Discovering Regularity in Mobility Patterns to Identify Predictable Aggregate Supply for Ridesharing

Transportation Research Record
1–11© National Academy of Sciences:
Transportation Research Board 2018
Article reuse guidelines:sagepub.com/journals-permissions

DOI: 10.1177/0361198118798720

journals.sagepub.com/home/trr Ivan Mendoza^{1,2}, Clas Rydergren³, and Chris M. J. Tampère¹

Abstract

Heterogeneous data collected by smartphone sensors offer new opportunities to study a person's mobility behavior. The mobility patterns extracted from the travel histories found in these data enable agents residing in mobile devices to model transitions between visited locations, so that upcoming trips can be predicted after observing a set of events, and assistance can be planned in advance. When several agents cooperate, the forecasted trips made by multiple users can provide a potential supply for shared mobility systems such as dynamic ridesharing. These trips must be sufficiently regular and frequent to be reliably announced as shareable trips. This paper describes a methodology to identify a predictable aggregate supply for ridesharing via mobility patterns discovered in users' travel histories. The methodology empirically quantifies measures like the regularity and frequency of these patterns on a dataset consisting of 967 users scattered across different geographical areas. The sample exhibits high heterogeneity with respect to these measures (hence, of predictability, regardless of the prediction method). This paper shows how frequency of trip patterns decreases, while regularity increases, when additional dimensions such as departure times are added to the analysis. It was concluded that the traveler flexibility with regard to accepting less regular trips is vital to discover a larger supply. These results provide insights to develop future applications able to take advantage of this approach, to increase ridesharing rates, allowing a critical mass to be more easily attained.

Today, some mobile apps can track a user's activities, producing logs for which the statistics are mainly used for sporting and health purposes; other apps in the "life-logging" category use sensors that allow users to track themselves during an entire day, logging the complete chain of events. Once processed, the heterogeneous data collected by these apps through the sensors of a mobile device, including but not limited to global positioning systems (GPS), accelerometers, and gyroscopes, can also provide insights about the person's mobility behavior. The distinctive information found in these data is the chain of activities carried out during the day, including visited locations, time spent at these places, and additional characteristics depending on the complexity of the algorithms.

Data-mining techniques allow extraction of mobility patterns from travel histories constructed from these logs, making it possible to identify regular transitions between repeatedly visited locations. These transitions can be conditioned to a particular detected state or context defined by attributes such as current location, time of day, or day of the week. The regularity of these patterns enables the prediction of future trips, so that

software agents residing in a user's mobile device can autonomously decide about planning service assistance in advance.

One of the transport services that could potentially benefit from this approach is dynamic ridesharing: when multiple agents cooperate within a region. If everyday car trips with available seats made by multiple users could be predicted for a certain time interval, day, or destination, information about a potential supply for ridesharing, matching an upcoming ride request, could be inferred and announced earlier. Ridesharing passengers could for instance visualize the expected supply, consisting of potential shareable trips to a destination of interest compatible with their own schedules. This visual

¹KU Leuven, L-Mob, Leuven Mobility Research Center, CIB, Leuven, Belgium

²Universidad del Azuay, Cuenca, Ecuador

³Department of Science and Technology, Linköping University, Norrköping, Sweden

Corresponding Author:

Address correspondence to Ivan Mendoza: ivan.mendoza@kuleuven.be

information could display hotspots where the expected trips will occur, so that passengers can find suitable pickup points to enable ridesharing. The possible software applications using this approach could provide an instrument to attain a critical mass for ridesharing services by promoting an increase in the number of participants. In this context, a critical mass refers to having a minimum number of drivers, so that passengers can find supply for a ride requirement; but at the same time, having a minimum number of passengers, enabling drivers that want to share their trip costs to find sufficient demand for requests. This paper proposes a solution from the supply perspective: assuming that without a supply, no passengers can choose to share rides, but without passengers, drivers could still offer rides and hence create a basis for a community of ridesharing users to grow steadily.

The objective of this paper is to define a methodology to estimate a supply of shareable trips for ridesharing that can be conditioned to a context, based on the analysis of trip frequency and regularity in travel histories. Then, a subset of these trips with a minimum level of regularity is used to produce visualizations of the potential supply. The key concept of trip regularity is explored empirically in a dataset of travel histories of 967 users scattered across different geographical areas. The present paper is organized as follows: the first section introduces the context of the research and states the paper's objective; a literature review of related works is then presented. The next section describes the proposed methodology, followed by a discussion of the results obtained from the empirical data, and finally the conclusions and potential future works are presented.

Literature Review

Ridesharing has received considerable attention in recent transport-related research; this alternative travel mode allows traffic congestion to be mitigated as the number of vehicles is reduced, bringing environmental and societal benefits. A review of some recent research and possible future directions on ridesharing can be found in Furuhashi et al., in which the authors also provide the following unified definition: "ridesharing is a transportation mode in which individual travelers share a vehicle for a trip, with the purpose of splitting travel costs among users with similar itineraries and time schedules" (1). Some of the suggested improvements included better ride-matching strategies, pricing systems, and multimodal integration so that ridesharing is combined with another travel mode to complete a trip. In this same article (1), it is also stated that: "A great number of choices for participants increases the importance of assistance by software agents that enable personalized travel planning

and execution."; so that our research uses this suggestion to propose an improvement for ridesharing. Selected relevant literature related to this research is studied in the following paragraphs.

An attempt to enable ridesharing through a recommender system, implemented as a web-based platform that uses large-scale smartphone mobility data is presented in Bicocchi and Mamei (2). Here, the authors extracted information from two datasets containing mobility traces to identify potential rides. Routines consisting of repeated transitions on certain days of the week between a user's frequent locations were found when mining trips endpoints, then rides were matched based on similarities between the origins and destinations of different users' routines. Recently, as detailed by Bicocchi et al. (3), the previous work was extended through a set of methods, which analyzed urban mobility traces to recognize matching rides along similar routes. A list of optimization alternatives for ride-matching can be found in Agatz et al. (4) and a review of techniques to extract different mobility patterns can be found in Lin and Hsu (5).

Cici et al. provided an upper bound for the potential reduction of traffic in three different cities through ridesharing (6). Mobility patterns concerning home and work locations found in data from different heterogeneous sources were extracted, then an algorithm for matching users with similar patterns considering additional constraints such as social distance was proposed. Research exploring the impact of ridesharing on congestion using mobile phone data is presented in Alexander and González (7). Here, the authors extracted the average daily origin-destination (OD) matrix per travel mode from mobile phone data to match trips with spatiotemporal similarities; the impacts on congestion were evaluated by considering different adoption rates.

The potential benefits of introducing meeting points in a ridesharing system to attain a critical mass are evaluated in Stiglic et al. (8). The authors used simulation to measure the impact of picking up and dropping off passengers at locations different than the actual origins or destinations, obtaining a significant increase in the number of matched trips. Later research by Stiglic et al. was extended by adding flexibility in departure and detour times (9). Finally, research performed in Goel et al. provides a method for choosing the best locations for these pickup points based on Voronoi diagrams (10).

Research evaluating regularity in mobility patterns can be found in Williams et al. (11), Wang et al. (12), and Zhong et al. (13). High regularity of trips leads to a low randomness or entropy in a choice set of destinations; these metrics are extensively used in literature for attribute selection, particularly in classification models such as decision trees (14).

The present paper extends these works by using travel data acquired from modern smartphone tracking apps: after a multi-step data-mining process, spatial characteristics of potential “sharable” trips are inferred and announced as hotspots to find ridesharing opportunities if patterns are sufficiently regular.

Methodology

In this section, the required methods to attain the paper’s objective are elaborated. These methods include the extraction of personal points of interest (POIs) and mobility patterns, when trips occur under certain conditions. Frequency and regularity measures are subsequently formulated so that a subset of candidate trips can be proposed as the potential aggregate supply for ridesharing.

Extracting Personal Points of Interest

To recognize recurrent visits to the same location in a travel history, the captured spatial coordinates of the trip’s endpoints cannot be used directly. Given that coordinates are estimates of a user’s position, they can be different on every new visit to the same location. Methods for detecting a trip’s endpoints can be found in Zheng et al. (15) and Mendoza et al. (16), and the problem of labeling endpoints as visits to known or new locations are detailed in Lin and Hsu (5) and Zheng et al. (15) among others (17–19). In this paper, POI refers to a frequently visited location in a user’s travel history. Typical approaches in literature to identify these spatial patterns use density-based clustering techniques such as DBSCAN (20) or OPTICS (21). The latter also allows a hierarchy to be obtained, facilitating the extraction of POIs at a desired scale (i.e., buildings, neighborhoods, regions, etc.). A general density-based procedure used to extract POIs will now be explained.

Let us assume a dataset H consisting of travel histories from multiple users with data collected by a smartphone app. Then, a travel history H^v , defined as a chain of trips $b_1, b_2, \dots, b_N$ between the endpoints where a user v stops during the day to start a new activity, is created. The following notation describes each trip i ’s characteristics:

- $b_{i.o}$ origin coordinates {latitude, longitude}, $\forall b_i \in H^v$
- $b_{i.d}$ destination coordinates {latitude, longitude}, $\forall b_i \in H^v$
- $b_{i.m}$ travel mode, $\forall b_i \in H^v$
- $b_{i.s}$ departure time in 24-h notation, $\forall b_i \in H^v$
- $b_{i.e}$ arrival time in 24-h notation, $\forall b_i \in H^v$
- $b_{i.w}$ day type, $w \in \{\text{weekday, weekend}\}$, $\forall b_i \in H^v$

Then, the spatial dataset containing coordinates of trips’ endpoints to be clustered is

$$S = \{x_1, x_2, \dots, x_{2N} \mid \forall b_i \in H^v : x_i \in \{b_{i.o}, b_{i.d}\}\}$$

The distance between two points, which is typically Euclidean after transforming coordinates to a Cartesian system (22), is denoted by $\text{dist}(x_1, x_2)$, then the neighborhood of a point x_i as shown in Figure 1a with points directly reachable at a radius $\varepsilon \geq 0$ is

$$N_\varepsilon(x_i) = \{x_j \in S \setminus \{x_i\} \mid \text{dist}(x_i, x_j) \leq \varepsilon\}$$

The complete neighborhood $N'_\varepsilon(x_i)$ shown in Figure 1b, consists of all merged points in neighborhoods of those previously discovered, that is,

$$N'_\varepsilon(x_i) = N_\varepsilon(x_i) \cup N_\varepsilon(x_j) \cup N_\varepsilon(x_k) \cup \dots \\ \forall x_j \in N_\varepsilon(x_i), \forall x_k \in N_\varepsilon(x_j), \dots$$

Algorithm 1 Find Complete Neighborhood

Function: Neighborhood ($x_i, N_\varepsilon(x_i)$, dataset S)
 For each unvisited $x_j \in S$
 Label x_j as visited
 If $\text{dist}(x_i, x_j) \leq \varepsilon$
 Set $N_\varepsilon(x_i) = \text{Neighborhood}(x_j, N_\varepsilon(x_i) \cup \{x_j\}, S \setminus \{x_j\})$
 Return $N_\varepsilon(x_i)$

Algorithm 2 Density-Based Clustering Structure

Function: DB-Clustering (dataset S)
 Set $C = \emptyset, k = 0, \varepsilon \geq 0, \text{minPts} \in \mathbb{Z}$
 For each unvisited $x_i \in S$
 Label x_i as visited
 Set $N'_\varepsilon(x_i) = \text{Neighborhood}(x_i, \{x_i\}, S \setminus \{x_i\})$
 Set $k = k + 1$
 Set $C = C \cup \text{newcluster}(N'_\varepsilon(x_i))$
 Return C

The process to obtain $N'_\varepsilon(x_i)$ is presented in Algorithm 1, and the final density-based procedure to find the set C of repeated locations in Algorithm 2, in which complete neighborhoods are included in the final result only if a minimum number of points (visits) specified by parameter $\text{minPts} \in \mathbb{Z}$ is reached, otherwise they are tagged as noise (see Figure 1, c and d). Then, the set of POIs in user v ’s travel history is

$$C^v \text{minPts}, \varepsilon = \{c_i \in C \mid \text{minPts} \leq |c_i|\}$$

Identifying Transitions between POIs

Following this process, the trip endpoints are tagged to identify the cluster they belong to, so that all visits to the same location have the same label. That is, assuming u_k is a unique identifier for cluster c_k and x'_i the label on point x_i is

Figure 1. (a) Neighborhood of a point; (b) complete neighborhood of the same point; (c) trip endpoints in a travel history; and (d) their clustering structure when $\minPts = 5$.

$$x'_i = \begin{cases} u_k & \text{if } x_i \in c_k \\ \text{undefined} & \text{otherwise} \end{cases}$$

Let u_0 be a label to denote endpoints at non-frequent locations, that is, those not included in a cluster. The following new notation is required for the updated trip i 's characteristics:

- $b_{i.o'}$ trip origin's label, $\forall b_i \in H^v$
- $b_{i.d'}$ trip destination's label, $\forall b_i \in H^v$
- u_0 label for "noise" points, in which $|c_0| < \minPts$

Two trips b_i and b_j by the same person are then assumed to have the same OD pair if $b_{i.o'} = b_{j.o'} \wedge b_{i.d'} = b_{j.d'}$, even though their actual endpoints' coordinates could be different. If the travel history is consistent it should be expected that $b_{i.d'} = b_{i+1.o'}$, $\forall b_i \in H^v$, producing a reliable chain of trips. After transitions between pairs of POIs are identified, regularity and frequency of these transitions with respect to their travel histories can be quantified.

Regularity of Transitions between POIs

In this paper, regularity of a certain trip characteristic (i.e., destination) is defined as the relative frequency of

that characteristic in a user's travel history, conditioned to a state defined by other characteristics; this relative aspect distinguishes regularity from frequency. Frequency is the number of events per unit of time; it denotes the probability that something happens, such that the expected number of events in a period is found as frequency $\times$ length of period. On the other hand, regularity is the relative frequency of an event i within a user's pattern: it denotes the probability that among all possible events in the given conditions, i is the current one. Values closer to 1 are more regular and hence easier to predict.

A high frequency does not necessarily imply high regularity or vice versa. For instance, an event (e.g., to travel from A to B) may be rare (e.g., only once per month), yet any time the person resides in A his next move may be to go to B, hence his move is 100% regular, conditional on residing in A. An agent observing this person for a while and identifying its current location to be A, may therefore have an easy job predicting B as the next destination. Inversely, a person might frequently travel between C and D (e.g., twice a day); nevertheless, when residing in C, there might be five more very frequent destinations other than D, which makes the trip C to D frequent but not very regular when only conditioned on residing in C.

Therefore, an agent that knows its current position in C cannot predict the next destination with confidence. The following functions provide different measures of regularity of a destination in a user v 's travel history.

Let $y_{i,j}$ be the number of trips in H^v between POIs u_i and u_j :

$$y_{i,j} = |\{b \in H^v | b.o' = u_i \wedge b.d' = u_j\}|$$

Let V be the set of all users with consistent travel histories. A passenger requesting a ride to a location q could potentially use trips toward its neighborhood:

$$N_\varphi(q) = \{b.d' \in H | \text{dist}(b.d', q) \leq \varphi\}, \text{ where } H = \bigcup_{v \in V} H^v$$

That is, $N_\varphi(q)$ contains all POIs destinations in every user's travel history near location q , where φ could denote the maximum distance a passenger is willing to walk from the drop-off location. Let P be the number of unique POIs in a user's travel history, the total number of inter-cluster trips to destination $u_j \in N_\varphi(q)$, including origins in u_0 is

$$y_{*,j}^* = \sum_{\substack{p=0 \\ j \neq p}}^P y_{p,j}$$

The regularity of a u_j with respect to other destinations in H^v is defined by

$$R(u_j) = P(u_j|v) = \frac{y_{*,j}^*}{|H^v|}$$

That is, the probability that user v visits location u_j without being conditioned to any current state. On the other hand, the frequency, measured in number of visits per day to u_j is

$$f(u_j) = \frac{y_{*,j}^*}{T^v}$$

where T^v , is the size of user v 's travel history in days. The total number of trips from an origin POI u_i , including destinations in u_0 is

$$y_{i,*} = \sum_{\substack{p=0 \\ i \neq p}}^P y_{i,p}$$

The regularity of transitions to u_j (OD regularity) from an origin u_i is then

$$R(u_j|u_i) = \frac{y_{i,j}}{y_{i,*}}$$

Furthermore, transitions conditioned to a certain departure time interval can be evaluated. First, the 24-h format of departure times used in the dataset is transformed to a discrete value, so that trips with similar times can be clustered in the same group. Consider a finite number of time segments ω during the day. For instance, with a time interval $\Delta t = 15$ min, a day consists of $\omega = 96$ segments. The resulting departure and arrival time periods are $b.s' = \frac{b.s}{\Delta t} + 1$ and $b.e' = \frac{b.e}{\Delta t} + 1$ respectively, with values $\{x \in \mathbb{Z} | 1 \leq x \leq 96\}$. Then the regularity of destination u_j conditioned to an origin and departure time is

$$R(u_j|u_i, s_k) = \frac{y_{i,j,k}}{y_{i,*},k}$$

where $y_{i,*},k$, represents the number of all trips starting from u_i at departure period s_k , defined as

$$y_{i,*},k = |\{b \in H^v | b.o' = u_i \wedge b.s' = s_k\}|$$

whereas $y_{i,j,k}$, considers only those to destination u_j , that is,

$$y_{i,j,k} = |\{b \in H^v | b.o' = u_i \wedge b.d' = u_j \wedge b.s' = s_k\}|$$

Finally, the frequency of these trips would be

$$f(u_j|u_i, s_k) = \frac{y_{i,j,k}}{T^v}$$

More dimensions can be added, although the number of trips constrained to the new conditions is significantly reduced, as well as their frequency. These new conditions can include day type: $b_i.w \in \{\text{weekday}, \text{weekend}\}$ and travel mode: $b_i.m \in \{\text{car}, \text{public transport}\}$.

The overall regularity $R(u_j|\theta)$ of destination u_j in travel history of user v , conditioned to a multidimensional state $\theta = (\theta_1, \theta_2, \dots)$, consists of any information sensed by an agent to be used when deciding to announce a potential trip.

Identifying the Aggregate Supply

Contrary to selecting the location with the highest probability to be the next trip's destination, as done by a typical predictor (i.e., a classifier), all patterns above a minimum regularity value are considered by the agent. Therefore, when a threshold ρ is applied to filter out unreliable trip predictions, the expected trips of user v conforming conditions stated in θ are

$$H^{v,\rho}(\theta) = \{b \in H^v | \forall i \in \theta : b.i = \theta_i \wedge R(b.d'|\theta) \geq \rho\}$$

For a specific passenger's request to location q , the fraction of the aggregate supply, that is, the car trips subset that could potentially be used for ridesharing is

$$H^p(\theta, q) = \{b \in H^p(\theta) | b.d' \in N_\phi(q) \wedge b.m = 'car'\}, \text{ where } H^p(\theta) = \bigcup_{v \in V} H^{v,p}(\theta)$$

The relevance of a trip pattern, when presented to a passenger, depends on how accurately it can be predicted and how frequent it is. Then the relevance of trip b_i can be quantified by predictability $\times$ frequency. Regularity denotes the probability of making a trip conditioned to some known state, and allows identifying which characteristics are relevant to correctly predict a specific destination; then predictability is a function of regularity and correlates positively (the more regular the more predictable), meaning they can be used interchangeably.

A first visualization highlights the origins of this predicted aggregate supply, so that these hotspots can be used by passengers as pickup points. In Figure 2a, origins of trips in $H^p(\theta, q)$ denoted by red circles, have different relevance for passengers, which is indicated by their radius size.

Let r_i be the circle's radius for trip i 's origin, then

$$r_i = \lambda R(b_i.d'|\theta) \times f(b_i.d'|\theta), \quad \forall b_i \in H^p(\theta, q)$$

where λ , is a configurable scaling parameter to correctly visualize the plot. Another useful piece of information is the predicted paths used by those trips so that passengers can find a ride on the way as shown in Figure 2b. The line weight of links in a network $G = (V, E)$ can represent the link relevance, calculated by

$$w_e = \lambda \sum_{b_i \in H^p(\theta, q)} R(b_i.d'|\theta) \times f(b_i.d'|\theta) z_{l,e}$$

where w_e is the line weight for link $e \in E$ for θ , and $z_{l,i} \in \{0, 1\}$ is a binary variable indicating whether link e

is included in the predicted path of trip b_i . The line weight is then determined by the number of trips using that link, and also by each trip's predictability and frequency.

Results and Discussion

The previous methods are now applied to a dataset with multiple users' travel histories collected via a smartphone tracking app called MOVES; the dataset is briefly described in the following.

Dataset Description

The MOVES dataset contains travel histories of users in different geographical areas represented by unimodal trip chains. The variation in size of travel histories with respect to the number of days and trips is presented in Figure 3, *a* and *b*.

The travel modes automatically inferred by the app are: walking, running, cycling, and motorized vehicles; there is no distinction, for instance, between car or public transport. The distributions of travel distance and time per travel mode are plotted in Figure 3, *c* and *d*; outliers corresponding to very infrequent trips' characteristics have been filtered out using Tukey's method (23). These plots exhibit high heterogeneity in data, nevertheless most trips are short both in time and distance.

To evidence the need of further filtering, as expected when treating big data, a summary of the main characteristics of this dataset is given:

- 967 unique registered users as well as travel histories;
- 692,306 registered trips in 1,070 days;

Figure 2. (a) Origins (red circles) of trips neighboring a location (blue circle), acting as ridesharing hotspots for which relevance is denoted by the radius size; and (b) trip's paths, where relevance is specified by line weight.

Figure 3. Heterogeneity in travel histories in MOVES dataset: (a) size in days; (b) in number of trips; (c) distribution of travel distance; and (d) time in the dataset.

- The time frame of travel histories ranges from 2 to 928 days.

The patterns studied in the next section were extracted from histories with at least 30 days of tracking, reducing the number of users to 573. The travel mode of trips was not considered in the study.

Extraction of Mobility Patterns

As parameter ε , corresponding to the search radius for the data-mining procedure, must be previously calibrated, trips with lengths smaller than this value were filtered out; this avoids trips with origins and destinations ending in the same POI. The heterogeneity in number of POIs per user, when $\varepsilon = 200$ m and $minPts = 5$ is presented in Figure 4a. Here, several POIs per user were found due to the low number of required visits specified by parameter $minPts$; nevertheless, the number of actual points of interest a user regularly visits is much smaller, as seen in Figure 4b, when only those with at least two visits per week were considered.

The following patterns correspond to repeated trips between the same pair of OD patterns after removing intra-cluster trips, when constrained to a minimum of observations denoted by parameter $minTrips$. The pattern's regularity, indicating the probability of the trip's destination conditioned to an origin, and the frequency

denoting the average number of times the pattern was observed per day are presented in Figure 4, c and d.

The most regular users ranked by the largest number of OD patterns are presented in Table 1, when a minimum regularity level of 0.5 and a minimum frequency of two trips per week are applied.

With respect to spatiotemporal regularity, the most regular users are displayed in Table 2 when adding departure time intervals, with parameters $minTrips = 5$ and $\Delta t = 15$ min. The corresponding regularity and frequency of these patterns can be seen in Figure 5, a and b. As noted, when more dimensions are added to the analysis, the number of trips and their frequency are reduced, therefore, the supply for more specific requests (e.g., those also including a desired arrival time) will be reduced; nevertheless, regularity and therefore predictability of patterns rises due to the increased availability of context information.

The most regular users with respect to OD-arrival patterns, which includes trips between repeated OD pairs arriving at regular times are displayed in Table 3; Figure 5, c and d show the corresponding frequency and regularity of these patterns. Accuracy in predictions is always wanted; nevertheless, if agents are less demanding when announcing upcoming trips, the potential aggregate supply would be sufficiently large to encourage passengers to participate. The findings support the idea that stronger trip patterns only involve a few POIs per travel

Figure 4. Variation in number of POIs per travel history: (a) those with a minimum of five occurrences in the entire history; (b) those with at least two visits per week; (c) regularity; and (d) frequency of the OD patterns.

Table 1. List of Most Regular Users with Respect to Their OD Patterns

User ID	OD patterns	Trips in patterns	Average frequency	Average regularity	POIs
U777	3	67	0.35	0.81	2
U992	3	109	0.54	0.71	3
U603	3	562	0.39	0.68	3
U000	3	236	0.3	0.65	2
U873	3	142	0.68	0.61	3
U211	3	166	0.32	0.6	3
U980	3	43	0.41	0.58	4
U552	2	87	0.29	0.93	2
U589	2	159	0.4	0.9	3
U519	2	176	0.61	0.85	3

Note: User ID = a unique user pseudo-identifier; OD patterns = the number of discovered patterns; Trips in patterns = the number of trips in history contained in patterns; Average frequency = the average daily frequency; Average regularity = the average regularity; POIs = the number of unique points of interest observed in the patterns.

history, as observed in the last two tables and as previously suggested by Figure 4b.

As the pickup hotspots of the potential aggregate supply conditioned to a hypothetical ride request include all patterns with minimum regularity, collecting several data

about the current context is necessary. However, an agent should select only a few characteristics to announce the predicted trips, as looking for too-specific and predictable patterns would produce a short or even empty supply.

Table 2. List of Most Regular Users with Respect to Their OD-Departure Patterns

User ID	ODD patterns	Trips in patterns	Average frequency	Average regularity	POIs
U516	4	53	0.44	0.99	3
U254	2	44	0.33	0.98	3
U925	2	21	0.31	0.96	3
U992	2	74	0.55	0.95	3
U098	2	33	0.34	0.94	3
U777	2	50	0.39	0.88	3
U040	2	93	0.37	0.87	3
U730	2	42	0.34	0.86	3
U282	2	40	0.43	0.85	3
U661	2	67	0.47	0.83	3

Note: User ID = a unique user pseudo-identifier; ODD patterns = the number of OD patterns within specific departure time periods; Trips in patterns = the number of trips in history contained in patterns; Average frequency = the average daily frequency; Average regularity = the average regularity; POIs = the number of unique points of interest observed in the patterns.

Figure 5. (a) Variation in regularity when adding departure period; (b) frequency of the OD-departure patterns; (c) variation in regularity when adding arrival period; and (d) frequency of the OD-arrival patterns.

Dealing with Other Datasets

The methods used in this paper can be used with datasets from other existing tracking apps, as long as trip chains can be retrieved containing spatiotemporal characteristics. The accuracy with respect to the collection of spatial

data may differ for each data set, because even though sensors and operating system programming interfaces are shared among mobile apps, each app may have its own inference and aggregation methods. Another required

Table 3. List of Most Regular Users with Respect to Their OD-Arrival Patterns

User ID	ODA patterns	Trips in patterns	Average frequency	Average regularity	POIs
U516	4	57	0.48	1	3
U966	3	80	0.33	1	3
U873	3	65	0.31	0.96	3
U867	2	38	0.35	1	3
U992	2	77	0.57	0.97	3
U254	2	43	0.33	0.96	3
U980	2	22	0.31	0.92	3
U282	2	37	0.39	0.89	3
U777	2	48	0.38	0.88	3
U040	2	91	0.36	0.86	3

Note: User ID = a unique user pseudo-identifier; ODA patterns = the number of OD patterns within specific arrival time periods; Trips in patterns = the number of trips in history contained in patterns; Average frequency = the average daily frequency; Average regularity = the average regularity; POIs = the number of unique points of interest observed in the patterns.

item of information not included in the previous analysis is used travel modes. MOVES, unfortunately, does not differentiate motorized vehicles, though in some cases, modes can be inferred from existing data such as way-points and instant speeds. It must be noted that databases consisting of raw data directly retrieved from sensors, should first be converted to a trip-chain format; this means that trip endpoints (stops) must be identified prior to data-mining procedures.

The main concern for research using user data is privacy. An appropriate data collection mechanism to certify that information is stored anonymously and with the user's explicit consent is compulsory. The methodology described in this paper does not require tracks or patterns to be stored in a remote server; instead these can reside in the smartphone for further use by software agents, avoiding the needless flow of data through insecure channels. Finally, the predicted aggregate supply consisting of the announced trips by other agents does not need to include any driver contact information.

Conclusions and Future Work

A methodology to identify mobility patterns at different levels from smartphone data collected by apps tracking a user's daily activities has been presented, together with measures to evaluate their relevance for ridesharing. The origins and trajectories of the predicted trips through these patterns represent hotspots for ridesharing in which passengers may find a potential aggregate supply conditioned to their ride requirements. For ridesharing, regularity that allows accurate prediction of trip patterns and frequency is important to determine realistic supply. It has been shown that mobility regularity can be identified from smartphone data via data-mining procedures, although heterogeneity in the travel histories is high. The number of patterns and their frequency has been found

to decrease when more trip characteristics are considered, although regularity and prediction accuracy with respect to a set of conditions will increase. Flexibility on accepting trips with lower regularity, as well as different pickup windows or locations from their actual origins, is important to discover a larger aggregate supply. The method described in this paper is believed to facilitate the design of future apps that will help increase ridesharing rates.

Some future directions might include adding travel mode to the mobility patterns, as the user's role in ridesharing depends on this characteristic. The potential of ridesharing could be evaluated in specific regions via other datasets or activity-based simulators. The latter approach could be attained through the generation of a synthetic population consisting of users with heterogeneous mobility behaviors, for instance, by transferring the distributions found in patterns extracted from datasets generated by tracking apps to an alternative dataset containing demographic data. Each possible scenario for the simulation would involve distinct ridesharing penetration rates, as well as thresholds of regularity and frequency, so that in each case the supply announced by an agent would be different.

Acknowledgments

This research project was financially supported by the National Secretariat of Higher Education, Science, Technology and Innovation of Ecuador (SENESCYT).

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: IM, CT; data collection: CR; analysis and interpretation of results: IM, CR, CT; draft manuscript preparation: IM. All authors reviewed the results and approved the final version of the manuscript.

References

1. Furuhata, M., M. Dessouky, F. Ordóñez, M. Brunet, X. Wang, and S. Koenig. Ridesharing: The State-of-the-Art and Future Directions. *Transportation Research Part B*, No. 57, 2013, pp. 28–46.
2. Bicocchi, N., and M. Mamei. Investigating Ride Sharing Opportunities through Mobility Data Analysis. *Pervasive and Mobile Computing*, Vol. 14, 2014, pp. 83–94.
3. Bicocchi, N., M. Mamei, A. Sassi, and F. Zambonelli. On Recommending Opportunistic Rides. *IEEE Transactions on Intelligent Transportation Systems*, 2017.
4. Agatz, N., A. Erera, M. Savelsbergh, and X. Wang. *Optimization for Dynamic Ride-Sharing: A Review*. North-Holland, 2012.
5. Lin, M., and W.-J. Hsu. Mining GPS Data for Mobility Patterns: A Survey. *Pervasive and Mobile Computing*, Vol. 12, 2014, pp. 1–16.
6. Cici, B., A. Markopoulou, E. Frias-Martinez, and N. Laoutaris. Assessing the Potential of Ride-Sharing Using Mobile and Social Data: A Tale of Four Cities. *Presented at the International Joint Conference on Pervasive and Ubiquitous Computing*, 2014, pp. 201–211.
7. Alexander, L. P., and M. C. González. Assessing the Impact of Real-Time Ridesharing on Urban Traffic Using Mobile Phone Data. *Proc., UrbComp*, 2015, pp. 1–9.
8. Stiglic, M., N. Agatz, M. Savelsbergh, and M. Gradisar. The Benefits of Meeting Points in Ride-Sharing Systems. *Transportation Research Part B: Methodological*, Vol. 82, 2015, pp. 36–53.
9. Stiglic, M., N. Agatz, M. Savelsbergh, and M. Gradisar. Making Dynamic Ride-Sharing Work: The Impact of Driver and Rider Flexibility. *Transportation Research Part E: Logistics and Transportation Review*, Vol. 91, 2016, pp. 190–207.
10. Goel, P., L. Kulik, and K. Ramamohanarao. Optimal Pick up Point Selection for Effective Ride Sharing. *IEEE Transactions on Big Data*, Vol. 3, No. 2, 2017, pp. 154–168.
11. Williams, M. J., R. M. Whitaker, and S. M. Allen. *Measuring Individual Regularity in Human Visiting Patterns*. 2012, pp. 117–122.
12. Wang, Y., N. J. Yuan, D. Lian, L. Xu, X. Xie, E. Chen, and Y. Rui. *Regularity and Conformity: Location Prediction Using Heterogeneous Mobility Data*. New York, NY, 2015, pp. 1275–1284.
13. Zhong, C., M. Batty, E. Manley, J. Wang, Z. Wang, F. Chen, and G. Schmitt. Variability in Regularity: Mining Temporal Mobility Patterns in London, Singapore and Beijing Using Smart-Card Data. *PLoS One*, Vol. 11, No. 2, 2016, pp. 1–17.
14. Caruana, R., and D. Freitag. Greedy Attribute Selection. In *Machine Learning Proceedings* (Cohen, W. W., and H. Hirsh, eds.), Morgan Kaufmann, San Francisco, CA, 1994, pp. 28–36.
15. Zheng, Y., L. Zhang, X. Xie, and W.-Y. Ma. *Mining Interesting Locations and Travel Sequences from GPS Trajectories*. ACM, New York, NY, 2009, pp. 791–800.
16. Mendoza, I., C. Tampère, and P. Mekerlé. Anticipatory Assistance for Real Time Ride-Sharing in Environments of Pervasive Computing. Presented at the 95th Annual Meeting of the Transportation Research Board, Washington, D.C., 2016.
17. Ashbrook, D., and T. Starner. Learning Significant Locations and Predicting User Movement with GPS. *Proc., Sixth International Symposium on Wearable Computers*, 2002, pp. 101–108.
18. Ashbrook, D., and T. Starner. Using GPS to Learn Significant Locations and Predict Movement across Multiple Users. *Personal and Ubiquitous Computing*, Vol. 7, No. 5, 2003, pp. 275–286.
19. Nurmi, P., and S. Bhattacharya. Identifying Meaningful Places: The Non-Parametric Way. In *Pervasive Computing* (Indulska, J., D. Patterson, T. Rodden, and M. Ott, eds.), Springer, Berlin Heidelberg, 2008, pp. 111–127.
20. Ester, M., H.-P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proc., 2nd Conference of ACM SIGKDD*, ACM Press, New York, 1996, pp. 226–231.
21. Ankerst, M., M. M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: Ordering Points to Identify the Clustering Structure. No. 28, 1999, pp. 49–60.
22. Butler, H., C. Schmid, D. Springmeyer, and J. Livni. *Spatial Reference*. EPSG Projection 31370 - Belgian Lambert 72. 2016.
23. Tukey, J. W. *Exploratory Data Analysis*. Addison-Wesley Publishing Company, 1977.

The Standing Committee on Urban Transportation Data and Information Systems (ABJ30) peer-reviewed this paper (18-06037).