Measuring Regularity of Individual Travel Patterns

Article *in* IEEE Transactions on Intelligent Transportation Systems · September 2017

DOI: 10.1109/TITS.2017.2728704

4 authors:

Gabriel Goulet-Langlois
Transport for London
5 PUBLICATIONS **10** CITATIONS
SEE PROFILE

Haris N. Koutsopoulos
Northeastern University
159 PUBLICATIONS **3,772** CITATIONS
SEE PROFILE

Zhan Zhao
Massachusetts Institute of Technology
5 PUBLICATIONS **24** CITATIONS
SEE PROFILE

Jinhua Zhao
Massachusetts Institute of Technology
31 PUBLICATIONS **241** CITATIONS
SEE PROFILE

Measuring Regularity of Individual Travel Patterns

Gabriel Goulet-Langlois, Haris N. Koutsopoulos, Zhan Zhao, and Jinhua Zhao

***Abstract*— Regularity is an important property of individual travel behavior, and the ability to measure it enables advances in behavior modeling, mobility prediction, and customer analytics. In this paper, we propose a methodology to measure travel behavior regularity based on the order in which trips or activities are organized. We represent individuals' travel over multiple days as sequences of "travel events"—discrete and repeatable behavior units explicitly defined based on the research question and the available data. We then present a metric of regularity based on entropy rate, which is sensitive to both the frequency of travel events and the order in which they occur. The methodology is demonstrated using a large sample of pseudonymised transit smart card transaction records from London, U.K. The entropy rate is estimated with a procedure based on the Burrows-Wheeler transform. The results confirm that the order of travel events is an essential component of regularity in travel behavior. They also demonstrate that the proposed measure of regularity captures both conventional patterns and atypical routine patterns that are regular but not matched to the 9-to-5 working day or working week. Unlike existing measures of regularity, our approach is agnostic to calendar definitions and makes no assumptions regarding periodicity of travel behavior. The proposed methodology is flexible and can be adapted to study other aspects of individual mobility using different data sources.**

***Index Terms*— Regularity, intrapersonal variability, travel behavior, smart card data, entropy rate.**

I. Introduction

TRAVEL behavior is dynamic and varies across individuals but also for the same person over time. Interpersonal variability refers to the heterogeneous spatiotemporal preferences of people, reflecting different sociodemographic attributes, home/work locations, and lifestyle preferences [26]. Intrapersonal variability describes longitudinal variability in the characteristics of the same individual's travel behavior from trip to trip, day to day, or week to week [13], [26], [31]. Sometimes it is referred to in the literature as intraindividual [15], or day-to-day variability [17], [21], [24]. Regularity refers to the extent to which individual travel behaviors repeat over time. A person's activity choices and their associated trips are not made randomly. According to activity-based travel theory, they are dictated by preferences, constraints, and needs which recur over time to some degree [20].

While conventional cross-sectional data, one-day travel diary surveys for example, can capture the interpersonal variability, measuring intrapersonal variability/regularity requires individual-level longitudinal data. Multi-day travel surveys, often used for activity-based modeling, provide such data but are costly to collect and hence usually constrained to small sample sizes and short observation periods. However, advances in urban sensing technologies afford the opportunity to collect traces of individual mobility on a large scale and over extended periods of time. New mobility data sources, such as mobile phone records and transit smart card records, enable detailed and reliable measurement of travel regularity. No existing definition and measure of behavior regularity align with the variety in people's routines and granularity which these new data sources can capture.

Central to the definition of regularity is the definition of a unit of analysis for which repetition is considered. This unit should be chosen in line with the attributes relevant to the research question of interest and consistent with the resolution of the available sensor data. Reference [15] use the term *behaviors* to describe components of travel behavior characterized by combinations of attributes, for example "driving a car to work". In this paper, we use the term "travel events" to refer to the same concept as [15]'s *behaviors*, but with a broader connotation. A travel event is a repeatable unit describing individual travel behavior, characterized by one or more attributes such as purpose, location, and duration. At the most basic level, a travel event is either a trip or an activity. Travel events can also be aggregated to different levels (e.g. daily or weekly) to form higher-level travel events. For example, for the analysis of individual daily routines, a travel event may be a combination of activities in one day. In this paper, if not specified otherwise, "travel events" are used to refer to the most basic building blocks of travel behavior—trips and activities.

Travel events do not occur in isolation. People's activity patterns govern the co-occurrence of multiple travel events. This is the basis of work on trip chaining behavior, e.g. [27], and activity-based models, e.g. [4]. Combinations of travel events reflect such activity patterns. Each event must be considered as part of this context. While some travel events are frequently repeated over time, their surrounding contexts may change from day to day [15]. This highlights that regularity depends not only on variability in the characteristics of a

Manuscript received April 2, 2017; revised July 11, 2017; accepted July 16, 2017. This work was supported by Transport for London. The Associate Editor for this paper was H. S. Mahmassani. *(Corresponding author: Jinhua Zhao.)*

G. Goulet-Langlois was with the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA. He is now with Transport for London, London SW7 2NJ, U.K.

H. N. Koutsopoulos is with the Department of Civil and Environmental Engineering, Northeastern University, Boston, MA 02115 USA.

Z. Zhao is with the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

J. Zhao is with the Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: jinhua@mit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2017.2728704

single event but also on the pattern in which multiple events are combined. In our approach, multiple travel events can be ordered over time and form "travel sequences".

In existing literature, some methods have been proposed to measure regularity by examining the periodic patterns of travel behavior [19], [32], [35]. However, periodicity is not equivalent to regularity. While periodicity only captures the cyclic repetitions of travel events at fixed time intervals (typically set as a day or a week), regularity refers to all forms of repetitions. Travel patterns may not necessarily repeat periodically or may repeat over unconventional periods not aligned with the typical day or week. To some extent, periodicity is a special type of regularity. The order in which an individual completes trips and activities is an integral component of the structure in their travel routines. A good metric of regularity should be sensitive to such sequential dependency in a travel sequence, without a predefined periodic cycle.

In this paper, we propose a new approach to measuring the regularity of travel behavior based on the order in which travel events are organized over time in travel sequences. The definition is not tied to an underlying calendar. Hence it is flexible. We demonstrate the approach using a large sample of transit smart card transaction records over a period of a month. The ability to measure regularity improves our understanding of travel behavior, facilitates advancements in behavior modeling, and enables the development of customer analytics for travel prediction, user segmentation, and targeted demand management.

The remainder of the paper is organized as follows. We present a literature review of the related work on intrapersonal variability/regularity in Section II. Section III proposes a sequential representation of travel behavior and develops its mathematical formulation. This is followed by a description of the proposed measure of regularity based on entropy rate in Section IV. The measure is demonstrated in Section V using smart card data from London, U.K. The paper is concluded with a discussion of future research directions and potential implications in Section VI, and a summary of the main findings in Section VII.

II. Literature Review

While the concept of travel behavior regularity is recognized as a critical dimension of travel behavior, approaches to measure such variability remain limited in scope. Specifically, many studies measure regularity based only on the extent to which single travel events are repeated, without consideration for how multiple events are combined. Some methods focus only on the relative frequency of trips. For example, [5] proposed a spatial repetition index corresponding to the percentage of activity locations which are visited more than once over a 7 day period. Based on survey data, this measure is computed for different time periods to evaluate the spatial stability of individual activity patterns at different times of the week. Based on smart card data, [18] identified the OD pairs that the card holder frequently travels as "regular OD" and the time of the trips between these regular ODs as "habitual time". They measured the regularity of transit users based on the percentage of a user's trips completed within habitual times and between regular ODs. Reference [23], using smart card data, evaluated the level of spatial and temporal variability of different users based on the frequency of trips made to different stops at different times of the day.

Other studies rely on the variance of different measures to quantify longitudinal variability. References [25] and [26] evaluated the variance in number of trips per day from a 7-day travel survey. Their results differentiated the part of the variance of trip generation rates associated with intrapersonal variability from the part associated with interpersonal variability. Reference [21] analyzed variability in the departure time of the first trip of the day. Relying on the concept of individual space-time prisms, they modeled the variance of first departure time so as to differentiate the part of the variance due to randomness, from the part due to changes in the time constraints dictating an individual's schedule. Similarly, [7] also attempted to dissect the variance of the first trip departure time by formulating a multilevel model for which the variance was decomposed into five parts: inter-individual variation, inter-household variation, spatial variation, temporal variation, and intra-individual variation. Like the frequency-based measures, these variance-based measures treat each trip independently and are not concerned with the sequence of multiple trips.

Accounting for combinations of travel events has long been recognized in the literature of travel behavior modeling as important. Some models rely on the assumption that activity and trip combinations are primarily a function of days of the week. For example, using the 7-day Toronto Travel Activity Panel Survey, [12] modeled the frequency of 15 non-home/work activity categories for the 7 days of the week using 7 independent models. In contrast, some studies model the relationship between different travel events more explicitly. Reference [29] modeled preplanned and spontaneous activity duration as well as number of trips by mode, using data from the 7-day activity survey in [12]. Their approach introduces same-day effects and next-day effects to capture the relationship between multiple activities. From a long-term perspective, [3] examined the relationship between successive activities for the same purpose (e.g. shopping) using a 6-week travel survey from Karlsruhe, Germany. They modeled the time elapsed between successive activities using a multivariate hazard model. Other studies used pattern recognition techniques to directly model the activity sequence as a whole, and such techniques include Walsh-Hadamard transformation [28], sequence alignment [16], and conditional random field [2]. These studies account, to various degrees, for the relationship between travel events to improve travel demand models. They use panel survey data and do not aim at measuring regularity in the order of travel events over time.

To measure regularity in combinations of travel events, many researchers, especially in the human mobility literature, proposed methods to uncover periodic patterns. Some studies use the Fourier transform to identify underlying periods of repetition in travel from digital traces of location collected over multiple weeks. Reference [19] found daily and weekly periods to be most significant in observing individuals' connection

to Wi-Fi access points (AP) on the Dartmouth campus. Reference [8] identified the same dominant periods using data from MIT's Reality Mining project. Reference [22] proposed a probabilistic measure of periodicity and demonstrated its robustness to noise and missing observations using GPS data, with superior performance over methods based on the Fourier transform.

The above studies account for repetition in combinations of travel events, by measuring the extent to which their co-occurrence map to a set calendar cycle (most often a weekly cycle). Other studies attempt to measure regularity explicitly by imposing a predefined cyclic period. For example, [35] proposed a measure of temporal irregularity in the intervals between a person's visits to a given location. They applied a weekly based measure to different data sources and found that the behavior captured from smart card data was most regular, while Wi-Fi data revealed the least regularity. Reference [32] presented another regularity measure also based on a weekly cycle. Given hourly information of a person's location over several weeks, they used the percentage of hours spent at the location most frequently visited during each hour of the week as the index of periodicity for the corresponding hour.

However, periodicity is not the same as regularity. Regularity indicates the degree to which sub-sequences of events are repeated, and these sub-sequences do not have to align with a particular cycle. This is especially relevant to sequences of activities, as activities are likely to be organized in a logical order. For example, visiting the doctor's office, going to the pharmacy to pick-up a prescription, and returning home are likely to occur in this logical order. The repetition of this sequence may not be periodic. Furthermore, [19], [32], [35], and [8] all discuss periodicity in the context of the most conventional cycles of repetition: the day and the week. We argue that regularity is an internal property of a travel sequence and should not depend on how the sequence aligns with the calendar. Some patterns may repeat on non-daily or weekly cycles. For example, certain types of employment (e.g. shift-workers, firefighters, doctors) may dictate working schedules which repeat on a cyclical unit other than the week. Periodicity measures computed on a weekly basis (as done by [32] and [35]) would fail to capture the true regularity in such cases. Similarly, a measure of daily periodicity may not be able to capture patterns spanning more than a calendar day, such as going out in the evening, sleeping at a friend's home, and then returning home the next day.

In conclusion, no index that captures repetition in the order in which events are observed has been introduced in the literature. In the following sections, we present a new metric for measuring the regularity of travel behavior that depends explicitly on the order in which travel events occur. As such, the metric avoids the issues inherent in existing periodicity-based measures which examine only co-occurring patterns of travel events and calendar events (i.e. hour, day, week).

III. Sequence Representation

Individual travel patterns can be conceptualized as a sequence of travel events. These events unfold over time with respect to a background calendar (time of day, day of the week, month). Travel events are characterized by different aspects of behavior, including location, time of day, mode, route, travel time, activity type (or travel purpose) and activity duration. For instance, an event defined as an activity occurs at a certain time of day (8 pm on Friday), for a certain duration (2 hours), at a certain location (downtown) and for a certain purpose. As recognized by [13]–[15], variations along these behavioral dimensions are not independent. For example, an individual's choice of mode or route will significantly influence the travel time for her morning commute, which impacts her departure time.

A key component of these sequences is the order in which events take place. An appropriate measure of regularity in a person's travel behavior should capture both, the extent of repetition in travel events and in the order in which they are performed. It is necessary to introduce a mathematical representation of travel sequences which captures the order of events to define such a regularity index. We model the mobility of each individual over multiple days as a random process, which represents how often and in what order travel events are generated. The notation follows that used by [9].

Let the stochastic process corresponding to the mobility of a given individual u be denoted by $\mathbf{X}_u$ and a travel event generated by this process by random variable X_u. Each travel event X_u assumes a discrete value x from the set of possible travel event outcomes E_u defined for individual u. x can be regarded as a unique identifier for a repeatable event. Two separate events assume the same value of x if and only if they have the same combinations of event attributes. X_u has a discrete probability distribution $p(x) = Pr\{X_u = x\}$ for $x \in E_u$.

For simplicity, subscript u is omitted and all remaining notation is defined with respect to a single individual. The stochastic process $\mathbf{X} = \{\ldots, X_{-1}, X_0, X_1, X_2, \ldots\}$ represents the ordered set of random variables X_i. Any finite sequence of this ordered set between event i and event j is denoted by the ordered subset $X_i^j = \{X_i, X_{i+1}, \ldots, X_{j-1}, X_j\}$, with $-\infty < i \leq j < \infty$ such that $X_i^j \subset \mathbf{X}$. Given a finite window of analysis, we observe a specific realization $x_i^j = \{x_i, x_{i+1}, \ldots, x_{j-1}, x_j\}$ of the finite random variable sequence X_i^j.

Informally, set E is akin to an alphabet from which a string of discrete events can be constructed. Different types of sequences, or strings, can be represented based on different definitions of travel events $x \in E$, driven by the aspects of behavior of interest. In practice, the specification of E is constrained by the available data. Different data provides information on varying aspects of travel and at various aggregation levels. For instance, smart card data provides location information at the stop level and the timing of the event, but no direct information on activity purpose.

For consistency and computation convenience, we assume all event attributes are discrete. This assumption is common for travel behavior analysis since many travel attributes are discrete by nature, such as purpose, location and time periods (e.g. morning peak, midday, afternoon peak). Attributes

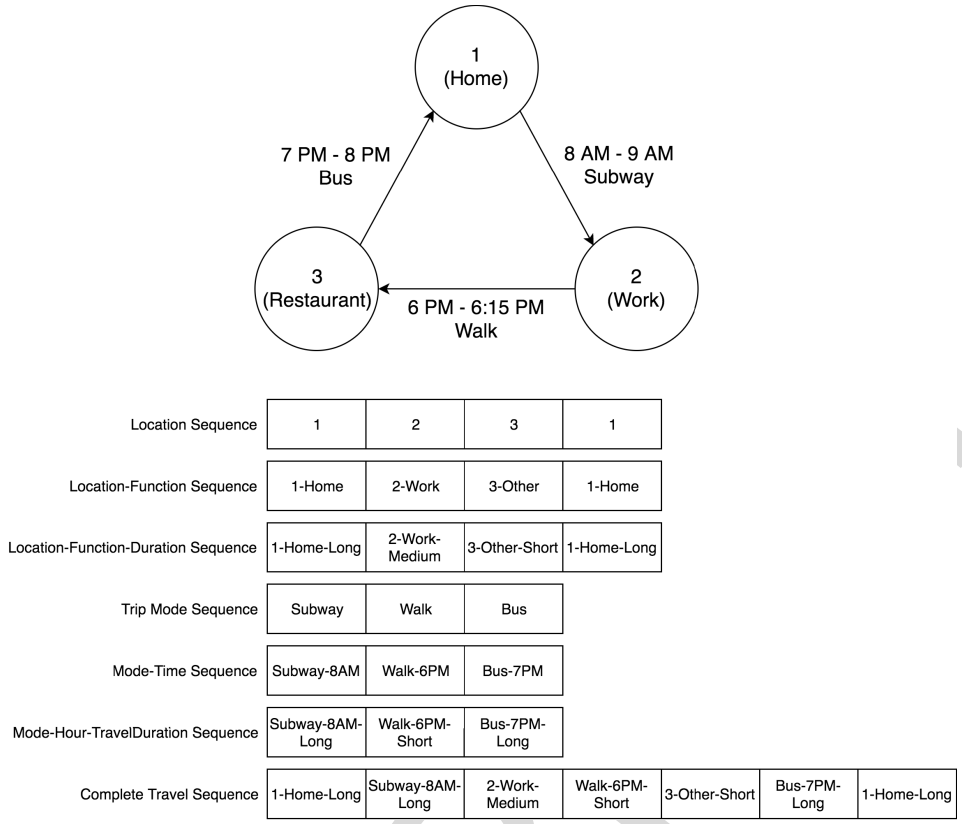


Fig. 1. Example of travel sequences.

that typically assume continuous values (e.g. activity duration) are discretized into a finite number of categories. The specification of these categories depends on both, the goal and the data of the analysis. While a larger number of categories can capture the variation of these attributes in finer detail, it can also make the specific values less repeatable and lead to a sparse distribution of $p(x)$. Ideally, these categories should meaningfully reflect behavioral choices. For example, using some clustering approach (e.g. Gaussian mixture model), the activity duration can be discretized into three categories - long, medium, short, and each of these categories is likely to be associated with certain activity types (e.g. home, work, other).

Fig. 1 shows how a person's travel over a day can be summarized as different travel sequences by changing the definition of travel events. For this example, we discretize activity duration into three categories - Long ($>$ 10 hours), Short ($<$ 3 hours) and Medium (between 2 and 10 hours), and travel duration into two categories - Long ($>$ 30 minutes) and Short ($<$ 30 minutes). We also characterize the trip start time using 24 hourly intervals. The level of discretization determines the granularity of travel events. Typically, finer granularity means that each travel event is more unique and less likely to repeat.

For many applications, a single aspect of travel behavior (i.e. purpose, location, or mode) is relevant. In these cases, the travel events only have a single attribute, and we may directly set the x value of an event to its attribute value. For example, the first sequence in Fig. 1 focuses on the locations visited by the person. This can be represented by defining set E as the set of all locations visited by the individual over the period of analysis. In this example, x_i^j is simply a series of location IDs.

In other contexts, it may be necessary to define events based on combinations of multiple attributes. For instance, location, function, and duration could be combined to differentiate between two activities observed in the same geographical area. In this case, the events x in set E are defined as compound outcomes of location, function, and purposes, as illustrated in the third sequence of Fig. 1.

At different levels of aggregation, multiple trips or activities can be grouped together to define a single event. For example, all trips made on the same day can be grouped into a single event to create a binary sequence representing when the person traveled across multiple days.

This representation provides a flexible approach to simplify and represent multidimensional travel behavior as a string of travel event symbols. These symbols are defined in line with the objective of the study so as not to distort or omit relevant information about aspects of travel of interest.

IV. Measurement of Regularity

As described in the previous section, we model the mobility of an individual over multiple days as a sequence of events generated by a random process $\mathbf{X}$. Through this abstraction, it is possible to characterize an individual's mobility by quantifying the nature of the random process $\mathbf{X}$. Many different

properties of process $\mathbf{X}$ may provide information about the individual's travel pattern. For example, consider a process $\mathbf{X}$ representing the activity sequence of an individual. In this case, the cardinality of set E informs us about the diversity of activities in which the individual engages, and the mode of probability distribution $p(x)$ reveals the individual's most frequent activity. This section introduces ways to measure such properties of $\mathbf{X}$ which can be used to describe regularity of a travel sequence.

A. Entropy vs Entropy Rate

First, we examine the extent of repetition of a travel sequence regardless of the order. Under this assumption, the regularity of a random process is solely determined by the probability distribution $p(x)$. Intuitively, on average, an outcome generated by a more regular process should be less uncertain and more predictable. In information theory, the level of randomness or unpredictability of a process can be measured using entropy. Entropy measures the average information, or surprise, provided by each realization of a random variable in bits. The entropy $H(X)$ of random variable X with probability distribution $p(x) = Pr\{X = x\}$ for $x \in E$ is defined by (1).

$$H(X) = -\sum_{x \in E} p(x) \log_2 p(x) \tag{1}$$

For the travel sequence problem, X represents the random variable associated with a travel event and E denotes the set of all possible travel event outcomes defined for a given individual. Entropy can be thought of as a measure of variance defined for categorical probability distributions. It accounts for both the number of possible outcomes (the cardinality of set E) and the relative frequency of outcomes. Hence, entropy equals 0 for a process with a single possible outcome (no uncertainty) and is highest when the probability distribution of a random variable with multiple outcomes is uniform (when all events are equally likely). Reference [30] used entropy to measure and contrast the complexity of activity patterns completed by individuals of different gender. The author points out that entropy is a good measure of the amount of heterogeneity in a categorical distribution, which is especially relevant when considering qualitative outcomes such as activities.

Although entropy is a good measure of repetition of isolated events in a travel sequence, it does not capture the extent to which ordered sub-sequences of events repeat over time. Travel sequences are not typically memoryless processes. Rather, the conditional distribution of an event X_i depends on the outcome of events $X_{i-1}, X_{i-2}, \ldots$ preceding it (i.e $p(X_i|X_{i-1}, X_{i-2}, \ldots) \neq p(X_i)$). For example, observing a visit to the doctor might significantly increase the likelihood of a visit to the pharmacy in the following event. Entropy rate accounts for the order of events in a travel sequence, or more formally for the memory in process $\mathbf{X}$. Entropy rate $H(\mathbf{X})$ of the random process $\mathbf{X}$ is defined as the asymptotic rate at which the entropy of sub-sequence X_1^n changes with increasing n [9], calculated using (2).

$$H(\mathbf{X}) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, X_3, \ldots, X_n) \tag{2}$$

where, $H(X_1, X_2, X_3, \ldots, X_n)$ denotes the entropy of the joint variable X_1^n defined for the subsequence $X_1, X_2, \ldots, X_n$. References [9] and [6] stated that this limit exists for all stationary random processes and is equal to

$$\begin{aligned} H(\mathbf{X}) &= \lim_{n \to \infty} H(X_n|X_{n-1}, \ldots, X_2, X_1) \\ &= \lim_{n \to \infty} -\sum_{x_1^n \in E^n} p_n(x_1^n) \log_2 \frac{p_n(x_1^n)}{p_n(x_1^{n-1})} \end{aligned} \tag{3}$$

where p_n denotes the joint probability distribution of a sub-sequence of length n. As described by (2) and (3), entropy rate measures the average entropy of each new event generated by random process $\mathbf{X}$, accounting for preceding events. It is measured as the entropy per event and has units in bits per event. The entropy rate of a random process with no memory is exactly equivalent to the entropy of the process as each new event is independent of the previous. As such, the entropy of a process is an upper bound for its entropy rate. In contrast, a process in which the outcome of an event X_i is perfectly determined by the previous events ($p(X_i = x|X_{i-1}, X_{i-2}, \ldots) = 1$) has an entropy rate of 0. Informally, entropy rate is the average measure of information, or surprise, associated with each additional event generated in a sequence of events. The more memory in a random process, the more information the previous events provide about the next event, and therefore the lower the entropy rate of the process. Also, memory in the random process is directly related to the order in which events are observed. Specifically, the more memory in a random process, the more the order of the events it generates tends to repeat. In line with these characteristics, the entropy rate is a good regularity measure of travel sequences because it is sensitive to not only the relative frequency of events but also the dependencies between multiple events.

Reference [32] used the entropy rate of hourly-location sequences derived from cell phone data to explore predictability in individual location patterns. Hourly-location sequences tend to have very low entropy rate because the location of a person during a given hour is highly related to their location in the previous hour. This is because individuals tend to visit a location for several hours consecutively (e.g. 8 hours at work or 14 hours at home). For these sequences, longer average activity durations are associated with low entropy rate. Hence, the high predictability reported by [32], albeit an interesting theoretical finding, is of limited practical use because it merely reflects the tendency of individuals to stay in a location for multiple hours. Nevertheless, their approach demonstrates how entropy rate can be used to quantify the dependencies between elements of the same sequence.

B. Estimation of Entropy Rate

Estimation of the entropy rate of a finite sequence can be computationally challenging. According to (3), the entropy rate is a function of the unknown joint probability distribution p_n of the sequence X_1^n. A naïve approach consists of estimating p_n from the observed frequency of combinations of symbols in X_1^n. This approach becomes computationally intractable as combinations of increasing length are considered.

Most entropy rate estimation approaches circumvent the issue of estimating p_n by relying on universal data compression algorithms [6], [9]. These algorithms, used to compress data generated from processes with unknown probability distributions and arbitrarily long memories, are known to achieve optimal lossless compression ratios (i.e. compression ratio equal to the entropy of the generating process). Hence, they can be used to estimate the amount of redundant, or repeated information in a sequence of symbols. For instance, text can be compressed by coding frequently repeated expressions. If for example, the 28-symbol phrase "the probability distribution" frequently occurs in the text, it can be coded by a single symbol.

Three families of lossless compression methods have been applied to entropy rate estimation. References [34] and [33] introduced the context-tree weighting (CTW) based entropy estimator. Reference [6] developed an estimation approach based on the Burrows-Wheeler transform (BWT), and [9] proposed different estimators based on the Lempel-Ziv (LZ) family of data compression algorithms. The estimators perform differently with respect to efficiency and bias depending on the property of the source and the sequence realization considered. Longer sequence realizations provide entropy estimates with smaller variance, and the variance of different approaches converges at different rates. The size of the alphabet also influences accuracy, with larger alphabets resulting in both, higher variance and potentially higher bias for an equal number of observations. Reference [9] presented an extensive comparison of LZ and CTW estimators using simulation for binary sequences. They concluded that the CTW estimator consistently provides more accurate and reliable results than LZ-based estimators. Reference [6] established an upper bound on the convergence rate of the BWT estimator for finite-alphabet, finite-memory processes and demonstrated that the BWT estimator performs better than an LZ-based estimator for binary sequences. No direct comparison of the CTW and BWT estimates has been reported in the literature.

The BWT estimator is simpler to implement. Hence, the BWT entropy estimator with uniform segmentation, as described by [6], is used for the case study presented in Section V. The authors prove almost-sure convergence of this estimator for stationary, ergodic random processes. These properties are assumed to hold for travel sequences described by the formulation previously introduced. Specifically, we assume that the underlying characteristics of an individual's mobility do not change over the period for which the individual is observed. This assumption would be violated if a long-term change (e.g. change in residential location or job) took place during the period of analysis.

The BWT entropy estimator is computed in two steps. First, the Burrows-Wheeler transform is applied to the finite sequence X_1^n of length n. Reference [1] provided an in-depth discussion of the transform, its properties, and implementation. Table I, adapted from [1], illustrates how the BWT operates. The BWT is applied to an example sequence *aardvark*, resulting in the transformed sequence *kavraad*. First, all rotations of the input sequence are listed and sorted alphanumerically. Then, the last symbol of each rotation is retained. BWT groups together outcomes (or symbols) which occur in similar contexts in the original sequence.

TABLE I
AN EXAMPLE OF BWT (ADAPTED FROM [1])

All Rotations	Sorted Rotations	
aardvark	aardvar**k**	k
ardvarka	ardvark**a**	a
rdvarkaa	arkaard**v**	v
dvarkaar	dvarkaa**r**	r
varkaard	kaardva**r**	r
arkaardv	rdvarka**a**	a
rkaardva	rkaardv**a**	a
kaardvar	varkaar**d**	d

Formally, the BWT of any stationary process $\mathbf{X}$ with finite memory results in a piecewise memoryless sequence. Reference [6] leverage this property of the transformed output to estimate the entropy rate of the process that generated the original sequence. Specifically, in the second step of the estimation, the transformed sequence is segmented into S segments s of uniform length, and the distribution of outcomes is estimated for each segment according to (4).

$$\hat{q}(x,s) = \frac{N_s(x)}{\sum_{y\in E} N_s(y)} \tag{4}$$

where $N_s(x)$ denotes the number of occurrences of symbol x in segment s. Given $\hat{q}(x,s)$, the entropy of each segment s is estimated by (5). Finally, the entropy rate of $\mathbf{X}$ is estimated by the average entropy of all segments using (6).

$$\log_2 \hat{q}(s) = \sum_{x\in E} N_s(x) \log_2 \hat{q}(x,s) \tag{5}$$

$$\hat{H}(\mathbf{X}) = -\frac{1}{n} \sum_{s\in S} \log_2 \hat{q}(s) \tag{6}$$

Reference [6] recommend that the length of each segment s is set as the integer value closest to $\sqrt{n}$. As mentioned above, the accuracy of the resulting estimate depends on both the length of the sequence observed and the number of different outcomes it contains.

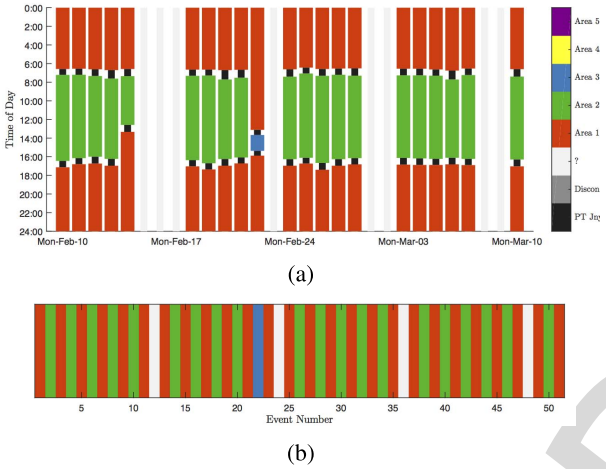
V. CASE STUDY

In this section, we demonstrate the proposed methodology described above using transit smart card data. Transport for London (TfL) provided the dataset used for this study. It consists of the smart card records of a sample of 99,925 pseudonymised cards observed between February 10th and March 10th 2014. The dataset covers Oyster transactions across all public transport modes including bus and rail. While the rail transactions contain entry and exit records with their associated stations and timestamps, the bus transactions only include boarding stop and time. Thus the alighting stop and time are inferred using the ODX method developed by [10]. ODX provides a set of complete public transport trips for each passenger.

For this case study, we are particularly concerned with the activities occurring between journeys, rather than the

TABLE II
ACTIVITY STATUS SUMMARY

Status	Semantics
-1	User activity location cannot be inferred because a non-PT trip was completed between observed PT journeys
0	User activity location cannot be inferred because origin or destination location are not known
1	User is at primary location
2	User is at secondary location
...	User is located at area ...



(a)

(b)

Fig. 2. Illustration of individual activity sequences. (a) Full Activity Sequence. (b) Simplified (durationless) activity sequence.

journeys themselves. Based on the approach described in [11], we obtain, for each passenger, a sequence of activity locations over 29 days. Each activity is associated with a location status value defined in Table II. While the activity purpose linked to each location is not explicitly inferred, the locations visited by a user are ordered on the amount of time spent at each location. Hence, the user's primary location aligns with the area in which the user spent the most time and the secondary location with the area where they spend the second most time. The user's location cannot always be inferred, either on days with no travel or because of unobserved trips made on other modes. Special non-location indices 0 and -1 are used to account for these cases. Consecutive days with no travel are represented by a single '0' status code. We do not consider activity duration for this application. If we consider other attributes, it would add to the granularity of travel events, and likely increase both the entropy and entropy rate of the sequence.

Fig. 2b illustrates the resulting sequence of activities completed by the user represented in Fig. 2a. Note that public transit trip events are excluded from the sequences for this particular case study as they always occur before a new area is visited. In general, trip events based on their attributes (e.g. mode, route, or duration) can be incorporate in the sequences with the activity events. In this way, an individual's compound behavior of where to go and how to get there can be examined in a single travel sequence.

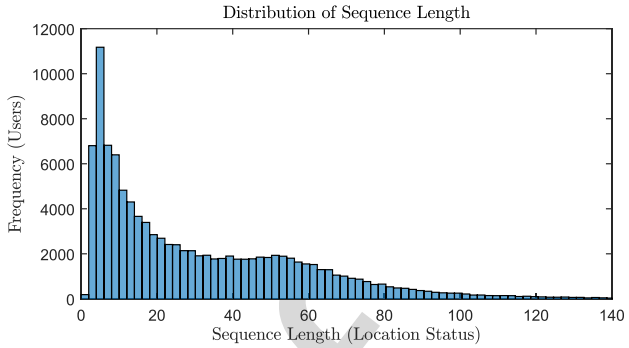


Fig. 3. Distribution of sequence length

The entropy $\hat{H}(X)$ and entropy rate $\hat{H}(\mathbf{X})$ associated with an observed user sequence x_1^n containing n events is estimated as described in Section IV. $\hat{H}(X)$ is computed according to (1), with the probability $p(x)$ of an event $x \in E$ estimated by

$$\widehat{p}(x) = n_x / n \tag{7}$$

where n_x represents the number of occurrences of x in the observed sequence X_1^n and n represents the length of the sequence observed. Fig. 3 shows the distribution of the sequence length for the sample of 99 925 users. Some users in the sample completed few trips over the 29-day period, and their intrapersonal variability cannot be analyzed from their smart card records. Consequently, all user-sequences shorter than 10 events are excluded from the regularity analysis presented next. The resulting sample contains 76 838 user-sequences.

The entropy and entropy rate distributions estimated from these sequences are presented in Fig. 4a and 4b respectively. The entropy rate $\hat{H}(\mathbf{X})$ of the sequence is estimated using the BWT method described in Section IV-B. As previously discussed, the value of entropy $\hat{H}(X)$ is equivalent to the entropy rate of a sequence with no memory (or for which the order of events is ignored). The entropy distribution has a mean of 2.5 bits and a standard deviation of 0.53 bits. As a reference, a fair coin toss has entropy of 1, and a fair six-sided dice roll has entropy of 2.6. Hence, on average, without accounting for the information provided by the order of events, a user-sequence is almost as random as a fair dice roll. Users at the low-end of the distribution tend to visit a few locations repeatedly, and are therefore more predictable, while those at the high end of the distribution visit many locations and are more unpredictable. An individual who traveled exclusively between home and work ($p(home) = p(work) = 0.5$) has an entropy of 1 bit, akin to a coin toss.

In contrast, the entropy rate distribution has a mean of 1.4 bits/event and a standard deviation of 0.42 bits/event. The 1.1-bit difference between the mean entropy and the mean entropy rate reflects the additional information provided by the order in which events take place. Considering the order in which events are generated, an average user-sequence is associated with only slightly more uncertainty than a coin toss. In others words, on average, the next event can be predicted

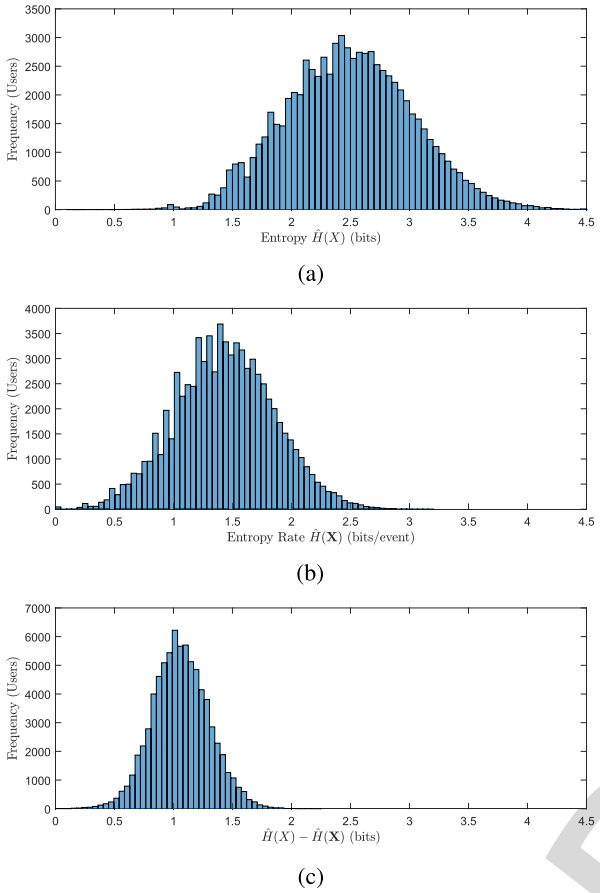


Fig. 4. Distribution of entropy measures across users. (a) Distribution of entropy. (b) Distribution of entropy rate. (c) Distribution of $(\hat{H}(X) - \hat{H}(\mathbf{X}))$.

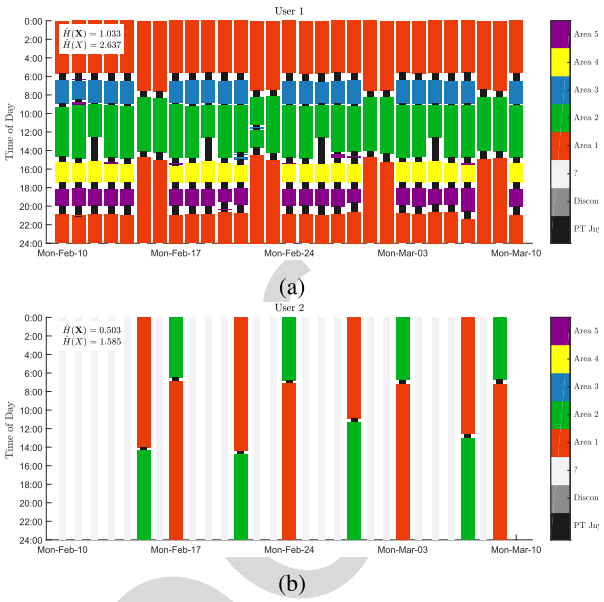


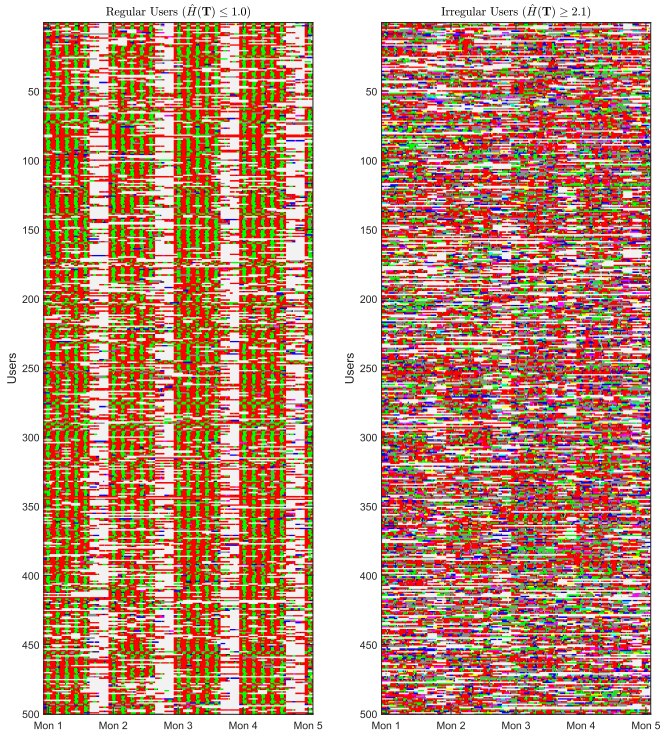
Fig. 5. Example location sequences of two users. (a) Activity sequence of user 1. (b) Activity sequence of user 2.

accurately almost 1 in 2 times when the order of events is considered, and only when 1 in 6 times when the order is not captured. Of course, the order of events does not provide the same amount of information for all individuals. As illustrated in Fig. 4c, for some users, the order of events provides almost no information, while for others it reduces the uncertainty by as much as 2 bits/event. Specifically, individuals who visit many locations frequently, but always in the same order will have relatively high entropy, but relatively low entropy rate. For reference, the individual used as an example earlier who traveled exclusively between home and work would have an entropy rate of 0 bit/event (as every new event is exactly determined by the previous one). In contrast, a coin toss has an entropy rate of 1 bit/event (same as its entropy) as there is no sequential dependency between events. Any individual whose travel pattern was exactly repeated over time would have an entropy rate of 0.

Fig. 5a illustrates the value of the entropy rate for a specific user who visited 5 locations almost exactly the same number of times, but consistently in the same order over the month-long observation period. The estimated entropy of the travel sequence of the user is 2.6, while its entropy rate is 1.0. The resulting 1.6 bit difference $\hat{H}(X) - \hat{H}(\mathbf{X})$ for this individual is at the high-end of the distribution shown in Fig. 4c. Additionally, while this individual's routine is not conventional, it is clearly regular as both the events and the order in which they are combined are repeated over time. This is reflected by the below average entropy rate of this sequence.

Fig. 5b illustrates another example of non-workday regularity captured by the entropy rate measure. On four separate occasions, the corresponding individual traveled from the primary location (i.e. red) to the secondary location (i.e. green) and made the reverse trip after one or two days without travel. In this specific example, the secondary location includes a terminal rail station, which the individual likely uses to leave London for the weekend. The user sometimes leaves on Friday, sometimes on Saturday and returns either on the following Sunday or Monday. Even though the pattern spans several days, and is not repeated periodically, its regularity is captured by the entropy rate of the sequence estimated to 0.503, below the sample average. This pattern would not accurately be captured by the standard periodicity measures reported in the literature, as it does not reoccur on the same days of the week from week to week.

Examination of other individuals shows that, in general, the entropy rate measures regularity accurately and can serve as a useful comparison metric. Fig. 6 compares two groups of 500 users whose sequence is longer than 40 events. The rows represent the sequence of an individual, while the columns correspond to different times of the 29 day period. The first group is randomly selected from all users with entropy rate below 1.0 bit. These regular users fall below the 10^{th} percentile of the entropy rate distribution for sequences longer than 40 events. The second group is randomly selected from all users with entropy rate above 2.1 bits. These irregular users fall above the 90^{th} percentile of the distribution. As expected, the sequences associated with regular users are characterized by the conventional working week structure. The irregular sequences contain much less repeated structure. It is important to note that while the dominant pattern in Fig. 6a



(a) (b)

Fig. 6. Comparison of users in the lower and upper 10^{th} percentile of irregularity. (a) Regular Users. (b) Irregular Users.

is associated with the typical working week, many non-conventional patterns, such as those illustrated above, are also qualified as regular. This demonstrates that the entropy rate can be used as an indicator of regularity, capturing the extent of repetition in events and in the order in which they appear, while not making any assumption about how the repetition fits with conventional calendar cycles such as a day or a week.

VI. Discussion

Regularity is an important property of individual travel behavior, and the ability to measure it is valuable for advances in behavior modeling, mobility prediction, and customer segmentation. First, our study shows that much of the uncertainty in travel choices (such as location choice) can be accounted for by considering the order of these choices. The difference between entropy and entropy rate can be used as a measure of the potential value of incorporating sequential dependency in behavior modeling. Second, regularity is closely tied with the concept of predictability. As shown by [32], the entropy rate makes it possible to compute a fundamental limit of predictability of individual travel behavior, which can be used to evaluate predictive behavior models. Third, regularity is one of the metrics that shed light on a person's lifestyle, because it captures patterns in the overall organization of these behavior components. For example, Fig. 6 shows two groups of users, one with consistent itineraries (mostly commuters) and one with flexible schedules. This makes the measure of regularity a useful metric for user segmentation. Finally, the proposed methodology is highly flexible and can be adapted for different scenarios. Representing behavior as a sequence of events makes it computationally convenient to measure and analyze certain properties of human behavior that would be difficult to quantify otherwise. This is particularly fitting for new mobility data sources (e.g. smart card data) which typically provide long series of event-driven observations of individual behavior with no semantic annotations (e.g. travel purposes or activity types from survey data). It is also possible to adapt our regularity measure to study other types of human behavior using other sources of data (telecommunication behavior using mobile phone data, shopping behavior using credit card data, etc.)

VII. Conclusion

This paper provides an in-depth discussion of regularity of human travel behavior. We hypothesize that the order in which an individual engages in trips and activities constitutes an integral characteristic of human travel behavior and that this characteristic should be captured in the definition of regularity. We present a measure of regularity based on entropy rate which is sensitive to the frequency of travel events and to the order in which events are observed. To apply this measure, we also propose a framework to represent individual travel behavior as a sequence of travel events. The methodology is demonstrated using a large sample of transit smart card records from London, U.K. The Burrows-Wheeler transform is used for the estimation of the entropy rate. The results show that on average the next travel event can be predicted accurately almost 1 in 2 times when the order of events is considered, and only 1 in 6 times when the order is not considered. They also confirm the hypothesis that the order of travel events is important and captures a component of regularity not considered in the periodicity-based methods. Furthermore, the findings reveal that travel regularity may follow atypical patterns which are not captured by either periodicity-based methods or activity-based models. The regularity measure we propose is useful to reveal such patterns through data mining because it does not require assumptions about the periodic interval or the structure of regularity in travel behavior. It is also flexible and hence, can be adapted to study other types of human behavior using similar types of traces.

References

[1] D. Adjeroh, T. Bell, and A. Mukherjee, *The Burrows-Wheeler Transform: Data Compression, Suffix Arrays, and Pattern Matching*. Boston, MA, USA: Springer, 2008.

[2] M. Allahviranloo and W. Recker, "Daily activity pattern recognition by using support vector machines with multiple classes," *Transp. Res. B, Methodol.*, vol. 58, pp. 16–43, Dec. 2013.

[3] C. R. Bhat, S. Srinivasan, and K. W. Axhausen, "An analysis of multiple interepisode durations using a unifying multivariate hazard model," *Transp. Res. B, Methodol.*, vol. 39, no. 9, pp. 797–823, Nov. 2005.

[4] J. L. Bowman and M. E. Ben-Akiva, "Activity-based disaggregate travel demand model system with activity schedules," *Transp. Res. A, Policy Pract.*, vol. 35, no. 1, pp. 1–28, Jan. 2001.

[5] R. N. Buliung, M. J. Roorda, and T. K. Remmel, "Exploring spatial variety in patterns of activity-travel behaviour: Initial results from the Toronto Travel-Activity Panel Survey (TTAPS)," *Transportation*, vol. 35, no. 6, pp. 697–722, Aug. 2008.

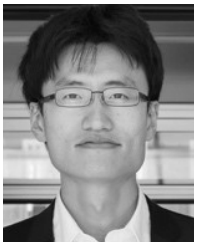
[6] H. Cai, S. R. Kulkarni, and S. Verdu, "Universal entropy estimation via block sorting," *IEEE Trans. Inf. Theory*, vol. 50, no. 7, pp. 1551–1561, Jul. 2004.

[7] M. Chikaraishi, A. Fujiwara, J. Zhang, and K. Axhausen, "Exploring variation properties of departure time choice behavior by using multilevel analysis approach," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2134, pp. 10–20, Dec. 2009.
[8] N. Eagle and A. Pentland, "Reality mining: Sensing complex social systems," *Pers. Ubiquitous Comput.*, vol. 10, no. 4, pp. 255–268, Mar. 2006.
[9] Y. Gao, I. Kontoyiannis, and E. Bienenstock, "Estimating the entropy of binary time series: Methodology, some theory and a simulation study," *Entropy*, vol. 10, no. 2, pp. 71–99, Jun. 2008.
[10] J. B. Gordon, H. N. Koutsopoulos, N. H. Wilson, and J. P. Attanucci, "Automated inference of linked transit journeys in London using fare-transaction and vehicle location data," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2343, pp. 17–24, Sep. 2013.
[11] G. Goulet-Langlois, H. N. Koutsopoulos, and J. Zhao, "Inferring patterns in the multi-week activity sequences of public transport users," *Transp. Res. C, Emerg. Technol.*, vol. 64, pp. 1–6, Mar. 2016.
[12] K. M. N. Habib and E. J. Miller, "Modelling daily activity program generation considering within-day and day-to-day dynamics in activity-travel behaviour," *Transportation*, vol. 35, no. 4, pp. 467–484, May 2008.
[13] S. Hanson and J. O. Huff, "Assessing day-to-day variability in complex travel patterns," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 891, pp. 18–24, Dec. 1982.
[14] S. Hanson and O. J. Huff, "Systematic variability in repetitious travel," *Transportation*, vol. 15, nos. 1–2, pp. 111–135, Mar. 1988.
[15] J. O. Huff and S. Hanson, "Repetition and variability in urban travel," *Geograph. Anal.*, vol. 18, no. 2, pp. 97–114, Apr. 1986.
[16] C.-H. Joh, T. Arentze, F. Hofman, and H. Timmermans, "Activity pattern similarity: A multidimensional sequence alignment method," *Transp. Res. B, Methodol.*, vol. 36, no. 5, pp. 385–403, Jun. 2002.
[17] H. Kang and D. M. Scott, "Exploring day-to-day variability in time use for household members," *Transp. Res. A, Policy Pract.*, vol. 44, no. 8, pp. 609–619, Oct. 2010.
[18] L. M. Kieu, A. Bhaskar, and E. Chung, "Passenger segmentation using smart card data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1537–1548, Jun. 2015.
[19] M. Kim and D. Kotz, "Periodic properties of user mobility and access-point popularity," *Pers. Ubiquitous Comput.*, vol. 11, no. 6, pp. 465–479, 2006.
[20] R. Kitamura and T. V. D. Hoorn, "Regularity and irreversibility of weekly travel behavior," *Transportation*, vol. 14, no. 3, pp. 227–251, Sep. 1987.
[21] R. Kitamura, T. Yamamoto, Y. O. Susilo, and K. W. Axhausen, "How routine is a routine? An analysis of the day-to-day variability in prism vertex location," *Transp. Res. A, Policy Pract.*, vol. 40, no. 3, pp. 259–279, Mar. 2006.
[22] Z. Li, J. Wang, and J. Han, "Mining event periodicity from incomplete observations," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 444–452.
[23] C. Morency, M. Trpanier, and B. Agard, "Measuring transit use variability with smart-card data," *Transp. Policy*, vol. 14, no. 3, pp. 193–203, May 2007.
[24] T. Neutens, M. Delafontaine, D. M. Scott, and P. De Maeyer, "An analysis of day-to-day variations in individual spacetime accessibility," *J. Transp. Geography*, vol. 23, pp. 81–91, Jul. 2012.
[25] E. I. Pas, "Intrapersonal variability and model goodness-of-fit," *Transp. Res. A, Gen.*, vol. 21, no. 6, pp. 431–438, Nov. 1987.
[26] E. I. Pas and F. S. Koppelman, "An examination of the determinants of day-to-day variability in individuals' urban travel behavior," *Transportation*, vol. 13, no. 2, pp. 183–200, Jun. 1986.
[27] F. Primerano, M. A. P. Taylor, L. Pitaksringkarn, and P. Tisato, "Defining and understanding trip chaining behaviour," *Transportation*, vol. 35, no. 1, pp. 1–2, Jan. 2008.
[28] W. W. Recker, M. G. McNally, and G. S. Root, "Travel/activity analysis: Pattern recognition, classification and interpretation," *Transp. Res. A, Gen.*, vol. 19, no. 4, pp. 279–296, Jul. 1985.
[29] M. J. Roorda and T. Ruiz, "Long- and short-term dynamics in activity scheduling: A structural equations approach," *Transp. Res. A, Policy Pract.*, vol. 42, no. 3, pp. 545–562, Mar. 2008.
[30] J. Scheiner, "The gendered complexity of daily life: Effects of life-course events on changes in activity entropy and tour complexity over time," *Travel Behaviour Soc.*, vol. 1, no. 3, pp. 91–105, Sep. 2014.
[31] S. Schönfelder, "Urban rhythms: Modelling the rhythms of individual travel behaviour," Ph.D. dissertation, Inst. Transp. Planning Syst., ETH Zurich, Zürich, Switzerland, 2006.
[32] C. Song, Z. Qu, N. Blumm, and A.-L. Barabsi, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, Feb. 2010.
[33] F. M. J. Willems, "The context-tree weighting method: Extensions," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 792–798, Mar. 1998.
[34] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653–664, May 1995.
[35] M. J. Williams, R. M. Whitaker, and S. M. Allen, "Measuring individual regularity in human visiting patterns," in *Proc. Int. Conf. Soc. Comput. (SocialCom) Privacy, Secur., Risk Trust (PASSAT)*, Sep. 2012, pp. 117–122.

Gabriel Goulet-Langlois received the M.Sc. degree in transportation from MIT in 2015. As part of his research, he developed methods to analyze travel patterns and user behavior from large ticketing data sets. He is currently building on this experience in a practical context as a Data Scientist with Transport for London. He is dedicated to improving customer experience and public transport planning through better use of data and technology.

Haris N. Koutsopoulos is currently a Professor with the Department of Civil and Environmental Engineering, Northeastern University, Boston, and a Guest Professor with the KTH Royal Institute of Technology, Stockholm. His current research interests include on the use of data from opportunistic and dedicated sensors to improve planning, operations, monitoring, and control of urban transportation systems, including public transportation. He is the Founder of the iMobility lab, which uses Information and Communication Technologies to address urban mobility problems. The Laboratory received the IBM Smarter Planet Award in 2012.

Zhan Zhao received the B.Eng. degree from Tongji University in 2011 and the master's degree in applied science from the University of British Columbia in 2013. He is currently pursuing the Ph.D. degree in interdepartmental doctoral program in transportation with the Massachusetts Institute of Technology (MIT), and a Graduate Research Assistant with the MIT Transit Lab. His research interests include travel behavior modeling, public transportation systems, and urban computing.



Jinhua Zhao is the Edward H. and Joyce Linde Associate Professor with City and Transportation Planning, MIT. He brings behavioral science and transportation technology together to shape travel behavior, design mobility systems, and reform urban policies. He directs the MIT Urban Mobility Lab (mobility.mit.edu) and MIT Transit Lab.

Measuring Regularity of Individual Travel Patterns

Gabriel Goulet-Langlois, Haris N. Koutsopoulos, Zhan Zhao, and Jinhua Zhao

***Abstract*— Regularity is an important property of individual travel behavior, and the ability to measure it enables advances in behavior modeling, mobility prediction, and customer analytics. In this paper, we propose a methodology to measure travel behavior regularity based on the order in which trips or activities are organized. We represent individuals' travel over multiple days as sequences of "travel events"—discrete and repeatable behavior units explicitly defined based on the research question and the available data. We then present a metric of regularity based on entropy rate, which is sensitive to both the frequency of travel events and the order in which they occur. The methodology is demonstrated using a large sample of pseudonymised transit smart card transaction records from London, U.K. The entropy rate is estimated with a procedure based on the Burrows-Wheeler transform. The results confirm that the order of travel events is an essential component of regularity in travel behavior. They also demonstrate that the proposed measure of regularity captures both conventional patterns and atypical routine patterns that are regular but not matched to the 9-to-5 working day or working week. Unlike existing measures of regularity, our approach is agnostic to calendar definitions and makes no assumptions regarding periodicity of travel behavior. The proposed methodology is flexible and can be adapted to study other aspects of individual mobility using different data sources.**

***Index Terms*— Regularity, intrapersonal variability, travel behavior, smart card data, entropy rate.**

I. Introduction

TRAVEL behavior is dynamic and varies across individuals but also for the same person over time. Interpersonal variability refers to the heterogeneous spatiotemporal preferences of people, reflecting different sociodemographic attributes, home/work locations, and lifestyle preferences [26]. Intrapersonal variability describes longitudinal variability in the characteristics of the same individual's travel behavior from trip to trip, day to day, or week to week [13], [26], [31]. Sometimes it is referred to in the literature as intraindividual [15], or day-to-day variability [17], [21], [24]. Regularity refers to the extent to which individual travel behaviors repeat over time. A person's activity choices and their associated trips are not made randomly. According to activity-based travel theory, they are dictated by preferences, constraints, and needs which recur over time to some degree [20].

While conventional cross-sectional data, one-day travel diary surveys for example, can capture the interpersonal variability, measuring intrapersonal variability/regularity requires individual-level longitudinal data. Multi-day travel surveys, often used for activity-based modeling, provide such data but are costly to collect and hence usually constrained to small sample sizes and short observation periods. However, advances in urban sensing technologies afford the opportunity to collect traces of individual mobility on a large scale and over extended periods of time. New mobility data sources, such as mobile phone records and transit smart card records, enable detailed and reliable measurement of travel regularity. No existing definition and measure of behavior regularity align with the variety in people's routines and granularity which these new data sources can capture.

Central to the definition of regularity is the definition of a unit of analysis for which repetition is considered. This unit should be chosen in line with the attributes relevant to the research question of interest and consistent with the resolution of the available sensor data. Reference [15] use the term *behaviors* to describe components of travel behavior characterized by combinations of attributes, for example "driving a car to work". In this paper, we use the term "travel events" to refer to the same concept as [15]'s *behaviors*, but with a broader connotation. A travel event is a repeatable unit describing individual travel behavior, characterized by one or more attributes such as purpose, location, and duration. At the most basic level, a travel event is either a trip or an activity. Travel events can also be aggregated to different levels (e.g. daily or weekly) to form higher-level travel events. For example, for the analysis of individual daily routines, a travel event may be a combination of activities in one day. In this paper, if not specified otherwise, "travel events" are used to refer to the most basic building blocks of travel behavior—trips and activities.

Travel events do not occur in isolation. People's activity patterns govern the co-occurrence of multiple travel events. This is the basis of work on trip chaining behavior, e.g. [27], and activity-based models, e.g. [4]. Combinations of travel events reflect such activity patterns. Each event must be considered as part of this context. While some travel events are frequently repeated over time, their surrounding contexts may change from day to day [15]. This highlights that regularity depends not only on variability in the characteristics of a

Manuscript received April 2, 2017; revised July 11, 2017; accepted July 16, 2017. This work was supported by Transport for London. The Associate Editor for this paper was H. S. Mahmassani. *(Corresponding author: Jinhua Zhao.)*

G. Goulet-Langlois was with the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA. He is now with Transport for London, London SW7 2NJ, U.K.

H. N. Koutsopoulos is with the Department of Civil and Environmental Engineering, Northeastern University, Boston, MA 02115 USA.

Z. Zhao is with the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

J. Zhao is with the Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: jinhua@mit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2017.2728704

single event but also on the pattern in which multiple events are combined. In our approach, multiple travel events can be ordered over time and form "travel sequences".

In existing literature, some methods have been proposed to measure regularity by examining the periodic patterns of travel behavior [19], [32], [35]. However, periodicity is not equivalent to regularity. While periodicity only captures the cyclic repetitions of travel events at fixed time intervals (typically set as a day or a week), regularity refers to all forms of repetitions. Travel patterns may not necessarily repeat periodically or may repeat over unconventional periods not aligned with the typical day or week. To some extent, periodicity is a special type of regularity. The order in which an individual completes trips and activities is an integral component of the structure in their travel routines. A good metric of regularity should be sensitive to such sequential dependency in a travel sequence, without a predefined periodic cycle.

In this paper, we propose a new approach to measuring the regularity of travel behavior based on the order in which travel events are organized over time in travel sequences. The definition is not tied to an underlying calendar. Hence it is flexible. We demonstrate the approach using a large sample of transit smart card transaction records over a period of a month. The ability to measure regularity improves our understanding of travel behavior, facilitates advancements in behavior modeling, and enables the development of customer analytics for travel prediction, user segmentation, and targeted demand management.

The remainder of the paper is organized as follows. We present a literature review of the related work on intrapersonal variability/regularity in Section II. Section III proposes a sequential representation of travel behavior and develops its mathematical formulation. This is followed by a description of the proposed measure of regularity based on entropy rate in Section IV. The measure is demonstrated in Section V using smart card data from London, U.K. The paper is concluded with a discussion of future research directions and potential implications in Section VI, and a summary of the main findings in Section VII.

II. Literature Review

While the concept of travel behavior regularity is recognized as a critical dimension of travel behavior, approaches to measure such variability remain limited in scope. Specifically, many studies measure regularity based only on the extent to which single travel events are repeated, without consideration for how multiple events are combined. Some methods focus only on the relative frequency of trips. For example, [5] proposed a spatial repetition index corresponding to the percentage of activity locations which are visited more than once over a 7 day period. Based on survey data, this measure is computed for different time periods to evaluate the spatial stability of individual activity patterns at different times of the week. Based on smart card data, [18] identified the OD pairs that the card holder frequently travels as "regular OD" and the time of the trips between these regular ODs as "habitual time". They measured the regularity of transit users based on the percentage of a user's trips completed within habitual times and between regular ODs. Reference [23], using smart card data, evaluated the level of spatial and temporal variability of different users based on the frequency of trips made to different stops at different times of the day.

Other studies rely on the variance of different measures to quantify longitudinal variability. References [25] and [26] evaluated the variance in number of trips per day from a 7-day travel survey. Their results differentiated the part of the variance of trip generation rates associated with intrapersonal variability from the part associated with interpersonal variability. Reference [21] analyzed variability in the departure time of the first trip of the day. Relying on the concept of individual space-time prisms, they modeled the variance of first departure time so as to differentiate the part of the variance due to randomness, from the part due to changes in the time constraints dictating an individual's schedule. Similarly, [7] also attempted to dissect the variance of the first trip departure time by formulating a multilevel model for which the variance was decomposed into five parts: inter-individual variation, inter-household variation, spatial variation, temporal variation, and intra-individual variation. Like the frequency-based measures, these variance-based measures treat each trip independently and are not concerned with the sequence of multiple trips.

Accounting for combinations of travel events has long been recognized in the literature of travel behavior modeling as important. Some models rely on the assumption that activity and trip combinations are primarily a function of days of the week. For example, using the 7-day Toronto Travel Activity Panel Survey, [12] modeled the frequency of 15 non-home/work activity categories for the 7 days of the week using 7 independent models. In contrast, some studies model the relationship between different travel events more explicitly. Reference [29] modeled preplanned and spontaneous activity duration as well as number of trips by mode, using data from the 7-day activity survey in [12]. Their approach introduces same-day effects and next-day effects to capture the relationship between multiple activities. From a long-term perspective, [3] examined the relationship between successive activities for the same purpose (e.g. shopping) using a 6-week travel survey from Karlsruhe, Germany. They modeled the time elapsed between successive activities using a multivariate hazard model. Other studies used pattern recognition techniques to directly model the activity sequence as a whole, and such techniques include Walsh-Hadamard transformation [28], sequence alignment [16], and conditional random field [2]. These studies account, to various degrees, for the relationship between travel events to improve travel demand models. They use panel survey data and do not aim at measuring regularity in the order of travel events over time.

To measure regularity in combinations of travel events, many researchers, especially in the human mobility literature, proposed methods to uncover periodic patterns. Some studies use the Fourier transform to identify underlying periods of repetition in travel from digital traces of location collected over multiple weeks. Reference [19] found daily and weekly periods to be most significant in observing individuals' connection

to Wi-Fi access points (AP) on the Dartmouth campus. Reference [8] identified the same dominant periods using data from MIT's Reality Mining project. Reference [22] proposed a probabilistic measure of periodicity and demonstrated its robustness to noise and missing observations using GPS data, with superior performance over methods based on the Fourier transform.

The above studies account for repetition in combinations of travel events, by measuring the extent to which their co-occurrence map to a set calendar cycle (most often a weekly cycle). Other studies attempt to measure regularity explicitly by imposing a predefined cyclic period. For example, [35] proposed a measure of temporal irregularity in the intervals between a person's visits to a given location. They applied a weekly based measure to different data sources and found that the behavior captured from smart card data was most regular, while Wi-Fi data revealed the least regularity. Reference [32] presented another regularity measure also based on a weekly cycle. Given hourly information of a person's location over several weeks, they used the percentage of hours spent at the location most frequently visited during each hour of the week as the index of periodicity for the corresponding hour.

However, periodicity is not the same as regularity. Regularity indicates the degree to which sub-sequences of events are repeated, and these sub-sequences do not have to align with a particular cycle. This is especially relevant to sequences of activities, as activities are likely to be organized in a logical order. For example, visiting the doctor's office, going to the pharmacy to pick-up a prescription, and returning home are likely to occur in this logical order. The repetition of this sequence may not be periodic. Furthermore, [19], [32], [35], and [8] all discuss periodicity in the context of the most conventional cycles of repetition: the day and the week. We argue that regularity is an internal property of a travel sequence and should not depend on how the sequence aligns with the calendar. Some patterns may repeat on non-daily or weekly cycles. For example, certain types of employment (e.g. shift-workers, firefighters, doctors) may dictate working schedules which repeat on a cyclical unit other than the week. Periodicity measures computed on a weekly basis (as done by [32] and [35]) would fail to capture the true regularity in such cases. Similarly, a measure of daily periodicity may not be able to capture patterns spanning more than a calendar day, such as going out in the evening, sleeping at a friend's home, and then returning home the next day.

In conclusion, no index that captures repetition in the order in which events are observed has been introduced in the literature. In the following sections, we present a new metric for measuring the regularity of travel behavior that depends explicitly on the order in which travel events occur. As such, the metric avoids the issues inherent in existing periodicity-based measures which examine only co-occurring patterns of travel events and calendar events (i.e. hour, day, week).

III. Sequence Representation

Individual travel patterns can be conceptualized as a sequence of travel events. These events unfold over time with respect to a background calendar (time of day, day of the week, month). Travel events are characterized by different aspects of behavior, including location, time of day, mode, route, travel time, activity type (or travel purpose) and activity duration. For instance, an event defined as an activity occurs at a certain time of day (8 pm on Friday), for a certain duration (2 hours), at a certain location (downtown) and for a certain purpose. As recognized by [13]–[15], variations along these behavioral dimensions are not independent. For example, an individual's choice of mode or route will significantly influence the travel time for her morning commute, which impacts her departure time.

A key component of these sequences is the order in which events take place. An appropriate measure of regularity in a person's travel behavior should capture both, the extent of repetition in travel events and in the order in which they are performed. It is necessary to introduce a mathematical representation of travel sequences which captures the order of events to define such a regularity index. We model the mobility of each individual over multiple days as a random process, which represents how often and in what order travel events are generated. The notation follows that used by [9].

Let the stochastic process corresponding to the mobility of a given individual u be denoted by $\mathbf{X}_u$ and a travel event generated by this process by random variable X_u. Each travel event X_u assumes a discrete value x from the set of possible travel event outcomes E_u defined for individual u. x can be regarded as a unique identifier for a repeatable event. Two separate events assume the same value of x if and only if they have the same combinations of event attributes. X_u has a discrete probability distribution $p(x) = Pr\{X_u = x\}$ for $x \in E_u$.

For simplicity, subscript u is omitted and all remaining notation is defined with respect to a single individual. The stochastic process $\mathbf{X} = \{\ldots, X_{-1}, X_0, X_1, X_2, \ldots\}$ represents the ordered set of random variables X_i. Any finite sequence of this ordered set between event i and event j is denoted by the ordered subset $X_i^j = \{X_i, X_{i+1}, \ldots, X_{j-1}, X_j\}$, with $-\infty < i \leq j < \infty$ such that $X_i^j \subset \mathbf{X}$. Given a finite window of analysis, we observe a specific realization $x_i^j = \{x_i, x_{i+1}, \ldots, x_{j-1}, x_j\}$ of the finite random variable sequence X_i^j.

Informally, set E is akin to an alphabet from which a string of discrete events can be constructed. Different types of sequences, or strings, can be represented based on different definitions of travel events $x \in E$, driven by the aspects of behavior of interest. In practice, the specification of E is constrained by the available data. Different data provides information on varying aspects of travel and at various aggregation levels. For instance, smart card data provides location information at the stop level and the timing of the event, but no direct information on activity purpose.

For consistency and computation convenience, we assume all event attributes are discrete. This assumption is common for travel behavior analysis since many travel attributes are discrete by nature, such as purpose, location and time periods (e.g. morning peak, midday, afternoon peak). Attributes

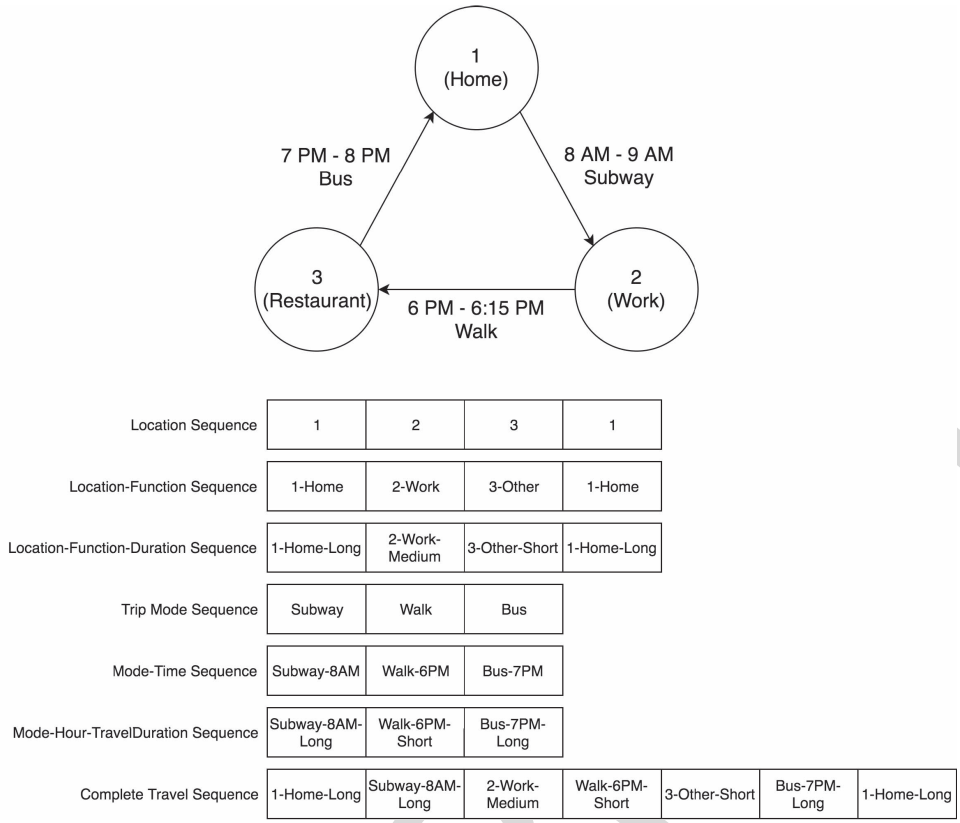


Fig. 1. Example of travel sequences.

that typically assume continuous values (e.g. activity duration) are discretized into a finite number of categories. The specification of these categories depends on both, the goal and the data of the analysis. While a larger number of categories can capture the variation of these attributes in finer detail, it can also make the specific values less repeatable and lead to a sparse distribution of $p(x)$. Ideally, these categories should meaningfully reflect behavioral choices. For example, using some clustering approach (e.g. Gaussian mixture model), the activity duration can be discretized into three categories - long, medium, short, and each of these categories is likely to be associated with certain activity types (e.g. home, work, other).

Fig. 1 shows how a person's travel over a day can be summarized as different travel sequences by changing the definition of travel events. For this example, we discretize activity duration into three categories - Long ($>$ 10 hours), Short ($<$ 3 hours) and Medium (between 2 and 10 hours), and travel duration into two categories - Long ($>$ 30 minutes) and Short ($<$ 30 minutes). We also characterize the trip start time using 24 hourly intervals. The level of discretization determines the granularity of travel events. Typically, finer granularity means that each travel event is more unique and less likely to repeat.

For many applications, a single aspect of travel behavior (i.e. purpose, location, or mode) is relevant. In these cases, the travel events only have a single attribute, and we may directly set the x value of an event to its attribute value. For example, the first sequence in Fig. 1 focuses on the locations visited by the person. This can be represented by defining set E as the set of all locations visited by the individual over the period of analysis. In this example, x_i^j is simply a series of location IDs.

In other contexts, it may be necessary to define events based on combinations of multiple attributes. For instance, location, function, and duration could be combined to differentiate between two activities observed in the same geographical area. In this case, the events x in set E are defined as compound outcomes of location, function, and purposes, as illustrated in the third sequence of Fig. 1.

At different levels of aggregation, multiple trips or activities can be grouped together to define a single event. For example, all trips made on the same day can be grouped into a single event to create a binary sequence representing when the person traveled across multiple days.

This representation provides a flexible approach to simplify and represent multidimensional travel behavior as a string of travel event symbols. These symbols are defined in line with the objective of the study so as not to distort or omit relevant information about aspects of travel of interest.

IV. Measurement of Regularity

As described in the previous section, we model the mobility of an individual over multiple days as a sequence of events generated by a random process $\mathbf{X}$. Through this abstraction, it is possible to characterize an individual's mobility by quantifying the nature of the random process $\mathbf{X}$. Many different

properties of process $\mathbf{X}$ may provide information about the individual's travel pattern. For example, consider a process $\mathbf{X}$ representing the activity sequence of an individual. In this case, the cardinality of set E informs us about the diversity of activities in which the individual engages, and the mode of probability distribution $p(x)$ reveals the individual's most frequent activity. This section introduces ways to measure such properties of $\mathbf{X}$ which can be used to describe regularity of a travel sequence.

A. Entropy vs Entropy Rate

First, we examine the extent of repetition of a travel sequence regardless of the order. Under this assumption, the regularity of a random process is solely determined by the probability distribution $p(x)$. Intuitively, on average, an outcome generated by a more regular process should be less uncertain and more predictable. In information theory, the level of randomness or unpredictability of a process can be measured using entropy. Entropy measures the average information, or surprise, provided by each realization of a random variable in bits. The entropy $H(X)$ of random variable X with probability distribution $p(x) = Pr\{X = x\}$ for $x \in E$ is defined by (1).

$$H(X) = -\sum_{x \in E} p(x) \log_2 p(x) \tag{1}$$

For the travel sequence problem, X represents the random variable associated with a travel event and E denotes the set of all possible travel event outcomes defined for a given individual. Entropy can be thought of as a measure of variance defined for categorical probability distributions. It accounts for both the number of possible outcomes (the cardinality of set E) and the relative frequency of outcomes. Hence, entropy equals 0 for a process with a single possible outcome (no uncertainty) and is highest when the probability distribution of a random variable with multiple outcomes is uniform (when all events are equally likely). Reference [30] used entropy to measure and contrast the complexity of activity patterns completed by individuals of different gender. The author points out that entropy is a good measure of the amount of heterogeneity in a categorical distribution, which is especially relevant when considering qualitative outcomes such as activities.

Although entropy is a good measure of repetition of isolated events in a travel sequence, it does not capture the extent to which ordered sub-sequences of events repeat over time. Travel sequences are not typically memoryless processes. Rather, the conditional distribution of an event X_i depends on the outcome of events $X_{i-1}, X_{i-2}, \ldots$ preceding it (i.e $p(X_i|X_{i-1}, X_{i-2}, \ldots) \neq p(X_i)$). For example, observing a visit to the doctor might significantly increase the likelihood of a visit to the pharmacy in the following event. Entropy rate accounts for the order of events in a travel sequence, or more formally for the memory in process $\mathbf{X}$. Entropy rate $H(\mathbf{X})$ of the random process $\mathbf{X}$ is defined as the asymptotic rate at which the entropy of sub-sequence X_1^n changes with increasing n [9], calculated using (2).

$$H(\mathbf{X}) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, X_3, \ldots, X_n) \tag{2}$$

where, $H(X_1, X_2, X_3, \ldots, X_n)$ denotes the entropy of the joint variable X_1^n defined for the subsequence $X_1, X_2, \ldots, X_n$. References [9] and [6] stated that this limit exists for all stationary random processes and is equal to

$$\begin{aligned} H(\mathbf{X}) &= \lim_{n \to \infty} H(X_n|X_{n-1}, \ldots, X_2, X_1) \\ &= \lim_{n \to \infty} -\sum_{x_1^n \in E^n} p_n(x_1^n) \log_2 \frac{p_n(x_1^n)}{p_n(x_1^{n-1})} \end{aligned} \tag{3}$$

where p_n denotes the joint probability distribution of a sub-sequence of length n. As described by (2) and (3), entropy rate measures the average entropy of each new event generated by random process $\mathbf{X}$, accounting for preceding events. It is measured as the entropy per event and has units in bits per event. The entropy rate of a random process with no memory is exactly equivalent to the entropy of the process as each new event is independent of the previous. As such, the entropy of a process is an upper bound for its entropy rate. In contrast, a process in which the outcome of an event X_i is perfectly determined by the previous events ($p(X_i = x|X_{i-1}, X_{i-2}, \ldots) = 1$) has an entropy rate of 0. Informally, entropy rate is the average measure of information, or surprise, associated with each additional event generated in a sequence of events. The more memory in a random process, the more information the previous events provide about the next event, and therefore the lower the entropy rate of the process. Also, memory in the random process is directly related to the order in which events are observed. Specifically, the more memory in a random process, the more the order of the events it generates tends to repeat. In line with these characteristics, the entropy rate is a good regularity measure of travel sequences because it is sensitive to not only the relative frequency of events but also the dependencies between multiple events.

Reference [32] used the entropy rate of hourly-location sequences derived from cell phone data to explore predictability in individual location patterns. Hourly-location sequences tend to have very low entropy rate because the location of a person during a given hour is highly related to their location in the previous hour. This is because individuals tend to visit a location for several hours consecutively (e.g. 8 hours at work or 14 hours at home). For these sequences, longer average activity durations are associated with low entropy rate. Hence, the high predictability reported by [32], albeit an interesting theoretical finding, is of limited practical use because it merely reflects the tendency of individuals to stay in a location for multiple hours. Nevertheless, their approach demonstrates how entropy rate can be used to quantify the dependencies between elements of the same sequence.

B. Estimation of Entropy Rate

Estimation of the entropy rate of a finite sequence can be computationally challenging. According to (3), the entropy rate is a function of the unknown joint probability distribution p_n of the sequence X_1^n. A naïve approach consists of estimating p_n from the observed frequency of combinations of symbols in X_1^n. This approach becomes computationally intractable as combinations of increasing length are considered.

Most entropy rate estimation approaches circumvent the issue of estimating p_n by relying on universal data compression algorithms [6], [9]. These algorithms, used to compress data generated from processes with unknown probability distributions and arbitrarily long memories, are known to achieve optimal lossless compression ratios (i.e. compression ratio equal to the entropy of the generating process). Hence, they can be used to estimate the amount of redundant, or repeated information in a sequence of symbols. For instance, text can be compressed by coding frequently repeated expressions. If for example, the 28-symbol phrase "the probability distribution" frequently occurs in the text, it can be coded by a single symbol.

Three families of lossless compression methods have been applied to entropy rate estimation. References [34] and [33] introduced the context-tree weighting (CTW) based entropy estimator. Reference [6] developed an estimation approach based on the Burrows-Wheeler transform (BWT), and [9] proposed different estimators based on the Lempel-Ziv (LZ) family of data compression algorithms. The estimators perform differently with respect to efficiency and bias depending on the property of the source and the sequence realization considered. Longer sequence realizations provide entropy estimates with smaller variance, and the variance of different approaches converges at different rates. The size of the alphabet also influences accuracy, with larger alphabets resulting in both, higher variance and potentially higher bias for an equal number of observations. Reference [9] presented an extensive comparison of LZ and CTW estimators using simulation for binary sequences. They concluded that the CTW estimator consistently provides more accurate and reliable results than LZ-based estimators. Reference [6] established an upper bound on the convergence rate of the BWT estimator for finite-alphabet, finite-memory processes and demonstrated that the BWT estimator performs better than an LZ-based estimator for binary sequences. No direct comparison of the CTW and BWT estimates has been reported in the literature.

The BWT estimator is simpler to implement. Hence, the BWT entropy estimator with uniform segmentation, as described by [6], is used for the case study presented in Section V. The authors prove almost-sure convergence of this estimator for stationary, ergodic random processes. These properties are assumed to hold for travel sequences described by the formulation previously introduced. Specifically, we assume that the underlying characteristics of an individual's mobility do not change over the period for which the individual is observed. This assumption would be violated if a long-term change (e.g. change in residential location or job) took place during the period of analysis.

The BWT entropy estimator is computed in two steps. First, the Burrows-Wheeler transform is applied to the finite sequence X_1^n of length n. Reference [1] provided an in-depth discussion of the transform, its properties, and implementation. Table I, adapted from [1], illustrates how the BWT operates. The BWT is applied to an example sequence *aardvark*, resulting in the transformed sequence *kavraad*. First, all rotations of the input sequence are listed and sorted alphanumerically. Then, the last symbol of each rotation is retained. BWT groups together outcomes (or symbols) which occur in similar contexts in the original sequence.

TABLE I
AN EXAMPLE OF BWT (ADAPTED FROM [1])

All Rotations	Sorted Rotations	
aardvark	aardvar**k**	k
ardvarka	ardvark**a**	a
rdvarkaa	arkaard**v**	v
dvarkaar	dvarkaa**r**	r
varkaard	kaardva**r**	r
arkaardv	rdvarka**a**	a
rkaardva	rkaardv**a**	a
kaardvar	varkaar**d**	d

Formally, the BWT of any stationary process $\mathbf{X}$ with finite memory results in a piecewise memoryless sequence. Reference [6] leverage this property of the transformed output to estimate the entropy rate of the process that generated the original sequence. Specifically, in the second step of the estimation, the transformed sequence is segmented into S segments s of uniform length, and the distribution of outcomes is estimated for each segment according to (4).

$$\hat{q}(x,s) = \frac{N_s(x)}{\sum_{y\in E} N_s(y)} \tag{4}$$

where $N_s(x)$ denotes the number of occurrences of symbol x in segment s. Given $\hat{q}(x,s)$, the entropy of each segment s is estimated by (5). Finally, the entropy rate of $\mathbf{X}$ is estimated by the average entropy of all segments using (6).

$$\log_2 \hat{q}(s) = \sum_{x\in E} N_s(x) \log_2 \hat{q}(x,s) \tag{5}$$

$$\hat{H}(\mathbf{X}) = -\frac{1}{n} \sum_{s\in S} \log_2 \hat{q}(s) \tag{6}$$

Reference [6] recommend that the length of each segment s is set as the integer value closest to $\sqrt{n}$. As mentioned above, the accuracy of the resulting estimate depends on both the length of the sequence observed and the number of different outcomes it contains.

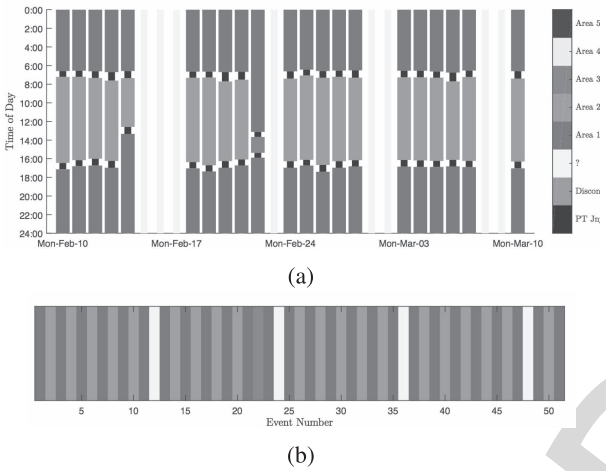
V. CASE STUDY

In this section, we demonstrate the proposed methodology described above using transit smart card data. Transport for London (TfL) provided the dataset used for this study. It consists of the smart card records of a sample of 99,925 pseudonymised cards observed between February 10th and March 10th 2014. The dataset covers Oyster transactions across all public transport modes including bus and rail. While the rail transactions contain entry and exit records with their associated stations and timestamps, the bus transactions only include boarding stop and time. Thus the alighting stop and time are inferred using the ODX method developed by [10]. ODX provides a set of complete public transport trips for each passenger.

For this case study, we are particularly concerned with the activities occurring between journeys, rather than the

TABLE II
ACTIVITY STATUS SUMMARY

Status	Semantics
-1	User activity location cannot be inferred because a non-PT trip was completed between observed PT journeys
0	User activity location cannot be inferred because origin or destination location are not known
1	User is at primary location
2	User is at secondary location
...	User is located at area ...



(a)

(b)

Fig. 2. Illustration of individual activity sequences. (a) Full Activity Sequence. (b) Simplified (durationless) activity sequence.

journeys themselves. Based on the approach described in [11], we obtain, for each passenger, a sequence of activity locations over 29 days. Each activity is associated with a location status value defined in Table II. While the activity purpose linked to each location is not explicitly inferred, the locations visited by a user are ordered on the amount of time spent at each location. Hence, the user's primary location aligns with the area in which the user spent the most time and the secondary location with the area where they spend the second most time. The user's location cannot always be inferred, either on days with no travel or because of unobserved trips made on other modes. Special non-location indices 0 and -1 are used to account for these cases. Consecutive days with no travel are represented by a single '0' status code. We do not consider activity duration for this application. If we consider other attributes, it would add to the granularity of travel events, and likely increase both the entropy and entropy rate of the sequence.

Fig. 2b illustrates the resulting sequence of activities completed by the user represented in Fig. 2a. Note that public transit trip events are excluded from the sequences for this particular case study as they always occur before a new area is visited. In general, trip events based on their attributes (e.g. mode, route, or duration) can be incorporate in the sequences with the activity events. In this way, an individual's compound behavior of where to go and how to get there can be examined in a single travel sequence.

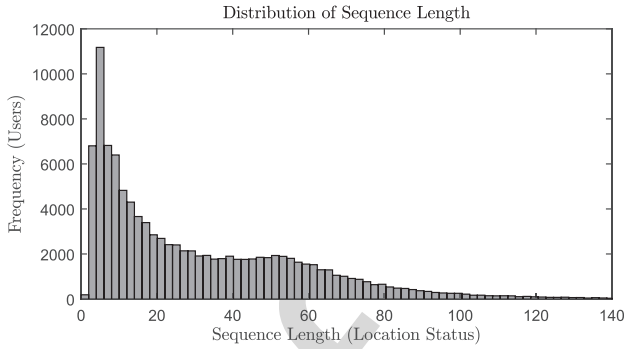


Fig. 3. Distribution of sequence length

The entropy $\hat{H}(X)$ and entropy rate $\hat{H}(\mathbf{X})$ associated with an observed user sequence x_1^n containing n events is estimated as described in Section IV. $\hat{H}(X)$ is computed according to (1), with the probability $p(x)$ of an event $x \in E$ estimated by

$$\widehat{p}(x) = n_x / n \tag{7}$$

where n_x represents the number of occurrences of x in the observed sequence X_1^n and n represents the length of the sequence observed. Fig. 3 shows the distribution of the sequence length for the sample of 99 925 users. Some users in the sample completed few trips over the 29-day period, and their intrapersonal variability cannot be analyzed from their smart card records. Consequently, all user-sequences shorter than 10 events are excluded from the regularity analysis presented next. The resulting sample contains 76 838 user-sequences.

The entropy and entropy rate distributions estimated from these sequences are presented in Fig. 4a and 4b respectively. The entropy rate $\hat{H}(\mathbf{X})$ of the sequence is estimated using the BWT method described in Section IV-B. As previously discussed, the value of entropy $\hat{H}(X)$ is equivalent to the entropy rate of a sequence with no memory (or for which the order of events is ignored). The entropy distribution has a mean of 2.5 bits and a standard deviation of 0.53 bits. As a reference, a fair coin toss has entropy of 1, and a fair six-sided dice roll has entropy of 2.6. Hence, on average, without accounting for the information provided by the order of events, a user-sequence is almost as random as a fair dice roll. Users at the low-end of the distribution tend to visit a few locations repeatedly, and are therefore more predictable, while those at the high end of the distribution visit many locations and are more unpredictable. An individual who traveled exclusively between home and work ($p(home) = p(work) = 0.5$) has an entropy of 1 bit, akin to a coin toss.

In contrast, the entropy rate distribution has a mean of 1.4 bits/event and a standard deviation of 0.42 bits/event. The 1.1-bit difference between the mean entropy and the mean entropy rate reflects the additional information provided by the order in which events take place. Considering the order in which events are generated, an average user-sequence is associated with only slightly more uncertainty than a coin toss. In others words, on average, the next event can be predicted

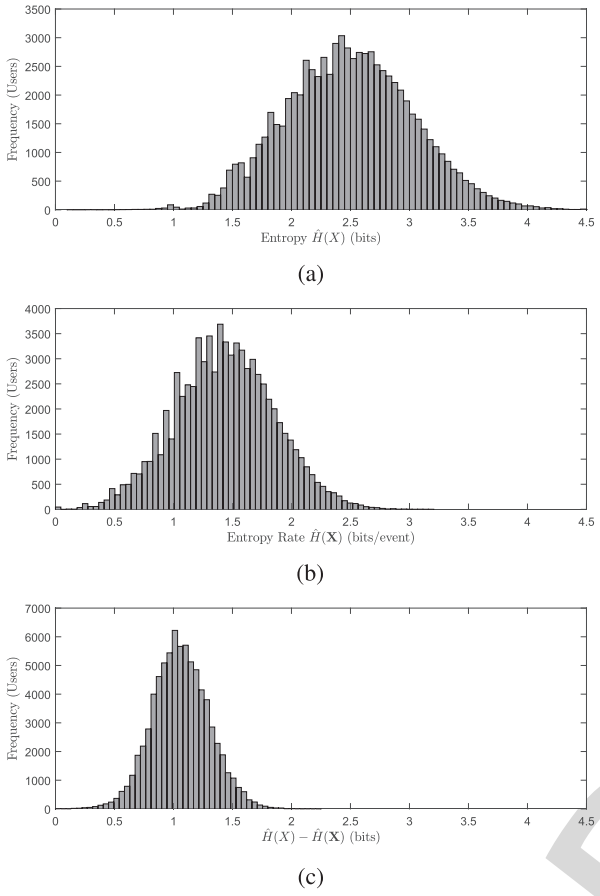


Fig. 4. Distribution of entropy measures across users. (a) Distribution of entropy. (b) Distribution of entropy rate. (c) Distribution of $(\hat{H}(X) - \hat{H}(\mathbf{X}))$.

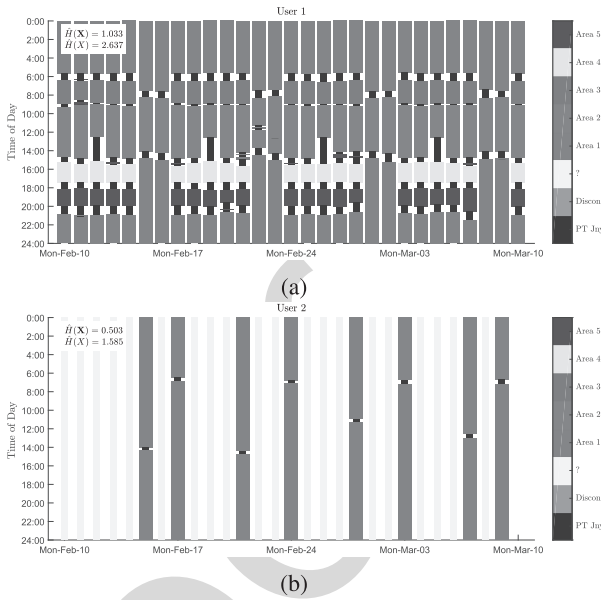


Fig. 5. Example location sequences of two users. (a) Activity sequence of user 1. (b) Activity sequence of user 2.

accurately almost 1 in 2 times when the order of events is considered, and only when 1 in 6 times when the order is not captured. Of course, the order of events does not provide the same amount of information for all individuals. As illustrated in Fig. 4c, for some users, the order of events provides almost no information, while for others it reduces the uncertainty by as much as 2 bits/event. Specifically, individuals who visit many locations frequently, but always in the same order will have relatively high entropy, but relatively low entropy rate. For reference, the individual used as an example earlier who traveled exclusively between home and work would have an entropy rate of 0 bit/event (as every new event is exactly determined by the previous one). In contrast, a coin toss has an entropy rate of 1 bit/event (same as its entropy) as there is no sequential dependency between events. Any individual whose travel pattern was exactly repeated over time would have an entropy rate of 0.

Fig. 5a illustrates the value of the entropy rate for a specific user who visited 5 locations almost exactly the same number of times, but consistently in the same order over the month-long observation period. The estimated entropy of the travel sequence of the user is 2.6, while its entropy rate is 1.0. The resulting 1.6 bit difference $\hat{H}(X) - \hat{H}(\mathbf{X})$ for this individual is at the high-end of the distribution shown in Fig. 4c. Additionally, while this individual's routine is not conventional, it is clearly regular as both the events and the order in which they are combined are repeated over time. This is reflected by the below average entropy rate of this sequence.

Fig. 5b illustrates another example of non-workday regularity captured by the entropy rate measure. On four separate occasions, the corresponding individual traveled from the primary location (i.e. red) to the secondary location (i.e. green) and made the reverse trip after one or two days without travel. In this specific example, the secondary location includes a terminal rail station, which the individual likely uses to leave London for the weekend. The user sometimes leaves on Friday, sometimes on Saturday and returns either on the following Sunday or Monday. Even though the pattern spans several days, and is not repeated periodically, its regularity is captured by the entropy rate of the sequence estimated to 0.503, below the sample average. This pattern would not accurately be captured by the standard periodicity measures reported in the literature, as it does not reoccur on the same days of the week from week to week.

Examination of other individuals shows that, in general, the entropy rate measures regularity accurately and can serve as a useful comparison metric. Fig. 6 compares two groups of 500 users whose sequence is longer than 40 events. The rows represent the sequence of an individual, while the columns correspond to different times of the 29 day period. The first group is randomly selected from all users with entropy rate below 1.0 bit. These regular users fall below the 10^{th} percentile of the entropy rate distribution for sequences longer than 40 events. The second group is randomly selected from all users with entropy rate above 2.1 bits. These irregular users fall above the 90^{th} percentile of the distribution. As expected, the sequences associated with regular users are characterized by the conventional working week structure. The irregular sequences contain much less repeated structure. It is important to note that while the dominant pattern in Fig. 6a

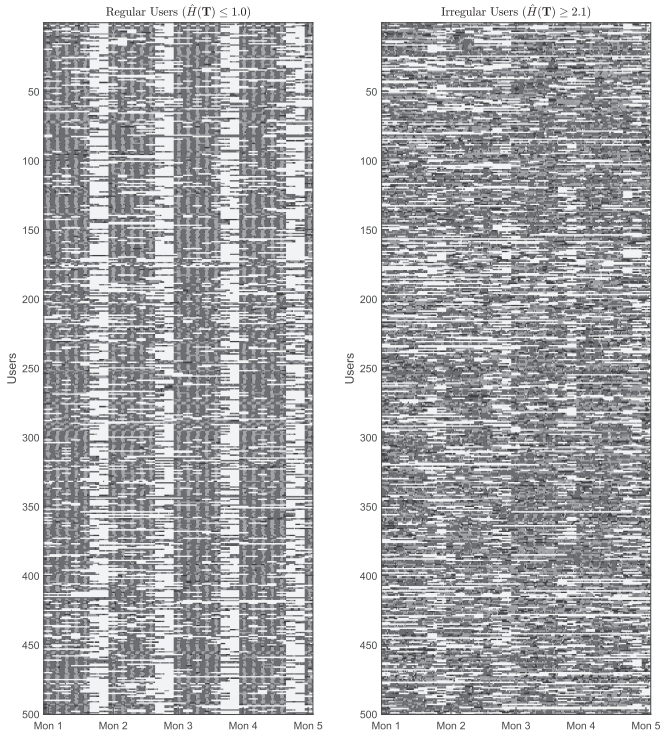


Fig. 6. Comparison of users in the lower and upper 10^{th} percentile of irregularity. (a) Regular Users. (b) Irregular Users.

is associated with the typical working week, many non-conventional patterns, such as those illustrated above, are also qualified as regular. This demonstrates that the entropy rate can be used as an indicator of regularity, capturing the extent of repetition in events and in the order in which they appear, while not making any assumption about how the repetition fits with conventional calendar cycles such as a day or a week.

VI. Discussion

Regularity is an important property of individual travel behavior, and the ability to measure it is valuable for advances in behavior modeling, mobility prediction, and customer segmentation. First, our study shows that much of the uncertainty in travel choices (such as location choice) can be accounted for by considering the order of these choices. The difference between entropy and entropy rate can be used as a measure of the potential value of incorporating sequential dependency in behavior modeling. Second, regularity is closely tied with the concept of predictability. As shown by [32], the entropy rate makes it possible to compute a fundamental limit of predictability of individual travel behavior, which can be used to evaluate predictive behavior models. Third, regularity is one of the metrics that shed light on a person's lifestyle, because it captures patterns in the overall organization of these behavior components. For example, Fig. 6 shows two groups of users, one with consistent itineraries (mostly commuters) and one with flexible schedules. This makes the measure of regularity a useful metric for user segmentation. Finally, the proposed methodology is highly flexible and can be adapted for different scenarios. Representing behavior as a sequence of events makes it computationally convenient to measure and analyze certain properties of human behavior that would be difficult to quantify otherwise. This is particularly fitting for new mobility data sources (e.g. smart card data) which typically provide long series of event-driven observations of individual behavior with no semantic annotations (e.g. travel purposes or activity types from survey data). It is also possible to adapt our regularity measure to study other types of human behavior using other sources of data (telecommunication behavior using mobile phone data, shopping behavior using credit card data, etc.)

VII. Conclusion

This paper provides an in-depth discussion of regularity of human travel behavior. We hypothesize that the order in which an individual engages in trips and activities constitutes an integral characteristic of human travel behavior and that this characteristic should be captured in the definition of regularity. We present a measure of regularity based on entropy rate which is sensitive to the frequency of travel events and to the order in which events are observed. To apply this measure, we also propose a framework to represent individual travel behavior as a sequence of travel events. The methodology is demonstrated using a large sample of transit smart card records from London, U.K. The Burrows-Wheeler transform is used for the estimation of the entropy rate. The results show that on average the next travel event can be predicted accurately almost 1 in 2 times when the order of events is considered, and only 1 in 6 times when the order is not considered. They also confirm the hypothesis that the order of travel events is important and captures a component of regularity not considered in the periodicity-based methods. Furthermore, the findings reveal that travel regularity may follow atypical patterns which are not captured by either periodicity-based methods or activity-based models. The regularity measure we propose is useful to reveal such patterns through data mining because it does not require assumptions about the periodic interval or the structure of regularity in travel behavior. It is also flexible and hence, can be adapted to study other types of human behavior using similar types of traces.

References

[1] D. Adjeroh, T. Bell, and A. Mukherjee, *The Burrows-Wheeler Transform: Data Compression, Suffix Arrays, and Pattern Matching*. Boston, MA, USA: Springer, 2008.

[2] M. Allahviranloo and W. Recker, "Daily activity pattern recognition by using support vector machines with multiple classes," *Transp. Res. B, Methodol.*, vol. 58, pp. 16–43, Dec. 2013.

[3] C. R. Bhat, S. Srinivasan, and K. W. Axhausen, "An analysis of multiple interepisode durations using a unifying multivariate hazard model," *Transp. Res. B, Methodol.*, vol. 39, no. 9, pp. 797–823, Nov. 2005.

[4] J. L. Bowman and M. E. Ben-Akiva, "Activity-based disaggregate travel demand model system with activity schedules," *Transp. Res. A, Policy Pract.*, vol. 35, no. 1, pp. 1–28, Jan. 2001.

[5] R. N. Buliung, M. J. Roorda, and T. K. Remmel, "Exploring spatial variety in patterns of activity-travel behaviour: Initial results from the Toronto Travel-Activity Panel Survey (TTAPS)," *Transportation*, vol. 35, no. 6, pp. 697–722, Aug. 2008.

[6] H. Cai, S. R. Kulkarni, and S. Verdu, "Universal entropy estimation via block sorting," *IEEE Trans. Inf. Theory*, vol. 50, no. 7, pp. 1551–1561, Jul. 2004.

[7] M. Chikaraishi, A. Fujiwara, J. Zhang, and K. Axhausen, "Exploring variation properties of departure time choice behavior by using multilevel analysis approach," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2134, pp. 10–20, Dec. 2009.
[8] N. Eagle and A. Pentland, "Reality mining: Sensing complex social systems," *Pers. Ubiquitous Comput.*, vol. 10, no. 4, pp. 255–268, Mar. 2006.
[9] Y. Gao, I. Kontoyiannis, and E. Bienenstock, "Estimating the entropy of binary time series: Methodology, some theory and a simulation study," *Entropy*, vol. 10, no. 2, pp. 71–99, Jun. 2008.
[10] J. B. Gordon, H. N. Koutsopoulos, N. H. Wilson, and J. P. Attanucci, "Automated inference of linked transit journeys in London using fare-transaction and vehicle location data," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2343, pp. 17–24, Sep. 2013.
[11] G. Goulet-Langlois, H. N. Koutsopoulos, and J. Zhao, "Inferring patterns in the multi-week activity sequences of public transport users," *Transp. Res. C, Emerg. Technol.*, vol. 64, pp. 1–6, Mar. 2016.
[12] K. M. N. Habib and E. J. Miller, "Modelling daily activity program generation considering within-day and day-to-day dynamics in activity-travel behaviour," *Transportation*, vol. 35, no. 4, pp. 467–484, May 2008.
[13] S. Hanson and J. O. Huff, "Assessing day-to-day variability in complex travel patterns," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 891, pp. 18–24, Dec. 1982.
[14] S. Hanson and O. J. Huff, "Systematic variability in repetitious travel," *Transportation*, vol. 15, nos. 1–2, pp. 111–135, Mar. 1988.
[15] J. O. Huff and S. Hanson, "Repetition and variability in urban travel," *Geograph. Anal.*, vol. 18, no. 2, pp. 97–114, Apr. 1986.
[16] C.-H. Joh, T. Arentze, F. Hofman, and H. Timmermans, "Activity pattern similarity: A multidimensional sequence alignment method," *Transp. Res. B, Methodol.*, vol. 36, no. 5, pp. 385–403, Jun. 2002.
[17] H. Kang and D. M. Scott, "Exploring day-to-day variability in time use for household members," *Transp. Res. A, Policy Pract.*, vol. 44, no. 8, pp. 609–619, Oct. 2010.
[18] L. M. Kieu, A. Bhaskar, and E. Chung, "Passenger segmentation using smart card data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1537–1548, Jun. 2015.
[19] M. Kim and D. Kotz, "Periodic properties of user mobility and access-point popularity," *Pers. Ubiquitous Comput.*, vol. 11, no. 6, pp. 465–479, 2006.
[20] R. Kitamura and T. V. D. Hoorn, "Regularity and irreversibility of weekly travel behavior," *Transportation*, vol. 14, no. 3, pp. 227–251, Sep. 1987.
[21] R. Kitamura, T. Yamamoto, Y. O. Susilo, and K. W. Axhausen, "How routine is a routine? An analysis of the day-to-day variability in prism vertex location," *Transp. Res. A, Policy Pract.*, vol. 40, no. 3, pp. 259–279, Mar. 2006.
[22] Z. Li, J. Wang, and J. Han, "Mining event periodicity from incomplete observations," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 444–452.
[23] C. Morency, M. Trpanier, and B. Agard, "Measuring transit use variability with smart-card data," *Transp. Policy*, vol. 14, no. 3, pp. 193–203, May 2007.
[24] T. Neutens, M. Delafontaine, D. M. Scott, and P. De Maeyer, "An analysis of day-to-day variations in individual spacetime accessibility," *J. Transp. Geography*, vol. 23, pp. 81–91, Jul. 2012.
[25] E. I. Pas, "Intrapersonal variability and model goodness-of-fit," *Transp. Res. A, Gen.*, vol. 21, no. 6, pp. 431–438, Nov. 1987.
[26] E. I. Pas and F. S. Koppelman, "An examination of the determinants of day-to-day variability in individuals' urban travel behavior," *Transportation*, vol. 13, no. 2, pp. 183–200, Jun. 1986.
[27] F. Primerano, M. A. P. Taylor, L. Pitaksringkarn, and P. Tisato, "Defining and understanding trip chaining behaviour," *Transportation*, vol. 35, no. 1, pp. 1–2, Jan. 2008.
[28] W. W. Recker, M. G. McNally, and G. S. Root, "Travel/activity analysis: Pattern recognition, classification and interpretation," *Transp. Res. A, Gen.*, vol. 19, no. 4, pp. 279–296, Jul. 1985.
[29] M. J. Roorda and T. Ruiz, "Long- and short-term dynamics in activity scheduling: A structural equations approach," *Transp. Res. A, Policy Pract.*, vol. 42, no. 3, pp. 545–562, Mar. 2008.
[30] J. Scheiner, "The gendered complexity of daily life: Effects of life-course events on changes in activity entropy and tour complexity over time," *Travel Behaviour Soc.*, vol. 1, no. 3, pp. 91–105, Sep. 2014.
[31] S. Schönfelder, "Urban rhythms: Modelling the rhythms of individual travel behaviour," Ph.D. dissertation, Inst. Transp. Planning Syst., ETH Zurich, Zürich, Switzerland, 2006.
[32] C. Song, Z. Qu, N. Blumm, and A.-L. Barabsi, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, Feb. 2010.
[33] F. M. J. Willems, "The context-tree weighting method: Extensions," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 792–798, Mar. 1998.
[34] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653–664, May 1995.
[35] M. J. Williams, R. M. Whitaker, and S. M. Allen, "Measuring individual regularity in human visiting patterns," in *Proc. Int. Conf. Soc. Comput. (SocialCom) Privacy, Secur., Risk Trust (PASSAT)*, Sep. 2012, pp. 117–122.

Gabriel Goulet-Langlois received the M.Sc. degree in transportation from MIT in 2015. As part of his research, he developed methods to analyze travel patterns and user behavior from large ticketing data sets. He is currently building on this experience in a practical context as a Data Scientist with Transport for London. He is dedicated to improving customer experience and public transport planning through better use of data and technology.

Haris N. Koutsopoulos is currently a Professor with the Department of Civil and Environmental Engineering, Northeastern University, Boston, and a Guest Professor with the KTH Royal Institute of Technology, Stockholm. His current research interests include on the use of data from opportunistic and dedicated sensors to improve planning, operations, monitoring, and control of urban transportation systems, including public transportation. He is the Founder of the iMobility lab, which uses Information and Communication Technologies to address urban mobility problems. The Laboratory received the IBM Smarter Planet Award in 2012.

Zhan Zhao received the B.Eng. degree from Tongji University in 2011 and the master's degree in applied science from the University of British Columbia in 2013. He is currently pursuing the Ph.D. degree in interdepartmental doctoral program in transportation with the Massachusetts Institute of Technology (MIT), and a Graduate Research Assistant with the MIT Transit Lab. His research interests include travel behavior modeling, public transportation systems, and urban computing.

Jinhua Zhao is the Edward H. and Joyce Linde Associate Professor with City and Transportation Planning, MIT. He brings behavioral science and transportation technology together to shape travel behavior, design mobility systems, and reform urban policies. He directs the MIT Urban Mobility Lab (mobility.mit.edu) and MIT Transit Lab.