior into account. In concrete terms, the prediction accuracy of common statistical prediction methods, like Markov Models, can be further improved by machine learning techniques such as Artificial Neural Networks. ANNs in particular are often used for pattern recognition, also in context of mobility prediction [19]. In our case, we use an ANN to identify atypical mobility behavior.

Focusing on individual spatio-temporal mobility prediction for medium-term horizons, we use the previously described findings to derive a prediction framework, which is robust against atypical behavior and offers real-world applicability. With help of ground truth data, we demonstrate that for some individuals mobility prediction can be improved with a preceding clustering scheme which uses a categorical similarity measure to identify similarities in mobility on a daily basis. Also, we do not limit the prediction on the step length, radius of gyration and relevant locations, but include the most

probable route, which is selected from an individual set of recorded trajectories.

### III. METHODOLOGY

Our proposed prediction framework can be illustrated as a multi-step process consisting of an *Inter-Day Prediction* and a subsequent *Intra-Day Prediction*.

The Inter-Day Prediction determines the relevant days from the recorded user profile, which are used for the Intra-Day Prediction. The Inter-Day Prediction itself consists of two sub-methods. One corresponds to the “traditional” approach, in which the order of clusters is given by the sequence of weekdays. We call this sub-method *Weekly Cycle Dependent* (WCD). To be independent from this given sequence, we introduce the *Weekly Cycle Independent* (WCI) sub-method, in which we cluster 24-hour periods regardless from the weekday’s names, but based on their similarity regarding travel behavior. Since this step eliminates the natural sequence of weekdays, we explore a Feedforward Artificial Neural Network to predict the anticipated sequence of previously formed clusters.

Based on the output of the Inter-Day Prediction the Intra-Day Prediction forecasts the day-specific travel pattern. In this second step, we present two approaches where one can be assigned to the category *Location Dependent* (LD) and the other to the category *Location Independent* (LI). The overall structure of our proposed framework is illustrated in Figure 1.

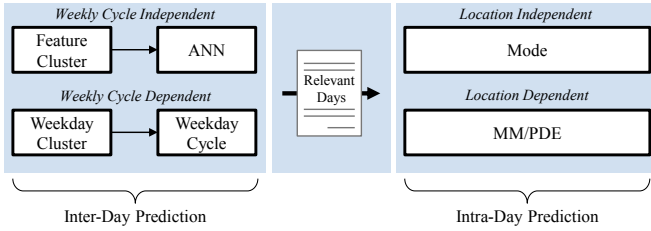


Fig. 1. Overview of the overall framework. The first step corresponds to the Inter-Day Prediction, consisting of the two alternative approaches: The first creates Feature Clusters (FC) and uses an ANN for cluster prediction, while the second uses the weekday cycle based on Weekday Clusters (WC). In the second step, the Intra-Day Prediction is applied on the basis of the selected data (days) in step one. Two alternative methods are presented here.

The goal of the framework’s design is to be applicable in different domains of location prediction, hence we pursue a simple way of providing spatio-temporal data to the system. We refer to the previously described characteristics in mobility patterns and define two elements in travel records: “visited places” and “transitions between locations”. In the next chapters of this paper, we refer to *Points of Interest* (POIs) and *Trip Identifiers* (TIDs) respectively. Note that this data input format allows a universal representation of spatio-temporal data, which could consist of ground truth GPS data, or as well as telephone pole data, as it has often been used in other research projects [15][20].

#### A. Inter-Day Prediction

The Inter-Day Prediction determines the relevant days from the recorded user profile, which are used for the Intra-Day

Prediction. The goal of Inter-Day Prediction is to determine which days in the profile records are relevant for the prediction of the next prediction period (24 hours). For this purpose, the profile data is clustered on a daily basis. Thus, the prediction is also conducted for 24 hour periods and repeated until the desired medium-term time span is reached.

1) *Weekly Cycle Dependent*: We implement the “traditional method” of forming day bins according to the weekday. We refer to this procedure as *Weekday Clustering* (WC). This approach assumes regularity on a weekly basis. The method is sometimes extended to distinguish mobility behavior between workdays and weekends [8] [9]. The sequence of days and hence the relevant data for the next day prediction are known a priori, as they are given by the order of weekdays. If a Tuesday is to be predicted, the data is based on all historical Tuesdays of the recorded mobility profile.

However, with this kind of clustering we observe poor performance for individuals with regularities outside one week, e.g. shift workers. Reference [6] illustrated on smart card data for public transportation in London that for some individuals a part of mobility behavior occurs outside of weekly periodicity.

2) *Weekly Cycle Independent*: To detect atypical regularities in a user’s travel history, we cluster days with similar movement patterns, regardless of day names, to be more robust against previously described atypical behavior. We employ a cluster analysis, which can be described in a three step process consisting of a similarity measure, fusion algorithm and determination of an adequate number of clusters. This procedure will be referred to as *Feature Clustering* (FC).

Boriah et al. [21] presented and evaluated several similarity measures for categorical data. We transfer their proposed method to quantify the similarity in daily movements with the *Goodall1* [21] similarity measure.

Therefore, let  $T \in \mathbb{N}$  denote the last time point of a day, which must be chosen according to the desired resolution, for example  $T = 24$  for a time point each hour or  $T = 1440$  for each minute. Further let  $\mathcal{P}_k$  be the set of all POIs that occurred on time point  $k$  for  $k = 1, \dots, T$  of the given user profile over all days in the data. Further let  $X = (X_1, \dots, X_T)$  and  $X' = (X'_1, \dots, X'_T)$  be the mobility data of two days, where  $X_k, X'_k \in \mathcal{P}_k$  describes the location at the  $k$ -th time point for  $k = 1, \dots, T$  of two days respectively.

The *similarity measure* (SM) is then calculated by:

$$SM(X, X') = \frac{1}{T} \sum_{k=1}^T S(X_k, X'_k), \quad (1)$$

where

$$S(X_k, X'_k) = \begin{cases} 1 - \sum_{q \in Q} p_k^2(q) & \text{if } X_k = X'_k \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

with

$$Q = \{\text{POI} \in \mathcal{P}_k \mid f_k(\text{POI}) \leq f_k(X_k)\}$$

and

$$p_k^2(X_k) = \frac{f_k(X_k)(f_k(X_k) - 1)}{N(N - 1)}. \quad (3)$$

$N \in \mathbb{N}$  denotes the total number of observed days of the user profile and  $f_k(x)$  denotes the total number of occurrences of POI  $x \in \mathcal{P}_k$  at the  $k$ -th point of time.

This similarity measure compares the locations for all corresponding time points of two days. Rarely visited locations are weighted higher to reduce the effect of an individual that mostly stayed in one POI, e.g. at home.

Since this is a similarity measure for categorical data, we utilize a hierarchical clustering algorithm. Finally, we use the *silhouette-coefficient* for cluster quality evaluation and to determine an optimal number of clusters [22]. It compares the mean intra-cluster dissimilarity and the mean inter-cluster dissimilarity of each object. To identify the optimal number of clusters, the amount of clusters to be formed is incremented in several cycles and the average silhouette coefficient is calculated. The optimal number is then determined by the highest average silhouette coefficient.

Figure 2 illustrates both cluster methods (WC and FC) for two exemplary user profiles from our vehicle data set (DS1). The respective quality of the cluster formation is illustrated by the entropy of the clusters formed. A lower entropy indicates a better clustering. The description of the entropy calculation is given in Equation (5).

With the formed Feature Clusters, a user's profile of 42 days results in a sequence of 42 clusters, which is exemplary illustrated in Figure 3. It is now necessary to determine the next cluster (day), since it is not given by the order of weekdays, as it is the case with the Weekly Cycle Dependent approach. To account for this, we assume that successive clusters provide information about subsequent clusters. The various regularities of the formed Feature Clusters can be taken from Figure 3. The same user profiles are used as in Figure 2.

To learn the regularities, a ratio of a dynamic number of known clusters to a following cluster is to be formed and learned. In order to improve prediction accuracy, additional categorical information should also be included to the actual cluster sequence. The information of known holidays and the current weekday will be considered in particular. The principle for two exemplary input clusters for a searched output cluster is illustrated in Figure 4. The cluster assigned to the respective day is indicated by each number and color.

The combination of different information types results in a complex non-linear relationship between input and output. An ANN is used to find existing patterns in the cluster sequence by using additional information at the same time. One advantage of the ANN is that additional information can be easily integrated into the forecast of the next cluster.

All described input factors of the ANN are categorical variables. In order to be able to process them, they must be *one-hot encoded* [23]. The output of the ANN shows which cluster is most likely to occur for the next day. Therefore it must also be one-hot encoded. In addition, the *Softmax-Function* [24] is applied to the output. Thus, the values of the individual outputs range between zero and one with the sum of all output neurons corresponding to one. As a result, the output values can be interpreted as probabilities for the cluster of the following day.

The coded data can now be used to train the ANN in order to find a functional relationship between the input and output clusters. Its structure is shown in Figure 5. To recognize the patterns in the respective profile, several ANNs with a different number of input clusters  $Cl_{t-n}$  are trained for each profile ( $n = 2, \dots, 7$ ). The one with the best cluster prediction performance is chosen. The trained ANN is then used to obtain a forecast of the next cluster for the upcoming day. All days in the recorded user profile assigned to this predicted cluster will now provide the input for the subsequent Intra-Day Prediction (see Figure 1).

### B. Intra-Day Prediction

The Intra-Day Prediction determines the specific POI for each time point of the day to be predicted. Therefore, it uses the outcome of the Inter-Day Prediction which consists of a list with selected days from the recorded mobility profile. The content (days) of the list can either be given by the Weekday Cluster (Weekly Cycle Dependent) or Feature Cluster (Weekly Cycle Independent). If it is based on WC, the Intra-Day Prediction is based on all days that are labeled identically to the corresponding weekday of the next day. If the Intra-Day Prediction is based on FC, the outcome consists of all days from the cluster predicted by the ANN (see Figure 5).

We are interested in comparing two different types of location prediction. For this purpose we employ 1) a location dependent and 2) a location independent prediction method:

1) *Location Dependent*: Our location dependent approach is based on a *Markov Model* (MM) combined with a *Kernel Density Estimation* (KDE). The MM is used to predict the next most likely location (POI). It is described by a set of states  $S = \{s_1, \dots, s_m\}$ , which represents the visited POIs of a user's profile. Changeovers between those states depend on their *transition probabilities* which are defined in a transition matrix  $P \in \mathbb{R}^{m \times m}$ . The transition probability  $(P)_{ij}$  depends on the quantity of trips driven from POI  $i$  to POI  $j$ .

The KDE is used for departure time prediction. In order to determine the most likely departure time for a specific location, KDE is used to determine periods, in which the temporal density of specific departure events is the highest. The set of considered departure events is defined by the previously predicted destination (POI) and a derived set of accountable TIDs. The density estimator  $\hat{f}_n(x)$  is defined as:

$$\hat{f}_n(x) = \frac{1}{nb} \sum_{i=1}^n K\left(\frac{x - x_i}{b}\right) \quad (4)$$

where  $K$  is the Gaussian kernel function,  $x_1, \dots, x_n \in \mathbb{R}$  is a random sample of length  $n \in \mathbb{N}$  and  $b > 0$  is the bandwidth. This two-stage process consisting of MM and KDE is used to predict location transitions for periods of up to one week. For a more detailed description of the location dependent Intra-Day Prediction, see [7]. It is adapted to work with the aforementioned Inter-Day Prediction approaches.

2) *Location Independent*: For reference, we employ a second model that is a variant of the compression method Prediction-by-Partial-Match of zeroth order, which has been



	Mo	Tu	We	Th	$WD_t$	Sa	Su
CW 1	1	1	2	1	1	3	3
CW 2	4	1	1	1	1	3	2
CW 3	4	1	1	1	1	3	3
CW 4	2	1	$Cl_{t-2}$	$Cl_{t-1}$	$Cl_t$		

Fig. 4. This plot illustrates the idea of cluster sequence prediction. It displays the sequence of clusters of all recorded days of a mobility profile. The number and color represents the determined cluster for the respective day. In this example there are two input days for the Input Layer ( $Cl_{t-2} = 1$ ,  $Cl_{t-1} = 1$ ) with the searched day being a Friday ( $WD_t = Fr$ ). In this case the ANN would predict  $Cl_t = 1$  as an output.

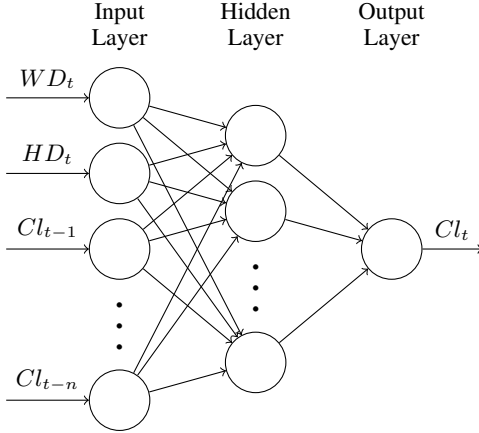


Fig. 5. Feedforward ANN for next cluster prediction. The Input Layer consists of a minimum of 4 and a maximum of 9 neurons. One corresponds to the weekday of the day looked for ( $WD_t$ ). A second neuron indicates whether the day of the week looked for is a public holiday ( $HD_t$ ). The remaining neurons correspond to the clusters of the previous days ( $Cl_{t-1}, \dots, Cl_{t-n}$ ). The Output Layer returns the searched cluster of the following day. By applying the Softmax-Function to the Output Layer, the probability for each possible cluster is determined.

used in a different form in [25] for mobility prediction. For our application we combine spatial and temporal information to divide weekdays in 1440 bins (one bin per weekday minute) and assign each bin to the recorded POI identifier (PID). The method uses no context (previous locations) and will, independent from the user's current location, predict an individual's next location to be the one that has been the most visited location at the day-specific time bin in the past. Note that in contrast to the location dependent prediction, this method is location independent, which especially for long predictions should be more robust against error propagation. As this method picks the most occurring POI identifier per time bin, we refer to this method with the mathematical operation *Mode* [6].

Depending on the individual behavior, the prediction results based on WCD and WCI can vary significantly. For this reason the framework follows an exhaustive search approach and creates predictions with both respective sub-methods. Based on the accuracy achieved for the present profile, the more suitable sub-method (WCD or WCI) is used for the final prediction of the profile.

## IV. EVALUATION

To evaluate the performance of the previously described framework, we will first discuss the characteristics of two data sets (DS1 and DS2) on hand and present the achieved prediction accuracy. We introduce three mobility prediction accuracy metrics for user specific predictions in conjunction with the corresponding user profile entropy.

### A. Data Pre-Processing

As described in Section III, we preprocess DS1 and DS2 to be processable for the introduced framework. As the data sets consist of GPS data, we apply the following processing steps:

*Identifying POIs:* POIs represent important and frequently visited places of an individual user. We use a mean shift approach to identify user specific POIs based on start and end points for driving records of DS1 and center points of dense location records in DS2. For DS1 arrivals and departures within a radius of empirically determined 500 meters are combined to one POI.

*Identifying TIDs:* Comparable to POIs, frequently driven trips get assigned to individual Trip IDs. Routes are characterized by a sequences of GPS points and corresponding edges between consecutively recorded coordinates. We declare route A and B as similar, when the average of the smallest distances between all edges of route A to its opposing coordinates of route B is below a certain threshold. The calculation is based on the Hausdorff distance and is comparable with methods in [26] and [27]. Later, we use the identified routes to predict not only destinations (POIs), but also the most likely route taken [7].

The pre-processing procedure yields data that contains aggregated information about POIs and the trips in between. Table I shows a small sample of processed data.

$POI_{Dep}$	$Time_{Dep}$	TID	$POI_{Arr}$	$Time_{Arr}$
$POI_1$	$t_1$	TID <sub>1</sub>	$POI_2$	$t_2$
$POI_2$	$t_3$	TID <sub>2</sub>	$POI_3$	$t_4$
$POI_3$	$t_5$	TID <sub>3</sub>	$POI_1$	$t_6$
$POI_1$	$t_7$	TID <sub>1</sub>	$POI_2$	$t_8$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

TABLE I  
PROCESSED GPS DATA, EACH ROW REPRESENTS A TRIP BETWEEN TWO POIS INCLUDING TIME OF DEPARTURE AND ARRIVAL.

Inherent to their collection method, both sets show different sources of errors, which has a measurable impact on their prediction accuracy outcome. Based on the definition given in [11], we apply similar prediction methods on data of different domains. Vehicle data refers to a specific car and is therefore considered as single-scale domain (unimodal). On the other hand, smartphone data is considered as multi-scale domain (multimodal), as the movement relates to several transportation systems.

### B. Mobility-Profile Analysis

In this section we identify and illustrate different mobility characteristics and distinctive properties of the collected user profiles.

1) *Observations*: Both data sets show differences in their characteristics, as displayed in Figure 6(a) and 6(b). Vehicles remain parked for around 61% in average at their most visited location while location data of smartphones indicate an average of 54% of dwell time for their most visited location.

For smartphone data more than 80% of the total recorded time was spent at one of the three most visited locations. The recorded vehicle data showed a 9% higher likeliness (89%) to find a vehicle at one of the three most visited parking locations.

Vehicle data is less complex compared to smartphone data as it features an unimodal movement record. The data shows fewer locations that we consider as relevant for the user and trajectories are in general limited to space that is accessible to vehicles. False data was observed when GPS signal strength was too weak or not existing, for instance in parking garages or tunnels.

Smartphone data is more complex compared to vehicle data as it consists of multimodal travel records and thus covers greater freedom of movement. The separation of “travel” and “movement” data is prone to incorrect classification and the identification of the particular transportation mode is a non-trivial task [3] [28] [29]. In addition, the available data set DS2 consists of time stamped GPS points without any indication of transportation mode and was collected without temporal consistency. To this end, the data set has been re-sampled to create profiles with one GPS point per minute. In order to be usable for our subsequent analysis, additional processing was necessary on the DS2 data set. We separated GPS records into movement data, for instance when a person travels or walks and non-movement data, when a person remains stationary. We assigned a trip when the data indicates an average speed of  $\geq 10$  km/h for at least 180 seconds.

2) *Determine Profile Regularity*: The maximum achievable prediction accuracy is limited by the randomness of the movement patterns within a user’s profile. Entropy measure is one way to quantify randomness and gives an indication about the maximal achievable prediction accuracy [6].

Literature proposes different approaches to estimate entropy in mobility data. Song et al. [20] were one of the first to use entropy to determine regularity in mobility profiles. They discovered that entropy indicates an upper limit to the predictability of human mobility. However, their approach is not independent of the spatial and temporal resolution on hand. As Burbey writes in [30], the form of representation of data has an impact on the prediction outcome. This is related to the findings of [6] and [31], showing that the spatial and temporal resolution of data influences the maximum predictability of movement data. Ikanovic et al. illustrated in [31] that the maximum achievable predictability increases with a lower temporal and spatial resolution of the same data.

Therefore we use an alternative temporal and local independent formula to calculate the entropy similar to [32]. We estimate the entropy  $E$  based on:

$$E = - \frac{\sum_{j=1}^T \sum_{i \in \mathcal{P}_j} \frac{l_{ij}}{N} \cdot \log \left( \frac{l_{ij}}{N} \right)}{T} \quad (5)$$

where the number of time slots in one day is  $T$ , the number of days is  $N$ ,  $\mathcal{P}_j$  is the set of all visited locations at time slot  $j$  and  $l_{ij}$  is the number of times location  $i \in \mathcal{P}_j$  dominates time slot  $j$  for  $j = 1, \dots, T$ .

Figure 7 illustrates the entropy distribution for both data sets in a histogram. It shows that the average entropy for DS1 is lower compared to DS2. This is due, among other things, to the higher level of mobility for smartphone data (DS2) than for vehicle data (DS1). The higher entropy is also indicated by the higher number of relevant POIs for DS2 as shown in Figure 6.

### C. Prediction Performance Analysis

An adequate comparison of different prediction models is challenging for several reasons: One reason is that most investigations use different data sets. Another is that there are diverse measurements that indicate the accuracy of a prediction method.

For the experiment, we present three different accuracy metrics to indicate the model’s prediction performance. For the measurement, the available profiles are cut in periods of seven consecutive weeks. In addition, we separate the last week from each mobility profile. This single week is not presented to the prediction models and serves for testing the prediction accuracy for each individual (out-of-sample test). Based on our previously given definition of medium-term mobility, which is characterized by prediction horizons of up to one week, the training data set has been dimensioned to create a 6:1 ratio of training and testing data, which is comparable to other research projects in mobility prediction [3] [33].

Under “Approach” Table II and III lists the different previously described combinations of Inter- and Intra-Day prediction methods and their mean and median prediction performances. The features (current location and current time of day) for the different approaches are the same.

The first accuracy metric illustrates the comparison of one predicted day to the first day of the separated test week (third column in Table II and III). The fourth column (“1 Week”) is used to indicate the accuracy that the framework achieved for the prediction of seven consecutive days, using the entire test week as comparison.

For an individual with a typical travel behavior, the separated test week would be a good representation of the remaining six weeks. However, this does not apply to individuals that show atypical travel behavior. Another source of error is that there might be a risk of separating a week that has no similarities to the remaining six weeks. For instance due to holidays, end of a semester, etc. In a best case scenario, the test week shares 100% similarity with the previous weeks. However, any deviation from regular behavior in the test week, would lead to a decrease in accuracy, even if an algorithm would have learned the past behavior correctly.

To account for this, we also compare the predicted week to the *average* mobility behavior of all seven weeks (third accuracy metric). For this purpose we compare the similarity of the predicted week with *each* of the seven recorded profile weeks. The normalized fraction of correct overlap accounts for



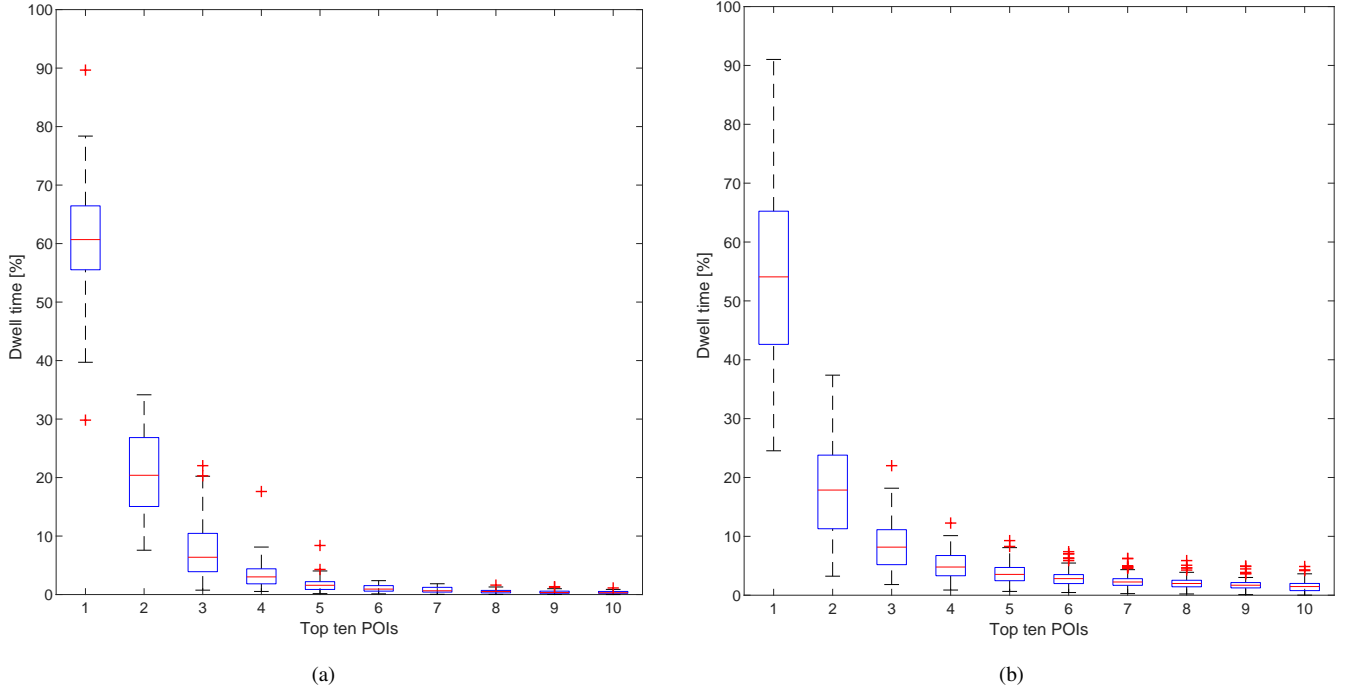


Fig. 6. (a) dwell time in percent per top ten location for vehicle data (DS1, 30 data samples). Around 89% of total record time is spend at one of the three most visited locations. (b) dwell time in percent per top ten location for smartphone data (DS2, 96 data samples). Around 80% of total record time is spend at one of the three most visited locations.

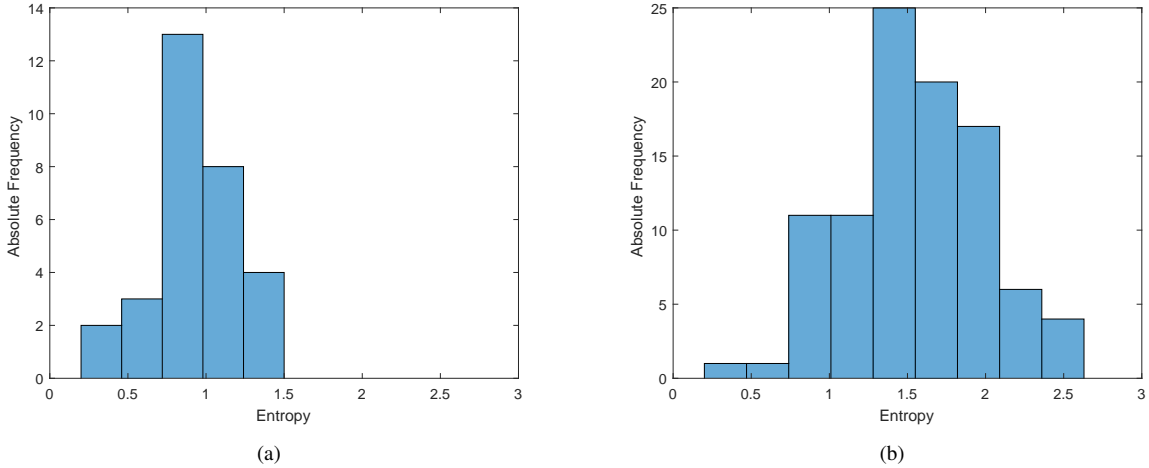


Fig. 7. (a) shows the histogram of the Entropy of DS1 while (b) shows the histogram of the Entropy of DS2. The entropy peak of DS1 is about 0.9, while DS2 is about 1.6.

the performance accuracy. This corresponds to an in-sample test.

Figure 1 shows that both Intra-Day Predictors (MM/KDE or Mode) are applicable with both Inter-Day Predictions (Weekly Cycle Dependent or Weekly Cycle Independent). This results in a total of four possible prediction outcomes. A fifth and sixth possibility arises when the respective Intra-Day Prediction automatically selects between the Inter-Day Predictions based on the better achieved accuracy (Auto-Select). This results in six different combinations of sub-methods per data set, whereby the achieved performance is compared with the three aforementioned accuracy metrics. Table II shows the achieved accuracy results for the proposed models for DS1,

while Table III shows the achieved accuracy results for the proposed models for DS2.

In general, the Mode operator achieves higher accuracy in all three accuracy metrics for both DS1 and DS2. Mode in conjunction with WCD achieves a slightly higher accuracy for one week and seven weeks than the Mode based on WCI. This can be attributed to the fact that most profiles do indeed show a weekly cycle based rhythm. Only for the one day comparison, the Mode with WCI in DS1 achieves a comparably high accuracy as the Mode with WCD.

Based on the exhaustive search approach, the framework calculates each Intra-Day Prediction for both Inter-Day Predictions. The better performing Inter-Day Prediction is chosen



Approach		1 Day	1 Week	7 Weeks
WCD + MM/KDE	Mean	52%	36%	34%
	Median	53%	34%	34%
WCI + MM/KDE	Mean	53%	41%	38%
	Median	56%	37%	35%
Auto-Select + MM/KDE	Mean	60%	44%	41%
	Median	66%	41%	38%
WCD + Mode	Mean	76%	68%	71%
	Median	81%	72%	72%
WCI + Mode	Mean	77%	66%	66%
	Median	86%	70%	68%
Auto-Select + Mode	Mean	83%	70%	69%
	Median	86%	75%	70%

TABLE II  
ACHIEVED PREDICTION ACCURACY BASED ON DS1.

Approach		1 Day	1 Week	7 Weeks
WCD + MM/KDE	Mean	46%	32%	28%
	Median	45%	31%	30%
WCI + MM/KDE	Mean	46%	38%	34%
	Median	47%	35%	32%
Auto-Select + MM/KDE	Mean	53%	40%	35%
	Median	56%	36%	33%
WCD + Mode	Mean	53%	55%	57%
	Median	57%	56%	57%
WCI + Mode	Mean	49%	49%	49%
	Median	52%	49%	48%
Auto-Select + Mode	Mean	58%	56%	55%
	Median	61%	58%	57%

TABLE III  
ACHIEVED PREDICTION ACCURACY BASED ON DS2.

(Auto-Select) based on the average prediction accuracy (of 1 Day, 1 Week, 7 Weeks). The decision is done individually for each user profile.

Hence, the identification of atypical travel behavior is conducted by selecting the better performing prediction method. If the prediction method for atypical movement performed better than the method for typical movement, the framework indicates that the respective individual shows an atypical movement behavior.

We conclude that location dependent prediction is more prone to errors caused by atypical travel behavior, as for 60% of our data profiles (DS1 and DS2), the ANN improved the prediction accuracy by 13.9% in average. Location independent prediction is less prone to atypical movement, as the accuracy for 22% of data profiles was improved by 5.8% in average through the ANN approach.

An insight into the prediction performance of the ANN itself is given in Figure 8. It shows the average prediction accuracy of seven consecutively predicted clusters (days), which corresponds to the accuracy metric “1 Week”. To determine the ANN’s predictive accuracy, we clustered the seven days of the out-of-sample data (Feature Clusters) and compared them to the predicted cluster of the ANN outcome. This results in a binary comparison of whether the correct cluster

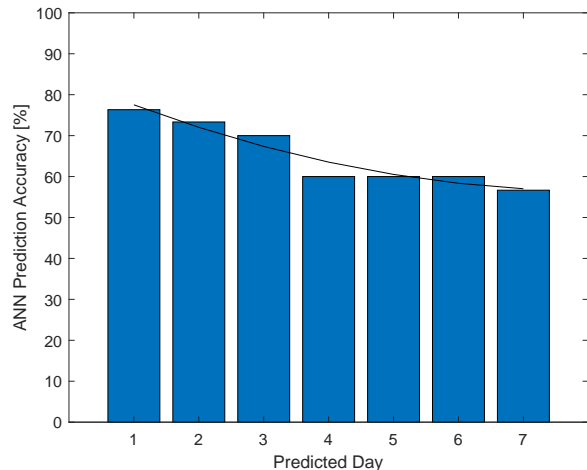


Fig. 8. Average cluster prediction performance of the ANN for a seven-day (cluster) prediction. Since incorrectly predicted clusters serve as input values for the next forecast, the prediction accuracy decreases (propagation error).

was predicted or not. It turns out that the forecast accuracy decreases over several days. This can be explained by the fact that the predicted cluster again serves as an input value for the ANN. An incorrectly predicted cluster thus influences the prediction of the next cluster, causing the error to propagate.

Figure 9 shows a detailed distribution of the achieved accuracy per user profile over the respective entropy. The plots (a), (b) and (c) in Figure 9 correspond to the data from row “Auto-Select + MM/KDE” of Table II and III, while the plots (d), (e) and (f) in Figure 9 correspond to the data from row “Auto-Select + Mode” of Table II and III. The colored marking of the data is used to distinguish between DS1 and DS2.

Figure 9 illustrates the correlation between prediction accuracy and entropy for the introduced accuracy metrics. Vehicle data (blue) shows a systematically lower entropy and hence a better prediction accuracy compared to smartphone data (red). We conclude a better predictability for vehicle data as vehicles are “less mobile” compared to human beings in terms of freedom of movement. Furthermore, we assume fewer detection errors in DS1 as it, other than DS2, requires no assumption about the transportation mode and hence should provide a better base for prediction.

The three different performance measures visualize the challenge of choosing an adequate performance indicator for location prediction data. While a day by day comparison of observation and prediction data is prone to outliers (Figure 9 (a) and (d)) for both MM/KDE and Mode, a comparison with the movement of an entire week shows a stronger correlation between entropy and prediction accuracy. Even though the mean and median prediction accuracy in DS1 and DS2 decreases for both MM/KDE and Mode. The “limits of predictability” become most obvious in Figure 9 (c) and (f), where one week of prediction data is compared to seven weeks of observation data. We expected a decrease of prediction accuracy for longer predictions due to error propagation. For the MM/KDE prediction variant in particular, we observe a higher decrease, which can be explained by the Markov

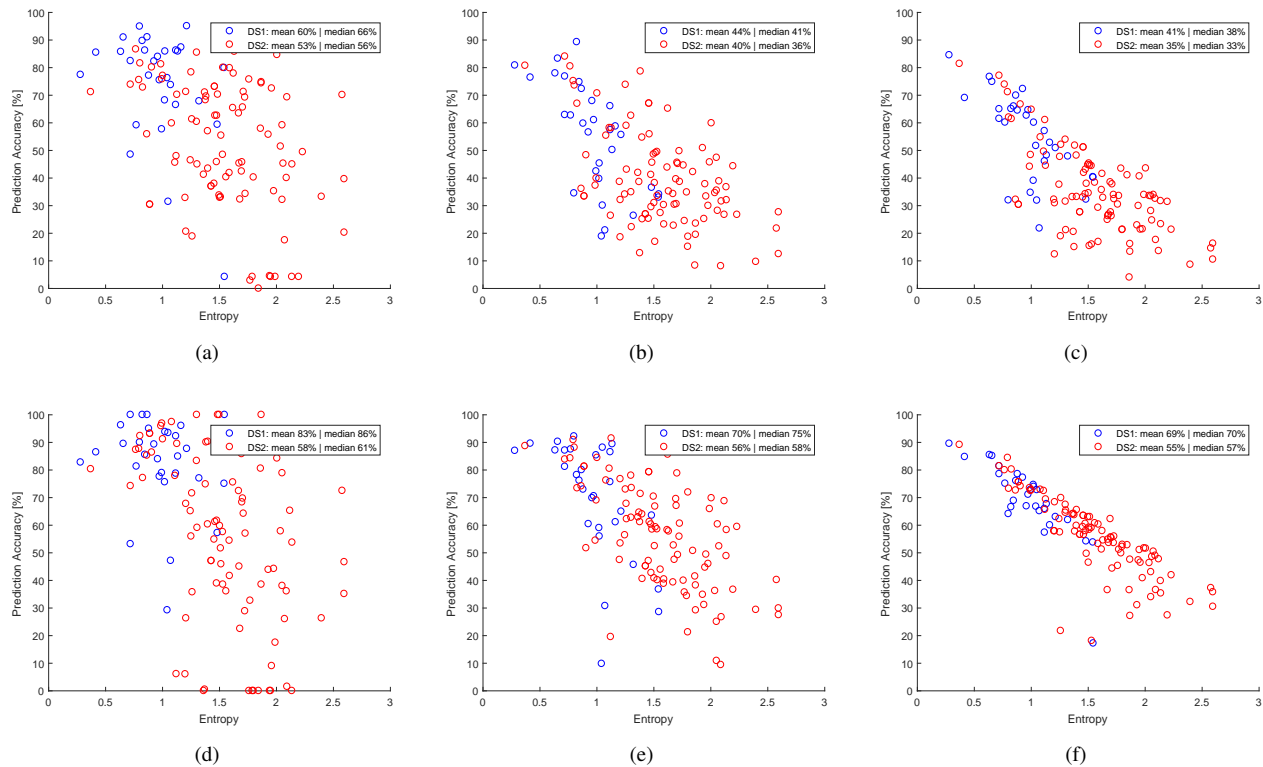


Fig. 9. Accuracy over Entropy: (a) 1 Day MM/KDE, (b) 1 Week MM/KDE, (c) 7 Weeks MM/KDE, (d) 1 Day Mode, (e) 1 Week Mode, (f) 7 Weeks Mode. Assessing the scatter plots from left to right illustrates that the shorter the comparison time frame between prediction accuracy and entropy, the smaller its correlation. This is reasonable as for the accuracy measure a day could have been chosen by coincidence that deviates from the average entropy of the entire profile data in (a) and (d). Also note that the mean and median prediction accuracy is higher in (a) and (d) which we explain by the fact that always the first day of prediction has been used for the calculation, which naturally has fewer propagation errors than the prediction of an entire week.

Model’s location dependency [8].

Also note that in Figure 9 (a) and (d), the prediction, regardless of the individual entropy, is in some cases completely incorrect. We remark that a simple predictor (Mode) did not perform better in these occasions, which implies that the user did something irregular.

## V. CONCLUSION AND OUTLOOK

In this paper we introduced a spatio-temporal mobility prediction framework, that focuses on a medium-term horizon, which typically covers periods of one week. With help of real-world data sets, we demonstrate that not in all cases a weekday-specific prediction, which is used by a number of state of the art prediction methods, leads to the best user specific mobility prediction. Instead, for individuals that follow atypical travel patterns, a feature related clustering method outperforms the weekday-specific prediction method. For 22% of individuals in our data sets, the framework detects atypical behavior and creates better prediction accuracy. We introduced an exhaustive search approach that computes and selects the best individual performing prediction method based on the achieved accuracy and adjusts the corresponding framework output leading to a better real-world applicability of the proposed framework.

Overall, we achieve a median accuracy of 75% for vehicle-based mobility profiles and 58% for smartphone based mobility profiles for an one-week mobility forecast and demonstrate

that the prediction outcome corresponds to the individuals profile entropy. We remark that these values must not be compared with performance indices of prediction methods that were applied for short-term prediction. We showed that our proposed framework tends to reach the entropy based upper limit of predictability. For further analysis, we intend to dynamically change the time frame of accountable data, which sets the basis for our prediction framework, to quantify the effect of historically old and new movement data. Also, future iterations of the proposed framework could include alternative sequence prediction models.

## REFERENCES

- [1] J. K. Laurila, D. Gatica-Perez, I. Aad, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, M. Miettinen *et al.*, “The mobile data challenge: Big data for mobile computing research,” in *Pervasive Computing*, no. EPFL-CONF-192489, 2012.
- [2] S. A. Stouffer, “Intervening opportunities: a theory relating mobility and distance,” *American sociological review*, vol. 5, no. 6, pp. 845–867, 1940.
- [3] Z. Zhao, H. N. Koutsopoulos, and J. Zhao, “Individual mobility prediction using transit smart card data,” *Transportation research part C: emerging technologies*, vol. 89, pp. 19–34, 2018.
- [4] Q. Lv, Y. Qiao, N. Ansari, J. Liu, and J. Yang, “Big data driven hidden markov model based individual mobility prediction at points of interest,” *IEEE Transactions on Vehicular Technology*, vol. 66, pp. 5204–5216, 2017.
- [5] A. Sadilek and J. Krumm, “Far out: Predicting long-term human mobility,” in *AAAI*, 2012.
- [6] G. Goulet-Langlois, H. N. Koutsopoulos, Z. Zhao, and J. Zhao, “Measuring regularity of individual travel patterns,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 5, pp. 1583–1592, 2018.

- [7] L. Menz, R. Herberth, C. Luo, F. Gauterin, A. Gerlicher, and Q. Wang, "An improved method for mobility prediction using a markov model and density estimation," in *Wireless Communications and Networking Conference (WCNC), 2018 IEEE*. IEEE, 2018, pp. 1–6.
- [8] Paul Baumann, "Human mobility and application usage prediction algorithms for mobile devices," Ph.D. dissertation, TU Dresden, Dresden, 19 August, 2016.
- [9] J. Scott, A. Bernheim Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges, and N. Villar, "Preheat: controlling home heating using occupancy prediction," in *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 2011, pp. 281–290.
- [10] Z. Ma, J. Xing, M. Mesbah, and L. Ferreira, "Predicting short-term bus passenger demand using a pattern hybrid approach," *Transportation Research Part C: Emerging Technologies*, vol. 39, pp. 148–163, 2014.
- [11] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini, "Human mobility: Models and applications," *Physics Reports*, 2018.
- [12] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez, "Next place prediction using mobility markov chains," in *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*. ACM, 2012, p. 3.
- [13] E. Jenelius and H. N. Koutsopoulos, "Urban network travel time prediction based on a probabilistic principal component analysis model of probe data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 2, pp. 436–445, 2018.
- [14] T. M. T. Do, O. Dousse, M. Miettinen, and D. Gatica-Perez, "A probabilistic kernel method for human mobility prediction with smartphones," *Pervasive and Mobile Computing*, vol. 20, pp. 13–28, 2015.
- [15] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González, "Unravelling daily human mobility motifs," *Journal of The Royal Society Interface*, vol. 10, no. 84, p. 20130246, 2013.
- [16] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, p. 462, 2006.
- [17] Marta C. Gonzales, Cesar A. Hidalgo, and Barabasi, "Understanding individual human mobility patterns," *Nature*, no. 453, 2008.
- [18] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, no. 10, p. 818, 2010.
- [19] V. Etter, M. Kafsi, E. Kazemi, M. Grossglauser, and P. Thiran, "Where to go from here? mobility prediction from instantaneous information," *Pervasive Mob. Comput.*, vol. 9, no. 6, pp. 784–797, Dec. 2013.
- [20] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.
- [21] S. Boriah, V. Chandola, and V. Kumar, "Similarity measures for categorical data: A comparative evaluation," in *Proceedings of the 2008 SIAM International Conference on Data Mining*. SIAM, 2008, pp. 243–254.
- [22] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53 – 65, 1987.
- [23] J. E. Beck and B. P. Woolf, "High-level student modeling with machine learning," in *Intelligent Tutoring Systems*, G. Gauthier, C. Frasson, and K. VanLehn, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 584–593.
- [24] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.
- [25] I. Burbey, "Predicting future locations and arrival times of individuals," Dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 26 April, 2011.
- [26] S. Atev, G. Miller, and N. P. Papanikolopoulos, "Clustering of vehicle trajectories," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 3, pp. 647–657, 2010.
- [27] P. Besse, B. Guillouet, J.-M. Loubes, and R. François, "Review and perspective for distance based trajectory clustering," *arXiv preprint arXiv:1508.04904*, 2015.
- [28] A. Jahangiri and H. A. Rakha, "Applying machine learning techniques to transportation mode recognition using mobile phone sensor data," *IEEE Trans. Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2406–2417, 2015.
- [29] B. Wang, L. Gao, and Z. Juan, "Travel mode detection using gps data and socioeconomic attributes based on a random forest classifier," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 5, pp. 1547–1558, 2018.
- [30] I. Burbey and T. L. Martin, "Predicting future locations using prediction-by-partial-match," in *Proceedings of the first ACM international workshop on Mobile entity localization and tracking in GPS-less environments*. ACM, 2008, pp. 1–6.
- [31] E. L. Ikanovic and A. Mollgaard, "An alternative approach to the limits of predictability in human mobility," *EPJ Data Science*, vol. 6, no. 1, p. 12, 2017.
- [32] S.-M. Qin, H. Verkasalo, M. Mohtaschemi, T. Hartonen, and M. Alava, "Patterns, entropy, and predictability of human mobility and life," *PLoS one*, vol. 7, no. 12, p. e51353, 2012.
- [33] V. Etter, M. Kafsi, E. Kazemi, M. Grossglauser, and P. Thiran, "Where to go from here? mobility prediction from instantaneous information," *Pervasive and Mobile Computing*, vol. 9, no. 6, pp. 784–797, 2013.



**Roland Herberth** received the B.Sc. and M.Sc. degrees in Mechatronics from University of Stuttgart in 2011 and 2015, respectively. He is currently pursuing the Ph.D. degree with Dr. Ing. h.c. F. Porsche AG, supervised by F. Gauterin from Karlsruhe Institute of Technology (KIT). His current research interests are Over-the-Air Diagnosis and Software Updates of Connected Cars.



**Leonhard Menz** received the B.Sc. in mechanical engineering and mechatronics from the University of Applied Sciences in Furtwangen in 2012 and the M.Eng. in international business and engineering from the SRH University Heidelberg in 2015. He is currently pursuing the Ph.D. degree with Dr. Ing. h.c. F. Porsche AG.



**Sidney Körper** received the B.Sc. in Mathematics and M.Sc. in Mathematics in Economics from Karlsruhe Institute of Technology (KIT) in 2014 and 2017, respectively. He is currently pursuing the Ph.D. degree with Dr. Ing. h.c. F. Porsche AG, supervised by O. Bringmann from the University of Tübingen. His current research interests are Predictive Maintenance and Diagnosis of Connected Cars.



**Chunbo Luo** is a Lecturer in Computer Science at the University of Exeter, UK. His research interests focus on developing model based and machine learning algorithms to solve networking and engineering problems such as wireless networks, unmanned aerial vehicles (UAVs) and computer vision. He received his PhD degree from the University of Reading, UK, for his work on the study of high performance cooperative wireless networks in 2011. He is a Fellow of the Higher Education Academy, an IEEE member and a BCS member. He has over

40 publications and has served as program chairs or technical session chairs in several international conferences including 2018 IEEE Data Science and Systems, 2017 I-SPAN, 2015 IEEE Globecom.



**Frank Gauterin** received the diploma degree in Physics from University of Münster, Germany, in 1989, and the Dr. rer. nat. (Ph.D.) degree in Physics from University of Oldenburg, Germany, in 1994. From 1989 to 2006, he was in research and development positions at Continental AG, Germany, leaving as the Director of NVH Engineering (noise, vibration and harshness). Since 2006, he has been a full professor with Karlsruhe Institute of Technology (KIT), Germany. He is currently the Head of the Institute of Vehicle System Technology and a Scientific Spokesperson of the KIT Mobility Systems Center. His research interests include vehicle control, vehicle dynamics, vehicle NVH, vehicle suspension, tire dynamics and tireroad-interaction as well as vehicle concepts, vehicle modeling, and identification methods.

His research interests include vehicle control, vehicle dynamics, vehicle NVH, vehicle suspension, tire dynamics and tireroad-interaction as well as vehicle concepts, vehicle modeling, and identification methods.



**Ansgar Gerlicher** is a Full Professor and Director of Research in Auto- / Mobile Applications & Security at the Stuttgart Media University, Germany. He is founder of the Research Lighthouse Responsive Media Experiences and the Institute of Mobility and Digital Innovation. Previously, he worked for several years as a Software Architect and Project Manager in the Telecommunication and Automotive Industries. He has published many papers and co-authored several books on mobile software development and computer science and media and was a Best Paper

Award Winner of the IEEE ICCE 2012. His research interests include mobile, distributed and embedded software architectures, frameworks and security. He received his PhD degree in real-time collaboration systems from the London College of Communication, UArts, UK.



**Qi Wang** is a Full Professor with the University of the West of Scotland (UWS), UK. He is Technical Co-Manager for EU Horizon 2020 projects SELF-NET and SliceNet, and Principal Investigator for a number of other projects funded by UK EPSRC etc. His research interests include wireless/mobile, multimedia, and software-defined networks. He has published more than 120 papers in these areas. He was a Best Paper Award Winner of IEEE ICCE 2014 and IEEE ICCE 2012, SOFTNETWORKING 2017, and SIGMAP 2014. He was also a winner of 2018

Scottish Knowledge Exchange Award. He received his PhD degree in Mobile Networking from the University of Plymouth, UK.