A Statistical Framework for Measuring the Temporal Stability of Human Mobility Patterns

Zhihang Dong

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Arts

University of Washington

2020

Reading Committee:

Adrian Dobra

Kyle Crowder

Nathalie Williams

Program Authorized to Offer Degree:
Department of Sociology

University of Washington

Abstract

A Statistical Framework for Measuring the Temporal Stability of Human Mobility Patterns

Zhihang Dong

Co-Chairs of the Supervisory Committee:
Adrian Dobra

Kyle Crowder

Despite the growing popularity of human mobility studies that collect GPS location data, the problem of determining the minimum required length of GPS monitoring has not been addressed in the current statistical literature. In this paper we tackle this problem by laying out a theoretical framework for assessing the temporal stability of human mobility based on GPS location data. We define several measures of the temporal dynamics of human spatiotemporal trajectories based on the average velocity process, and on activity distributions in a spatial observation window. We demonstrate the use of our methods with data that comprise the GPS locations of 185 individuals over the course of 18 months. Our empirical results suggest that GPS monitoring should be performed over periods of time that are significantly longer than what has been previously suggested. Furthermore, we argue that GPS study designs should take into account demographic groups.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to Friedrich Nietzsche, who said:

> "*In order to be able thus to misjudge, and thus to grant left-handed veneration to our classics, people must have ceased to know them. This, generally speaking, is precisely what has happened. For, otherwise, one ought to know that there is only one way of honoring them, and that is to continue seeking with the same spirit and with the same courage, and not to weary of the search.*
>
> ---
>
> *Friedrich Nietzsche*
>
> *Untimely Meditations (1876)*

There are those moments in our lives that infiltrate us with illusions of regret. Learning thyself is, in many occasions, painful and cruel. Nevertheless, my spirits of seeking truth, love and science will never die.

DEDICATION

To Yijun

Love alters not with his brief hours and weeks

But bears it out even to the edge of doom.

To Mom and Dad

Section 1

MAIN ARTICLE

1.1 Introduction

Recent developments on global positioning systems (GPS) for wearable technology such as smartphones have drawn a great amount of interest from scientists studying the effects of environmental influences on different population groups [21, 17, 20, 12, 2, 30, 13, 27, 10, 23, 22]. A recent article [18] documents more than 100 studies from 20 disciplines that collect and analyze human time-stamped GPS location data. This type of data is key for learning about the places where people routinely spend their time during activities of daily living in order to establish their relationship with socio-economic outcomes, crime victimization, and physical and mental well-being. There have been extensive studies on the social stratification of mobility, such as health disparities of different neighborhoods, mental health, and substance abuse intervention [9, 24, 27], on the assessment of human spatial behavior and spatiotemporal contextual exposures [17, 20, 12], on the characterization of the relationship between geographic and contextual attributes of the environment (e.g., the built environment) and human energy balance (e.g., diet, weight, physical activity) [2, 30], on the study of segregation, environmental exposure, and accessibility in social science research [13], or on the understanding of the relationship between health-risk behavior in adolescents (e.g., substance abuse) and community disorder [27, 1, 28].

Notwithstanding a general consensus across disciplines about the tremendous potential of GPS location data for studying human mobility, very little is currently known about how long a GPS study should last. There is an inherent trade-off between collecting location data from people for longer vs. shorter periods of time. Recording more GPS locations yields more information about the locations where an individual spends their time, as well

as about the frequency, duration and timing of their visits to these places. However, an individual's participation in a GPS study comes with burdens that often become significant if accumulated over longer periods of time: the individual needs to carry the device recording the data (a GPS tracker) everywhere they go, and needs to make sure the device is properly charged at all times and functions properly. Until recently, most GPS study designs stipulated mandatory regular visits to project coordination sites to download data from the location trackers, to replace batteries, and replace the GPS tracking devices that were lost or were malfunctioning. While some of these issues have been addressed by using specialized apps on smartphones to collect GPS data and wirelessly transmit them into secure cloud databases, the costs of distributing smartphones to study participants, data plans, software development, and cloud computing are quite significant. In addition, there are important privacy considerations related to recording locations that might sensitive for study participants for long periods of time. For these reasons, it is desirable to design GPS studies that are as short as possible to reduce the costs of the projects and the burden of study participants, while in the same time still providing guarantees that sufficient location data have been collected to properly address the research aims.

Despite the constant growth in the number of human mobility studies that collect GPS location data in the last 20 years, the question about the determination of the amount of time of GPS monitoring has not been asked until recently [29]. In this paper, the authors argue that an effective GPS study should last until a minimum of 14 to 15 days of valid GPS data have been collected. While this finding is relevant for numerous research groups that, in the past, have designed GPS studies with a duration of 7 days (see [29] and the references therein), two weeks seems to severely underestimate the duration of other, more recent, GPS studies whose duration is significantly longer. For example, [4] and [19] represent studies that tracked adolescents in the San Francisco Bay area for one month. Another study [8] employs a more complex three site design that comprises five assessments that take place every six months over two years of follow-up for participants enrolled in Chicago, and three assessments that take place every six months over one year of follow-up for participants

enrolled in Jackson and New Orleans. During each assessment, participants wear a GPS tracker for two weeks. Thus this study [8] records GPS locations for a total of 10 weeks and 6 weeks, respectively, but splits the period of observation into several contiguous two week periods of GPS monitoring. These longer periods of observation time were suggested in [16] who found 17 weeks to be an adequate period of time to monitor human mobility based on geotagged social media data.

In this paper we lay out a theoretical framework for assessing the temporal stability of human mobility based on GPS location data. Such a framework is missing from the current statistical literature. Previous work [29, 16] on the assessment of the duration of GPS observation periods is based on empirical findings, and lack any theoretical underpinnings. We address this gap by introducing several measures of the temporal dynamics of spatiotemporal trajectories of individuals. We illustrate the use of these measures with publicly available data from a study that recorded GPS locations of 185 individuals that live in a city in Switzerland over the course of 18 months.

1.2 Methods

The spatiotemporal trajectory of an individual in a reference time frame $[t_{\min}, t_{\max}]$ and spatial observation window $\mathcal{W} \subset \mathbb{R}^2_+$ is a curve

$$X^{[t_{\min},t_{\max}]} \;=\; \{X(t) = (x_1(t), x_2(t)) : t \in [t_{\min}, t_{\max}]\} \subseteq \mathcal{W}, \tag{1.1}$$

where $x_1(\cdot)$ and $x_2(\cdot)$ represent the longitude and latitude coordinates, respectively, and $X(t)$ is the location visited by this individual at time t. We assume that this curve is smooth: $x_1(\cdot)$ and $x_2(\cdot)$ have continuous derivatives. The length of the curve in Eq. (1.1) is defined as [6]:

$$\mathsf{L}(X^{[t_{\min},t_{\max}]}) \;=\; \int_{t_{\min}}^{t_{\max}} \sqrt{\left(\frac{\mathrm{d}x_1(t)}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}x_2(t)}{\mathrm{d}t}\right)^2}\,\mathrm{d}t. \tag{1.2}$$

The complete trajectory $X^{[t_{\min},t_{\max}]}$ is never observed in the real world. Instead, n observation times $t_1, \ldots, t_n$ are sampled from a distribution on $[t_{\min}, t_{\max}]$ with density $\rho(\cdot)$, and

the corresponding locations $X(t_1), \ldots, X(t_n)$ on the curve $X^{[t_{\min}, t_{\max}]}$ are recorded. These locations are realizations of a random variable $X(T)$ where $T \sim q(\cdot)$. Ideally we would like T to follow a uniform distribution to have the same chance of recording a visited location anywhere in the reference time frame $[t_{\min}, t_{\max}]$. Due to technological limitations (e.g., GPS devices running out of power), heterogeneous built environments that prevent GPS devices to obtain a location (e.g., skyscapers in downtown areas or buildings without windows and WIFI coverage), or human behavioral factors (e.g., individuals turning off their GPS devices around certain locations sensitive to them) the distribution of T can be far from the uniform distribution.

We assume that GPS positional data from K study participants were recorded. We denote by $X_k^{[t_{\min}, t_{\max}]} = \{X_k(t) : t \in [t_{\min}, t_{\max}]\}$ the unobserved spatiotemporal trajectory of the k-th study participant. The observation times in the reference time frame $[t_{\min}, t_{\max}]$ can vary between study participants. The GPS data for the k-th study participant are the time stamped longitude and latitude locations:

$$\{X_{k,i} = X_k(t_{k,i}) : i = 1, \ldots, n_k\}, \tag{1.3}$$

where $n_k \geq 1$, the time $t_{k,i}$ was sampled from a distribution with density $\rho_k(\cdot)$ independently of the rest of the observation times, and $t_{\min} \leq t_{k,1} \leq \ldots t_{k,n_k} \leq t_{\max}$. Here $t_{k,i}$ represents the time when the i-th location of study participant k was recorded. Our framework allows for the possibility of having different reference time frames for various groups of study participants.

1.2.1 Measuring the temporal stability of human mobility patterns

One possible measure of the dynamics of the spatiotemporal trajectory $X^{[t_{\min}, t_{\max}]}$ is the average velocity $V(\tau)$ at time τ which is a function $V(\tau)$ of the length of the subcurve $X^{[t_{\min}, t_{\min}+\tau]}$ of $X^{[t_{\min}, t_{\max}]}$ from Eq. (1.1):

$$V(\tau) \;=\; \frac{1}{\tau}\mathsf{L}(X^{[t_{\min}, t_{\min}+\tau]}), \tag{1.4}$$

for $\tau \in (0, t_{\max} - t_{\min}]$ and $V(0) = 0$. A sample estimator of the average velocity for the k-th study participant is

$$\widehat{V}_k(\tau) \;=\; \frac{1}{\tau} \sum_{\{i: t_{k,i+1} \leq \tau\}} \|X_{k,i+1} - X_{k,i}\|. \tag{1.5}$$

where $\|X_{k,i+1} - X_{k,i}\|$ represents an estimate of the distance traveled between times $t_{k,i}$ and $t_{k,i+1}$. The average velocity is a straighforward way to quantify the dynamical characteristics of an individual, hence its stability can be used as an intuitive, easy to understand measure of temporal stability.

In what follows we will assume that study participants traveled in a straight line or "as the crow flies" between two consecutive observed GPS locations. This is the simplest assumption one can make which leads to an easy way of calculating Great Circle (WGS84 ellipsoid) distances between two spatial locations [3]. However, this assumption underestimates actual distances traveled, and consequently underestimates the average velocity. More accurate approximations of distances traveled can be defined based on the shortest distances between two locations on a road network that spans the spatial observation window $\mathcal{W}$. Calculating distances based on a road network is more complex than calculating straight line distances, and involves significant GIS work since the maximum speed of travel on different segments of road needs to be taken into account [7]. Nevertheless, as the span of time between two consecutive observed locations becomes shorter, the difference between the road network and straight line distances decrease.

More generally, consider a stochastic process $Z = \{Z(\tau) : \tau \in [0, t_{\max} - t_{\min}]\}$, where $Z(\tau)$ is a mapping $f(\cdot)$ of the subcurve $X^{[t_{\min}, t_{\min}+\tau]}$ into $\mathbb{R}_+$. The mapping $f(\cdot)$ is chosen such that $\lim_{\tau \to (t_{\max} - t_{\min})} Z(\tau) = Z(t_{\max} - t_{\min})$. We define the absolute percentage error (APE, henceforth) $\phi(Z; \tau)$ which measures the error made when approximating $Z(t_{\max} - t_{\min})$ with $Z(\tau)$ for $\tau \in [0, t_{\max} - t_{\min}]$:

$$\phi(Z; \tau) \;=\; \frac{|Z(\tau) - Z(t_{\max} - t_{\min})|}{Z(t_{\max} - t_{\min})}.$$

We quantify the temporal stability of the process Z by introducing a related process called

the last crossing time process $\mathsf{LCT}_Z = \{\mathsf{LCT}_Z(\gamma) : \gamma \geq 0\}$, where

$$\mathsf{LCT}_Z(\gamma) = \max\{\tau \in [0, t_{\max} - t_{\min}] : \phi(Z; \tau) > \gamma\}. \tag{1.6}$$

In Eq. (1.6), $\mathsf{LCT}_Z(\gamma)$ is the last time when the APE made when $Z(t_{\max} - t_{\min})$ is approximated with $Z(\tau)$ is above a threshold γ. The last crossing time is well defined since $\lim_{\tau \to (t_{\max} - t_{\min})} \phi(Z; \tau) = 0$.

Consider the process $Z_k = \{Z_k(\tau) : \tau \in [0, t_{\max} - t_{\min}]\}$ associated with the k-th study participant, $Z_k(\tau) = f\left(X_k^{[t_{\min}, t_{\min}+\tau]}\right)$, and let $\widehat{Z}_k$ be its sample estimator based on the positional data in Eq. (1.3). The average velocity in Eq. (1.4) and its sample estimator in Eq. (1.5) are examples of processes Z_k and $\widehat{Z}_k$. A sample estimator of the last crossing time $\mathsf{LCT}_{Z_k}(\gamma)$ is

$$\widehat{\mathsf{LCT}}_{Z_k}(\gamma) = \max_{i=1,\ldots,n_k} \left\{ t_{k,i} - t_{\min} : \phi(\widehat{Z}_k; t_{k,i} - t_{\min}) > \gamma \right\}. \tag{1.7}$$

We note that $\widehat{Z}_k(\tau)$ in the APE $\phi(\widehat{Z}_k; \tau)$ is determined based on the locations recorded for the k-th study participant before time τ: $\{X_{k,i} : t_{\min} \leq t_{k,i} \leq \tau\}$. As an illustration, Figure 1.1 shows estimates of the average velocity of an individual in the MDC data, together with the last crossing time estimate at $\gamma = 0.1$. The threshold γ is a precision threshold specified by the user. It reflects the analyst's requirement on how stable the estimator (1.7) has to be. By decreasing γ, the stability of this estimator increases. Smaller values of γ correspond with a more stable estimator. For example, the choice $\gamma = 0.1$ expresses a 10% relative error to a long term study.

The last crossing time of the APE associated with a process that is a function of the spatiotemporal trajectory of a study participant represents a measure of this individual's mobility. Study participants that have more irregular mobility patterns (e.g., regular travel to locations at various distances from the individual's residence that change after a few days or weeks) are expected to have larger last crossing times compared to study participants that travel to the same locations each week. An example individual with a very regular mobility pattern that travels every day from his home to his office and back by following the same route, and goes nowhere else will record an APE equal to 0 after one day which leads to last

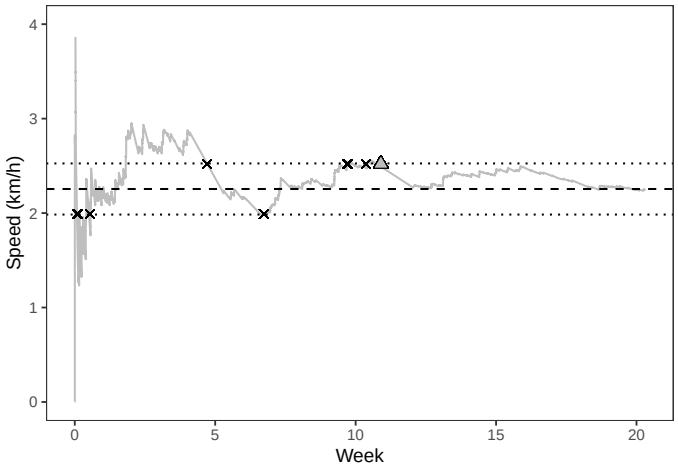

Figure 1.1: Estimate of the average velocity (gray curve) of an individual in the MDC data over $t_{\max} = 21$ weeks. The dashed line indicates the value of $\widehat{V}(t_{\max})$, and the two dotted lines represent the lower bound $(1-\gamma)\widehat{V}(t_{\max})$ and the upper bound $(1+\gamma)\widehat{V}(t_{\max})$ for $\gamma = 0.1$. These bounds correspond with times τ for which the APE $\phi(V;\tau) \leq \gamma$. The crosses denote the times τ for which $\phi(V;\tau) = \gamma$. The last crossing time for $\gamma = 0.1$ is marked with a triangle, and occurs at the end of week 10.

crossing times of less than one day in Eq. (1.7).

Previous work [29] on the temporal stability of spatiotemporal trajectories has used the mean absolute percentage error (MAPE) which is the average of the APE across study participants:

$$\overline{\phi}_K(\tau) = \frac{1}{K}\sum_{k=1}^{K} \phi(\widehat{Z}_k;\tau). \tag{1.8}$$

We define two measures of the overall temporal stability of the spatiotemporal trajectories of multiple study participants. The first overall measure is the last crossing time process $\mathsf{LCT}_{\overline{\phi}_K} = \{\mathsf{LCT}_{\overline{\phi}_K}(\gamma) : \gamma \geq 0\}$ of the MAPE process $\overline{\phi}_K = \{\overline{\phi}_K(\tau) : \tau \in [0, t_{\max} - t_{\min}]\}$. We refer to this measure as $\mathsf{LCT} - \mathsf{MAPE}(Z)$. The second overall measure is defined as the average of the last crossing times of the APE of $\widehat{Z}_k$ for $k = 1, \ldots, K$, i.e. $\overline{\mathsf{LCT}}_K = \{\overline{\mathsf{LCT}}_K(\gamma) :$

$\gamma \geq 0\}$ where

$$\overline{\mathsf{LCT}}_K(\gamma) \;=\; \frac{1}{K}\sum_{k=1}^{K}\mathsf{LCT}_{Z_k}(\gamma).$$

We denote this second measure by $\overline{\mathsf{LCT}} - \mathsf{APE}(Z)$. These two measures are the same only if they are calculated for a single study participant ($K = 1$). They are useful for comparing the temporal regularity of mobility patterns of groups of study participants (e.g., younger vs. older individuals, men vs. women, high SES vs. low SES).

1.2.2 The activity distribution of human mobility patterns

The average velocity associated with the spatiotemporal trajectory of an individual does not provide any information about the spatial configuration of locations visited. Consider two example individuals that drive without stopping with the same speed for a long period of time. The first example individual drives back and forth between two places A_1 and A_2. The second example individual drives in a cycle from a place A_1 to another place A_2, then to places A_3 and A_4, then back to place A_1. Since the spatiotemporal trajectory of the second individual involves two additional places, more sample locations will be needed to understand the mobility pattern of the second individual compared to the mobility pattern of the first individual. However, the mobility patterns of these two example individuals will be indistinguishable based on the last crossing time process associated with their average velocity processes. We address this issue by introducing a distribution of the locations visited by an individual.

We assume that the observation window $\mathcal{W}$ is partitioned into a set of grid cells $\mathcal{G} = \{G_1, \ldots, G_N\}$. Each location $X(t)$ on the curve $X^{[t_{\min}, t_{\max}]}$ representing the spatiotemporal trajectory of an individual is mapped into a grid cell $G(t) \in \mathcal{G}$. The observed locations for this individual mapped into $\mathcal{G}$ are the sequence of grid cells $g_1 = G(t_1), \ldots, g_n = G(t_n)$ that are realizations of a random variable $G(T)$ where T is a random variable on $[t_{\min}, t_{\max}]$ with a distribution with density $\rho(\cdot)$.

We define the activity distribution $\pi = (\pi_1, \ldots, \pi_N)$ over the grid cells $\mathcal{G}$. Here π_j represents the proportion of time in $[t_{\min}, t_{\max}]$ spent by an individual in cell $G_j \in \mathcal{G}$. We assume that T follows a uniform distribution on $[t_{\min}, t_{\max}]$, and define:

$$\pi_j = \mathsf{P}(G(T) = G_j), \quad \text{for } j = 1, \ldots, N. \tag{1.9}$$

The activity distributions associated with the two example individuals we introduced earlier can differentiate between their mobility patterns if the grid cells in which A_3 and A_4 do not coincide with the grid cells of A_1 and A_2, and will show that the first example individual did not spend any time in the grid cells associated with A_3 and A_4. To employ activity distributions we need to have a method for recovering them from the available data.

The simplest estimator $\widehat{\pi} = (\widehat{\pi}_1, \ldots, \widehat{\pi}_N)$ of the activity distribution π is based on the relative frequency of visitation of the grid cells $\mathcal{G}$:

$$\widehat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(g_i = G_j), \quad \text{for } j = 1, \ldots, N.$$

However, this estimator of π is reasonable only if T follows a uniform distribution as in Eq. (1.9). When T follows an arbitrary distribution with density $\rho(\cdot)$, a better approach is to use a weighted average estimator $\widetilde{\pi} = (\widetilde{\pi}_1, \ldots, \widetilde{\pi}_N)$ where:

$$\widetilde{\pi}_j = \frac{\sum_{i=1}^{n} \rho^{-1}(t_i) \mathbf{1}(g_i = G_j)}{\sum_{\ell=1}^{n} \rho^{-1}(t_\ell)}, \quad \text{for } j = 1, \ldots, N. \tag{1.10}$$

Although this estimator can be shown to be statistically consistent, it requires knowledge of the density $\rho(\cdot)$. There are many methods for estimating $\rho(\cdot)$ from the data such as histograms or kernel density estimators [26]. We suggest using an estimation method that assumes that the distribution of T is approximated by a piecewise uniform distribution. We take $t_0 = t_{\min}$ and $t_{n+1} = t_{\max}$. If T is approximately uniform in $[t_{i-1}, t_{i+1}]$ for $i = 1, \ldots, n$, then $\rho^{-1}(t_i) \approx t_{i+1} - t_{i-1}$. This is a reasonable assumption if the times when locations are collected are roughly equally spaced in time (e.g., a location is collected every 10 minutes) since the mean of t_i is $(t_{i+1} - t_{i-1})/2$. Thus an estimator of $\rho(\cdot)$ is

$$\widehat{\rho}(t_i) = \frac{\omega(t_i)}{\sum_{\ell=1}^{n} \omega(t_\ell)}, \ \omega(t_i) = \frac{1}{t_{i+1} - t_{i-1}}, \quad \text{for } i = 1, \ldots, n.$$

The weighted average estimator from Eq. (1.10) becomes

$$\begin{aligned}\widehat{\pi}_{o,j} &= \frac{\sum_{i=1}^{n}\omega^{-1}(t_i)\mathbf{1}(g_i = G_j)}{\sum_{\ell=1}^{n}\omega^{-1}(t_\ell)} \\ &= \frac{\sum_{i=1}^{n}(t_{i+1}-t_{i-1})\mathbf{1}(g_i = G_j)}{t_{\max}-t_{min}+t_n-t_1}, \quad \text{for } j = 1,\ldots,N.\end{aligned} \tag{1.11}$$

We call $\widehat{\pi}_o = (\widehat{\pi}_{o,1},\ldots,\widehat{\pi}_{o,N})$ the ordinary proportional time estimator of the activity distribution π. This estimator relies on the assumption that the length of the time intervals in which an individual transitions between two grid cells is added to the time spent in both the grid cell they leave from, and the grid cell they arrive in. More specifically, assume that the consecutive observation times t_i and t_{i+1} are such that $g_i \neq g_{i+1}$. Then $\widehat{\pi}_o$ allocates $(t_{i+1}-t_i)$ to the total time spent in both g_i and g_{i+1}.

We introduce a second estimator $\widehat{\pi}_c = (\widehat{\pi}_{c,1},\ldots,\widehat{\pi}_{c,N})$ of the activity distribution π:

$$\widehat{\pi}_{c,j} = \frac{\sum_{i=2}^{n}(t_i - t_{i-1})\mathbf{1}(g_i = g_{i-1} = G_j)}{\sum_{i=2}^{n}(t_i - t_{i-1})\mathbf{1}(g_i = g_{i-1})}, \quad \text{for } j = 1,\ldots,N. \tag{1.12}$$

We call $\widehat{\pi}_c$ the conservative proportional time estimator. This estimator is more conservative than the ordinary proportional time estimator $\widehat{\pi}_o$ from Eq. (1.11) in the sense that any time interval defined by consecutive observation times t_i and t_{i+1} such that $g_i \neq g_{i+1}$ is ignored. That is, the time spent in a grid cell is calculated only based on time intervals in which an individual is known to have remained in that cell.

We show two important properties of the ordinary and the conservative proportional time estimators. First, we prove that both estimators are asymptotically equivalent. Second, we prove that both estimators are statistically consistent, that is, they will eventually recover the true activity distribution π if sufficient location data are available. These properties rely on the assumptions (S1), (S2) and (S3) below:

(S1) The length of the time intervals between consecutive observation times $\max_{i=1,\ldots,n-1}|t_{i+1}-t_i| \to 0$ as the sampling rate $n \to \infty$.

(S2) The sampling period is such that $t_1 \to t_{\min}$ and $t_n \to t_{\max}$ when $n \to \infty$.

(S3) The number of transitions between grid cells is finite, i.e., there exists $M < \infty$ such that $\sum_{t \in [t_{\min}, t_{\max}]} \mathbf{1}(G(t_+) \neq G(t_-)) \leq M$, where $G(t_-)$ and $G(t_+)$ are the left and right limits of $G(\cdot)$ at t.

Assumptions (S1) and (S2) describe the meaning of asymptotics in our context. They imply that the observation times $t_1, \ldots, t_n$ will eventually be dense in the reference time frame, i.e., there will not exist a fixed region of $[t_{\min}, t_{\max}]$ without any observation times when $n \to \infty$. Assumption (S3) requires that the spatiotemporal trajectory $X^{[t_{\min}, t_{\max}]}$ is sufficiently smooth such that it will not jump between grid cells infinitely often.

Theorem 1.2.1 (Asymptotic Equivalence Rule with Large Sampling Rate) *Under assumptions (S1), (S2) and (S3), the ordinary proportional time estimator $\widehat{\pi}_o$ from Eq.* (1.11) *and the conservative proportional time estimator $\widehat{\pi}_c$ from Eq.* (1.12) *are asymptotically the same.*

The proof of this result is given in Appendix A.1.1. We can also show that the same assumptions imply that the two estimators are statistically consistent.

Theorem 1.2.2 (Convergence Rule with Large Sampling Rate) *Under assumptions (S1), (S2) and (S3), the ordinary proportional time estimator $\widehat{\pi}_o$ from Eq.* (1.11) *and the conservative proportional time estimator $\widehat{\pi}_c$ from Eq.* (1.12) *converge to the true activity distribution π from Eq.* (1.9)*.*

The proof of this result is given in Appendix A.1.2.

1.2.3 Measuring the temporal stability of human activity distributions

We are interested in determining the temporal stability of the activity distribution of an individual. We assume that the reference time frame $[t_{\min}, t_{\max}]$ is divided into $D_{\max}$ time periods of equal lengths (e.g., days or weeks). We denote by $\pi^{(d)}$ the activity distribution from Eq. (1.12) associated with time period D, $D = 1, \ldots, D_{\max}$. Then $\pi^{(D)}$ can be viewed as an

N-dimensional random vector whose distribution reflects the variability from time period to time period of the individual's mobility patterns. With this understanding, we are interested in determining the expectation $\bar{\pi} = \mathbb{E}(\pi^{(D)})$. We call $\bar{\pi}$ the time period activity distribution (e.g., daily or weekly activity distribution). The j-th component of $\bar{\pi}$ is interpreted as the average proportion of time spent by the individual in grid cell G_j in a given time period (a day or a week).

A simple estimator of $\bar{\pi}$ is

$$\widehat{\bar{\pi}}(D) = \frac{1}{D}\sum_{d=1}^{D} \widehat{\pi}^{(d)}, \quad \text{for } D = 1, \ldots, D_{\max}, \tag{1.13}$$

where $\widehat{\pi}^{(d)}$ is the ordinary proportional time estimator $\widehat{\pi}_o$ from Eq. (1.11) or the conservative proportional time estimator $\widehat{\pi}_c$ from Eq. (1.12).

Because $\widehat{\bar{\pi}}(D)$ is a consistent estimator of $\bar{\pi}$, the error we make when approximating $\bar{\pi}$ with $\widehat{\bar{\pi}}(D)$ decreases as we observe the spatiotemporal trajectory of the individual for a larger number of time periods $D_{\max}$. We define the last crossing time of the sequence of estimators $\{\widehat{\bar{\pi}}(D) : D = 1, \ldots, D_{\max}\}$ as follows:

$$\widehat{\mathsf{LCT}}_{\mathsf{dist}}(\gamma) = \max_{D=1,\ldots,D_{\max}} \left\{ D : \|\widehat{\bar{\pi}}(D) - \widehat{\bar{\pi}}(D_{\max})\|_1 > \gamma \right\}, \tag{1.14}$$

where $\|v\|_1$ is the usual L_1 norm for a vector v, i.e., $\|v\|_1 = \sum_i |v_i|$. Note in Eq. (1.14) we used the fact that $\|\widehat{\bar{\pi}}(D)\|_1 = 1$ for any D.

The last crossing time in Eq. (1.14) is a measure of the temporal stability of the entire time period activity distribution $\bar{\pi}$. Individuals that spend approximately the same amount of time in the same places in every time period need to be observed for a smaller number of time periods to calculate estimator $\widehat{\bar{\pi}}(D)$ with the same APE compared to individuals with heterogeneous mobility patterns that spend different amounts of times at locations that change substantially across time periods. Therefore $\widehat{\mathsf{LCT}}_{\mathsf{dist}}(\gamma)$ will be smaller for individuals whose time period to time period mobility changes less, and larger for individuals with irregular mobility patterns.

The disadvantage of using the last crossing time in Eq. (1.14) as a measure of temporal stability comes from the fact that it gives the same weight to the error made when estimating the proportion of time spent in grid cells in which an individual spends a lot of their time, and to the grid cells in which the individual rarely visits. The number of grid cells with a large proportion of time spent in them is likely significantly smaller than the total number of grid cells N because most people tend to spend time at their residence, to their work place and perhaps in a few other select locations. For this reason, the error made when estimating the proportion of time spent in grid cells with sparse presence could dominate the overall APE of $\widehat{\bar{\pi}}(D)$, and lead to larger values of $\widehat{\mathsf{LCT}}_{\mathsf{dist}}(\gamma)$. To remedy this issue, we define a new measure of temporal stability that focuses on the grid cells in which an individual spends larger proportions of time.

We define the ranking time period activity distribution $\bar{r} = (\bar{r}_1, \cdots, \bar{r}_N)$ associated with $\bar{\pi}$ by replacing each component of $\bar{\pi}$ with the sum of those components of $\bar{\pi}$ that are no larger than that component, as follows [5]:

$$\bar{r}_j = \sum_{l=1}^{N} \bar{\pi}_l \mathbf{1}(\bar{\pi}_l \leq \bar{\pi}_j), \quad \text{for } j = 1, \ldots, N. \tag{1.15}$$

The α-level set ($\alpha \in [0, 1]$) of $\bar{r}$ is defined to consist of all the grid cells whose corresponding components in $\bar{r}$ exceed α:

$$L_\alpha = \{G_j : \bar{r}_j \geq \alpha\}. \tag{1.16}$$

It turns out that the α-level set covers grid cells whose total sum of components of $\bar{\pi}$ is larger than $1 - \alpha$:

$$\sum_{G_j \in L_\alpha} \bar{\pi}_j \geq 1 - \alpha.$$

Levels sets have an easy to understand interpretation: for a given level α, say $\alpha = 0.7$, all the grid cells with a ranking time period activity distribution above 0.7 will jointly cover at least $(1 - 0.7) \cdot 100 = 30\%$ of the time in the time period. Values of α closer to 1 lead to level sets L_α with a smaller coverage that comprise only the grid cells in which the individual

spends the largest amounts of time. Values of α close to 0 lead to level sets L_α with a larger coverage that comprise the majority of grid cells the individual spent time in.

Let $\widehat{\bar{r}}(D)$ be the ranking distribution of the estimator $\widehat{\bar{\pi}}(D)$ of $\bar{\pi}$ in Eq. (1.13), and $L_\alpha(D)$ be the α-level set associated with $\widehat{\bar{r}}(D)$ as in Eq. (1.16). Given a level $\alpha \in [0,1]$ and a stability threshold $\gamma > 0$, we define the last crossing time of the sequence of level sets $\{L_\alpha(D) : D = 1, \ldots, D_{\max}\}$ as follows:

$$\widehat{\mathsf{LCT}}_{\mathsf{level},\alpha}(\gamma) = \max_{D=1,\ldots,D_{\max}} \left\{ D : \frac{\|L_\alpha(D) \triangle L_\alpha(D_{\max})\|}{\|L_\alpha(D_{\max})\|} > \gamma \right\}, \tag{1.17}$$

where $\triangle$ denotes the symmetric difference of two sets, and $\|\cdot\|$ denotes the number of elements in a set.

The LCT of the level sets from Eq. (1.17) is a measure of temporal stability of the time period activity distribution $\bar{\pi}$ that takes into account only the error made when estimating the time spent in the grid cells in which an individual spent most of their time. For the same value of γ, $\widehat{\mathsf{LCT}}_{\mathsf{level},\alpha}(\gamma)$ is decreasing as the level α is increasing.

1.3 Application

The data we analyze comes from Nokia's Mobile Data Challenge (MDC) [11, 14, 15]. This was a mobile computing research initiative focusing on generating a deeper scientific understanding of social and behavioral patterns related to mobile technologies. The study took place in Switzerland, and collected various types of longitudinal information including time stamped GPS data from the cell phones of 185 study participants over the course of 18 months. Demographic data such as age and sex is also available. There are approximately 57.5 million GPS location records. The average length of observation for study participants was about 55 weeks. These data are publicly available upon request from the Idiap Research Institute.

Most activities of daily living of the study participants took place in a rectangular area that we partitioned into 4000^2 square grid cells with sides of length 28 meters. The locations that do not belong to this spatial observation window were dropped. These locations typically

correspond with longer trips took by study participants away from their places of residency. Figure 1.2 displays summaries of the GPS locations that fall in our chosen spatial observation window.

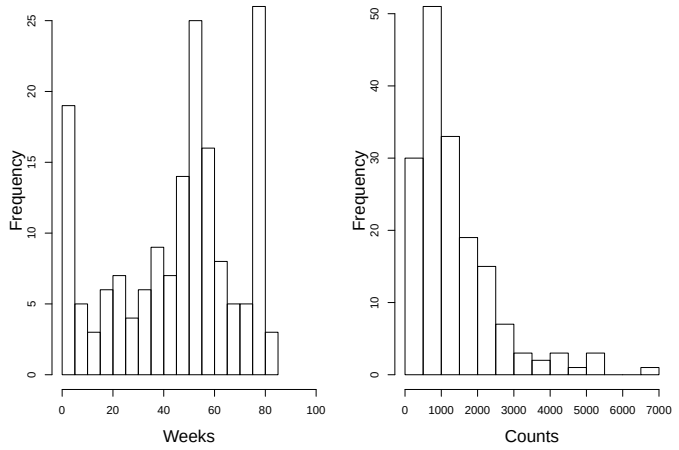

Figure 1.2: Summary information of the GPS location data. Left panel: histogram of the total length of observation for each study participant expressed in weeks. Right panel: histogram of the average number of GPS locations per week for each study participant.

For each study participant, we calculated three measures of temporal stability of their mobility patterns: the last crossing time of the average velocity (LCT-velocity) as defined in Eq. (1.5) and Eq. (1.7), the last crossing time of the activity distribution (LCT-distribution) as defined in Eq. (1.14), and the last crossing time of the level sets of the weekly activity distribution as defined in Eq. (1.17). In the calculation of LCT-distribution and LCT-level set, we used the ordinary proportional time estimator defined in Eq. (1.11). We chose to use the ordinary proportional time estimator over the conservative proportional time estimator because the conservative proportional time estimator disregards the pairs of consecutive time points that are located in different grid cells. The conservative proportional time estimator would most likely yield a smaller sample size compared to the ordinary proportional time estimator. We used $\alpha = 0.2$ in the determination of level sets, and $\gamma = 0.2$ as the stability threshold for all three measures. The results are summarized in Table 1.1.

Table 1.1: Means, medians and sample standard deviations of three measures of temporal stability of mobility patterns. The unit of time is weeks.

Mobility Measure	Mean	Median	St. Dev.
LCT-velocity	30.04	26	17.29
LCT-distribution	37.18	37	16.06
LCT-level set ($\alpha = 0.2$)	17.69	17	9.50

About 30 weeks of observation is needed until the mobility patterns stabilize according to the LCT-velocity measure. A longer period of time, 37 weeks, is needed until the weekly activity distribution stabilizes. The increased length of the period of observation for this measure is not surprising since it is based on an estimated of the full weekly activity distribution in $N = 4000^2$ grid cells. About half of this observation time (18 weeks) is needed to obtain estimates of the 0.2-level set of the weekly activity distribution which comprise the grid cells in which the study participants spend 80% of their weekly time.

We exemplify how the α-level set L_α from Eq. (1.16) and its corresponding LCT-level set $\widehat{\mathsf{LCT}}_{\mathsf{level},\alpha}(0.2)$ from Eq. (1.17) change for different values of $\alpha \in [0, 1]$. To this end, we define an adjacency graph G_{grid} whose vertices are the $N = 4000^2$ grid cells in the spatial observation window. Two grid cells are connected by an edge in G_{grid} if they share an edge or a corner in their arrangement in the spatial observation window [25, 3]. We denote by $G_{\mathsf{grid}}(L_\alpha)$ the subgraph of G_{grid} defined by the grid cells in L_α. We chose a study participant, and determined the level set L_α, the last crossing time $\widehat{\mathsf{LCT}}_{\mathsf{level},\alpha}(0.2)$ and the number of connected components of $G_{\mathsf{grid}}(L_\alpha)$ for $\alpha \in \{0.1, 0.2, \ldots, 1\}$ – see Figure 1.3. For smaller values of α, L_α contains grid cells in which the study participant spend the largest proportion of time. When $\alpha \in \{0.1, 0.2, 0.3, 0.4\}$, $G_{\mathsf{grid}}(L_\alpha)$ has one connected component which implies that the grid cells that belong to L_α are spatially adjacent, and define a single area in which the study participant spends larger amounts of time. The corresponding values of $\widehat{\mathsf{LCT}}_{\mathsf{level},\alpha}(\gamma)$ are less than 20 weeks which represents the length of observation time needed for reliably

detecting this spatial area. For $\alpha \in \{0.5, 0.6\}$, $G_{\mathsf{grid}}(L_\alpha)$ has two connected components, and for $\alpha \in \{0.7, 0.8\}$, $G_{\mathsf{grid}}(L_\alpha)$, $G_{\mathsf{grid}}(L_\alpha)$ has three connected components. Thus this study participant spends their time in grid cells that define two or three spatially contiguous areas. Since these areas include grid cells in which the study participant spends smaller proportions of their weekly time, the length of the observation time needed to identify these areas doubles to about 40 weeks. For $\alpha = 1$, $G_{\mathsf{grid}}(L_\alpha)$ has 72 connected components because L_α includes grid cells in which the study participant spends very little time. Figure 1.3 shows that approximately 70 weeks of observation time are needed to detect these grid cells. The same type of plots constructed for other study participants show similar relationships between α, L_α, and $\widehat{\mathsf{LCT}}_{\mathsf{level},\alpha}(0.2)$.

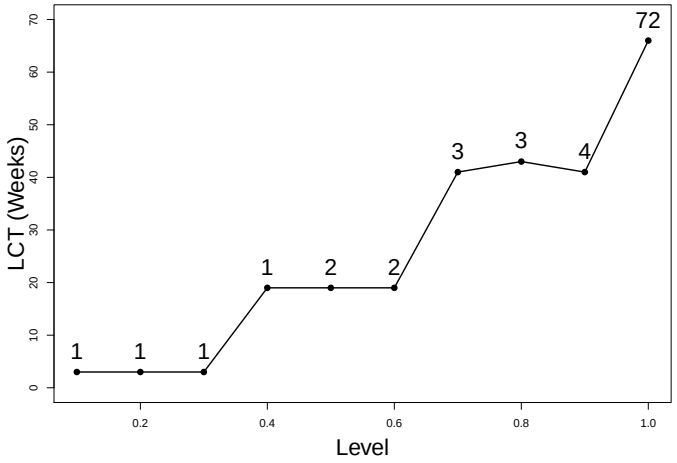

Figure 1.3: Values of the LCT-level sets $\widehat{\mathsf{LCT}}_{\mathsf{level},\alpha}(0.2)$ for $\alpha \in \{0.1, 0.2, \ldots, 1\}$ for an MDC study participant. The unit of time is weeks. The number of connected components of $G_{\mathsf{grid}}(L_\alpha)$ defined by the α-level sets L_α are shown above the curve.

Next we want to determine whether the temporal stability of activity distributions varies by the demographic characteristics of the population. We group the study participants by sex (male, female) and age group (young age 15–34 years old, middle age 35–54 years old, and old age $\geq$55 years old). For each of these five demographic groups, we calculated the average of the last crossing times of the activity distribution $\widehat{\mathsf{LCT}}_{\mathsf{level},\alpha}(0.2)$ for every

$\alpha \in \{0.1, 0.2, \ldots, 1\}$. The resulting curves are presented in Figure 1.4. The last crossing times at all levels are similar for men and women (see the top left panel). As such, there do not seem to be any sex-based differences in the temporal stability of men and women who live in Switzerland. However, since Switzerland is known to be a country with very high equality between the two sexes, this finding might not extend to other countries with profound sex inequality.

In the top right and bottom panels of Figure 1.4, we find evidence that the average last crossing times decrease with age especially for levels below 0.5. This means that mobility patterns are more regular, and consequently are more temporally stable for older study participants compared to younger study participants. The average last crossing times are larger and become very similar across demographic groups for levels above 0.5 compared to smaller levels below 0.5. Thus study participants that belong to any of the five demographic groups tend to visit locations they do not typically visit. Longer observation periods are needed to successfully determine these locations. Nevertheless, in order to identify the areas in which study participants spend most of their time, Figure 1.4 suggests that 10 weeks of observation of GPS locations should suffice for individuals older than 55. Middle age individuals require about 15 weeks of observation time, while young individuals require about 20 weeks.

1.4 Discussion

The contribution we made in this paper is two fold. On the theoretical side, we proposed the use of last crossing time processes associated with spatiotemporal trajectories of individuals to assess the temporal stability of their mobility patterns. We defined several measures of the temporal dynamics of spatiotemporal trajectories based on the average velocity process, and on human activity distributions in a spatial observation window. We defined the ordinary and the conservative proportional time estimators of human activity distributions, and proved that they are consistent and asymptotically equivalent. We introduced the time period and the ranking time period activity distributions that capture the change in human activity

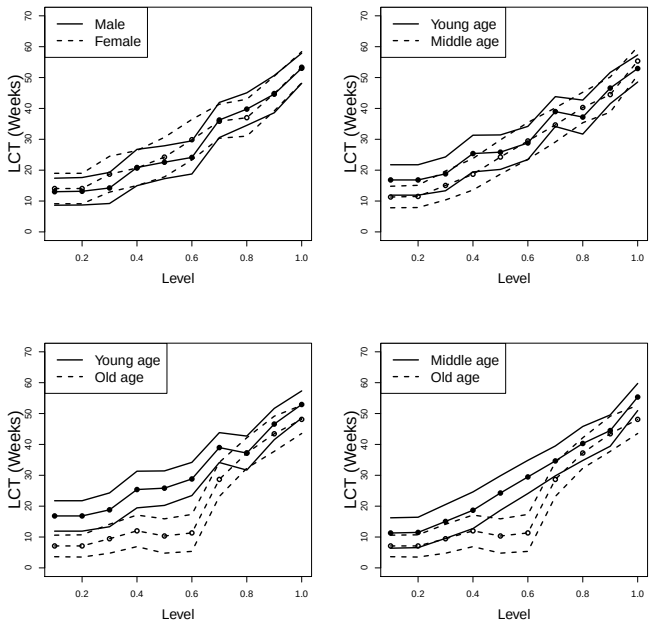

Figure 1.4: Mean values and 90% confidence intervals of the LCT-level sets $\widehat{\mathsf{LCT}}_{\mathsf{level},\alpha}(0.2)$ for $\alpha \in \{0.1, 0.2, \ldots, 1\}$ calculated for five demograhic groups: sex (male, female), and age (young, middle, old).

distributions across time periods. We presented related estimators based on GPS location data.

On the empirical side, we analyzed GPS location data collected over a period of 18 months. The previous empirical study [29] that focused on assessing the duration of GPS studies is based on data collected over 30 days. By using our new statistical methods and GPS data collected over a much longer period of time, we determined that GPS monitoring needs to be done for at least 15 weeks which represents a minimum study duration about 7 times longer than the 14 days minimum duration recommended in [29]. We also put forward the idea that the duration of GPS studies should be assessed by demographic groups. We determined that younger population groups should be monitored for longer periods of time compared to middle age population groups because of their more irregular patterns of mobility. On the other hand, shorter monitoring periods might be needed for older population groups that exhibit mobility patterns that are temporally more stable. We also suggest using our methods to assess the need for different time spans of GPS monitoring for men and women in countries with a known history of inequality between the two sexes. To the best of our knowledge, differential periods of GPS data collection based on demographic groups has not been discussed before. Our work suggests that GPS study designs should take demographic groups into account.

Funding

The work of Z.D. and A.D. was partially supported by the National Science Foundation Grant DMS/MPS-1737746 to University of Washington. Y.C. received partial support from the National Science Foundation Grant DMS-1810960 and National Institutes of Health Grant U01-AG016976. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgment

Portions of the research in this paper used the MDC Database made available by Idiap Research Institute, Switzerland and owned by Nokia.

BIBLIOGRAPHY

[1] L.A. Basta, T.S. Richmond, and D.J. Wiebe, *Neighborhoods, daily activities, and measuring health risks experienced in urban environments*, Social Science & Medicine 71 (2010), pp. 1943–1950.

[2] D. Berrigan, J.A. Hipp, P.M. Hurvitz, P. James, M.M. Jankowska, J. Kerr, F. Laden, T. Leonard, R.A. McKinnon, T.M. Powell-Wiley, E. Tarlov, S.N. Zenk, and The TREC Spatial and Contextual, Measures and Modeling Work Group, *Geospatial and contextual approaches to energy balance and health*, Annals of GIS 21 (2015), pp. 157–168.

[3] R.S. Bivand, E. Pebesma, and V. Gómez-Rubio, *Applied Spatial Data Analysis with R*, Springer, New York, 2013.

[4] H. Byrnes, B.A. Miller, C.N. Morrison, D.J. Wiebe, M. Woychik, and S.E. Wiehe, *Association of environmental indicators with teen alcohol use and problem behavior: Teens' observations vs. objectively-measured indicators*, Health & Place 43 (2017), pp. 151–157.

[5] Y.C. Chen, *Generalized cluster trees and singular measures*, Annals of Statistics 47 (2019), pp. 2174–2203.

[6] R. Courant and F. John, *Introduction to Calculus and Analysis*, Vol. I, Springer, New York, 1991.

[7] A. Dobra and N.E. Williams, *Spatiotemporal detection of unusual human population behavior using mobile phone data*, PLoS ONE 10 (2015), p. e0120449.

[8] D.T. Duncan, D.A. Hickson, W.C. Goedel, D. Callander, B. Brooks, Y.T. Chen, H. Hanson, R. Eavou, A.S. Khanna, B. Chaix, S. Regan, D.P. Wheeler, K.H. Mayer, S.A. Safren, M.S. Carr, C. Draper, V. Magee-Jackson, R. Brewer, and J.A. Schneider, International Journal of Environmental Research and Public Health 16 (2019), p. 1922.

[9] K. Elgethun, M.G. Yost, C.T. Fitzpatrick, T.L. Nyerges, and R.A. Fenske, *Comparison of Global Positioning System (GPS) tracking and parent-report diaries to characterize children's time–location patterns*, Journal of Exposure Science and Environmental Epidemiology 17 (2007), pp. 196–206.

[10] B. Entwisle, *Putting people into place*, Demography 44 (2007), pp. 687–703.

[11] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez, and J. Laurila, *Towards Rich Mobile Phone Datasets: Lausanne Data Collection Campaign*, in *Proc. ACM Int. Conf. on Pervasive Services (ICPS), Berlin*, July. 2010.

[12] M.P. Kwan, *The uncertain geographic context problem*, Annals of the Association of American Geographers 102 (2012), pp. 958–968.

[13] M.P. Kwan, *Beyond space (as we knew it): Toward temporally integrated geographies of segregation, health, and accessibility*, Annals of the Association of American Geographers 103 (2013), pp. 1078–1086.

[14] J.K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T. Do, O. Dousse, J. Eberle, and M. Miettinen, *The Mobile Data Challenge: Big Data for Mobile Computing Research*, in *Proc. Mobile Data Challenge Workshop (MDC) in conjunction with Int. Conf. on Pervasive Computing, Newcastle*, June. 2012.

[15] J.K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.M.T. Do, O. Dousse, J. Eberle, and M. Miettinen, *From big smartphone data to worldwide research: The Mobile Data Challenge*, Pervasive and Mobile Computing 9 (2013), pp. 752–771.

[16] J.H. Lee, A.W. Davis, S.Y. Yoon, and K.G. Goulias, *Activity space estimation with longitudinal observations of social media data*, Transportation 43 (2016), pp. 955–977.

[17] S.A. Matthews and T.C. Yang, *Spatial polygamy and contextual exposures (SPACEs): Promoting activity space approaches in research on place and health.*, The American Behavioral Scientist 57 (2013), pp. 1057–1081.

[18] J.D. Mazimpaka and S. Timpf, *Trajectory data mining: A review of methods and applications*, Journal of Spatial Information Science 13 (2016), pp. 61–99.

[19] C.N. Morrison, H.F. Byrnes, B.A. Miller, E. Kaner, S.E. Wiehe, W.R. Ponicki, and D. Wiebe, *Assessing individuals' exposure to environmental conditions using residence-based measures, activity location-based measures, and activity path-based measures*, Epidemiology 30 (2019), pp. 166–176.

[20] C. Perchoux, B. Chaix, S. Cummins, and Y. Kestens, *Conceptualization and measurement of environmental exposure in epidemiology: Accounting for activity space related to daily mobility*, Health & Place 21 (2013), pp. 86–93.

[21] D.B. Richardson, N.D. Volkow, M.P. Kwan, R.M. Kaplan, M.F. Goodchild, and R.T. Croyle, *Spatial turn in health research*, Science 339 (2013), pp. 1390–1392.

[22] M. Šimon, P. Vašát, H. Daňková, P. Gibas, and M. Poláková, *Mobilities and commons unseen: spatial mobility in homeless people explored through the analysis of GPS tracking data*, GeoJournal (2019), pp. 1–17.

[23] M. Šimon, P. Vašát, M. Poláková, P. Gibas, and H. Daňková, *Activity spaces of homeless men and women measured by gps tracking data: A comparative analysis of Prague and Pilsen*, Cities 86 (2019), pp. 145–153.

[24] G.M. Vazquez-Prokopec, D. Bisanzio, S.T. Stoddard, V. Paz-Soldan, A.C. Morrison, J.P. Elder, J. Ramirez-Paredes, E.S. Halsey, T.J. Kochel, and T.W. Scott, *Using GPS technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment*, PloS One 8 (2013), p. e58802.

[25] L.A. Waller and C.A. Gotway, *Applied Spatial Statistics for Public Health Data*, John Wiley & Sons, Hoboken, NJ, 2004.

[26] L. Wasserman, *All of Nonparametric Statistics*, Springer Texts in Statistics, Springer, New York, 2007.

[27] S.E. Wiehe, M.P. Kwan, J. Wilson, and J.D. Fortenberry, *Adolescent health-risk behavior and community disorder*, PloS One 8 (2013), p. e77667.

[28] S.E. Wiehe, A.E. Carroll, G.C. Liu, K.L. Haberkorn, S.C. Hoch, J.S. Wilson, and J.D. Fortenberry, *Using gps-enabled cell phones to track the travel patterns of adolescents*, International Journal of Health Geographics 7 (2008), pp. 22–22.

[29] S.N. Zenk, S.A. Matthews, A.N. Kraft, and K.K. Jones, *How many days of Global Positioning System (GPS) monitoring do you need to measure activity space environments in health research?*, Health & Place 51 (2018), pp. 52–60.

[30] S.N. Zenk, A.J. Schulz, S.A. Matthews, A. Odoms-Young, J. Wilbur, L. Wegrzyn, K. Gibbs, C. Braunschweig, and C. Stokes, *Activity space environment and dietary and physical activity behaviors: a pilot study*, Health & Place 17 (2011), pp. 1150–1161.

Appendix A

APPENDIX

A.1 Proofs of theoretical results

A.1.1 Proof of Theorem 1.2.1

We note that the ordinary proportional time estimator in Eq. (1.11) can be written as

$$\widehat{\pi}_{o,j} = \frac{\frac{1}{2}\sum_{i=2}^{n-1}(t_{i+1} - t_{i-1})\mathbf{1}(g_i = G_j)}{\frac{1}{2}(\mathcal{T} + t_n - t_1)}, \tag{A.1}$$

where $\mathcal{T} = t_{\max} - t_{\min}$. We will first show that the denominators of $\widehat{\pi}_{o,j}$ and $\widehat{\pi}_{c,j}$ are asymptotically the same. Assumption (S2) implies that $\frac{1}{2}(\mathcal{T} + t_n - t_1) \to \mathcal{T}$, which shows the asymptotic behavior of the denominator of $\widehat{\pi}_{o,j}$. For $\widehat{\pi}_{c,j}$, we have

$$\begin{aligned}
\sum_{i=2}^{n}(t_i - t_{i-1})\mathbf{1}(g_i = g_{i-1}) &= \sum_{i=2}^{n}(t_i - t_{i-1}) - \sum_{i=2}^{n}(t_i - t_{i-1})\mathbf{1}(g_i \neq g_{i-1}), \\
&= \mathcal{T} - \sum_{i=2}^{n}(t_i - t_{i-1})\mathbf{1}(g_i \neq g_{i-1}), \\
&\geq \mathcal{T} - M \max_i |t_{i+1} - t_i|, \\
&\to \mathcal{T},
\end{aligned}$$

where M is the constant from assumption (S3). The limit in the above equation is due to assumption (S1). Thus, the denominators of $\widehat{\pi}_{o,j}$ and $\widehat{\pi}_{c,j}$ are asymptotically the same. Next we focus on the numerators of the two estimators.

The numerator of $\widehat{\pi}_{c,j}$ can be written as

$$\sum_{i=2}^{n}(t_{i+1} - t_i)\mathbf{1}(g_{i+1} = g_i = G_j) = \sum_{i=2}^{n} A_i,$$

where $A_i = (t_{i+1} - t_i)\mathbf{1}(g_{i+1} = g_i = G_j)$. Let $B_i = \frac{t_{i+1}-t_{i-1}}{2}\mathbf{1}(g_i = G_j)$. Using Eq. (A.1), the numerator of $\widehat{\pi}_{o,j}$ can be written as

$$\frac{1}{2}\sum_{i=2}^{n-1}(t_{i+1} - t_{i-1})\mathbf{1}(g_i = G_j) = \sum_{i=2}^{n-1} B_i.$$

When $g_{i-1} = g_i = g_{i+1} = G_j$, we have $2B_i = A_i + A_{i-1}$. By assumption (S3), there are at most $2M$ number of time points t_i such that the equality $g_{i-1} = g_i = g_{i+1} = G_j$ does not hold. Thus

$$\sum_{i=2}^{n-1} B_i\mathbf{1}(g_{i-1} = g_i = g_{i+1} = G_j) \geq \sum_{i=2}^{n-1} B_i - 2M \cdot \max_i |t_{i+1} - t_i|,$$

which implies that

$$\begin{aligned}\widehat{\pi}_{o,j} &\to \frac{1}{\mathcal{T}}\sum_{i=2}^{n-1} B_i\mathbf{1}(g_{i-1} = g_i = g_{i+1} = G_j), \\ &= \frac{1}{\mathcal{T}}\sum_{i=2}^{n-1}\frac{A_i + A_{i-1}}{2}\mathbf{1}(g_{i-1} = g_i = g_{i+1} = G_j). \end{aligned} \tag{A.2}$$

Again, using the fact that there are at most $2M$ number of time points t_i such that the equality $g_{i-1} = g_i = g_{i+1} = G_j$ does not hold, we obtain

$$\sum_{i=2}^{n-1} A_i\mathbf{1}(g_{i-1} = g_i = g_{i+1} = G_j) \geq \sum_{i=2}^{n} A_i - (2M+1) \cdot \max_i |t_{i+1} - t_i|,$$

$$\sum_{i=2}^{n-1} A_{i-1}\mathbf{1}(g_{i-1} = g_i = g_{i+1} = G_j) \geq \sum_{i=2}^{n} A_i - (2M+1) \cdot \max_i |t_{i+1} - t_i|.$$

It follows that

$$\begin{aligned}\widehat{\pi}_{c,j} &= \frac{\sum_{i=2}^{n}(t_i - t_{i-1})\mathbf{1}(g_i = g_{i-1} = G_j)}{\sum_{i=2}^{n}(t_i - t_{i-1})\mathbf{1}(g_i = g_{i-1})} \\ &\to \frac{1}{\mathcal{T}}\sum_{i=2}^{n} A_i, \\ &\to \frac{1}{\mathcal{T}}\sum_{i=2}^{n-1}\frac{A_i + A_{i-1}}{2}\mathbf{1}(g_{i-1} = g_i = g_{i+1} = G_j),\end{aligned}$$

which is the same limit in Eq. (A.2) we obtained for $\widehat{\pi}_{o,j}$. Therefore the numerators of $\widehat{\pi}_{o,j}$ and $\widehat{\pi}_{c,j}$ are asymptotically the same, which proves that $\widehat{\pi}_{o,j}$ and $\widehat{\pi}_{c,j}$ are asymptotically equal.

A.1.2 Proof of Theorem 1.2.2

Theorem 1.2.1 proves that the two estimators are asymptotically equivalent. Thus, we only need to derive the convergence of one of the two estimators to the true activity distribution $\pi = (\pi_1, \ldots, \pi_N)$ from Eq. (1.9). In what follows we focus on the conservative proportional time estimator.

Without loss of generality, we assume that there exist $K \geq 1$ disjoint time intervals in which the individual is inside grid cell G_j, i.e., there are $[a_1, b_1], \cdots, [a_K, b_K]$ such that $a_i < b_i < a_{i+1}$ for $i = 1, \ldots, K-1$, $t_{min} \leq a_1$, $b_K \leq t_{\max}$ and

$$\{t : G(t) \in G_j\} = [a_1, b_1] \cup \cdots \cup [a_K, b_K].$$

Since, in the definition of the true activity distribution π, T follows a uniform distribution on the reference time frame $[t_{\min}, t_{\max}]$, we can express π_j as

$$\pi_j = \mathsf{P}(G(T) \in G_j) = \sum_{k=1}^{K} \mathsf{P}(T \in [a_k, b_k]) = \frac{1}{\mathcal{T}} \sum_{k=1}^{K} (b_k - a_k).$$

As before, $\mathcal{T} = t_{\max} - t_{\min}$.

For the interval $[a_k, b_k]$, we let t_{i_*} be the first observation time after a_k, and $t_{i_{**}}$ be the last observation time before b_k:

$$t_{i_*} \geq a_k, \quad t_{i_*-1} < a_k, \quad t_{i_{**}+1} > b_k, \quad t_{i_{**}} \leq b_k.$$

Because $G(t) \in G_j$ for all $t \in [a_k, b_k]$, we have $g_i \in G_j$ for all $i \in \{i_*, i_* + 1, \ldots, i_{**}\}$. The conservative proportional time estimator estimates the length of the interval $[a_k, b_k]$ based on the length of the interval $[t_{i_*}, t_{i_{**}}]$. The corresponding error is

$$\begin{aligned} |(b_k - a_k) - (t_{i_{**}} - t_{i_*})| &\leq t_{i_*} - a_k + b_k - t_{i_{**}}, \\ &\leq (t_{i_*} - t_{i_*-1}) + (t_{i_{**}+1} - t_{i_{**}}), \\ &\leq 2 \max_{i=1,\ldots,n-1} |t_{i+1} - t_i| \to 0, \end{aligned}$$

due to assumption (S1).

By applying the above argument to each interval $[a_k, b_k]$, $k = 1, \ldots, K$, we conclude that

$$\sum_{i=2}^{n}(t_i - t_{i-1})\mathbf{1}(g_i = G_j) \to \sum_{k=1}^{K}(b_k - a_k).$$

Because

$$\sum_{i=2}^{n}(t_i - t_{i-1})\mathbf{1}(g_i = G_j) \geq \sum_{i=2}^{n}(t_i - t_{i-1})\mathbf{1}(g_i = g_{i-1} = G_j) - M \cdot \max_{i=1,\ldots,n-1} |t_{i+1} - t_i|,$$

we further conclude that

$$\sum_{i=2}^{n}(t_i - t_{i-1})\mathbf{1}(g_i = g_{i-1} = G_j) \to \sum_{k=1}^{K}(b_k - a_k).$$

This proves the convergence of the conservative proportional estimator to the true activity distribution:

$$\begin{aligned}
\widehat{\pi}_{c,j} &\to \frac{\sum_{i=2}^{n}(t_i - t_{i-1})\mathbf{1}(g_i = g_{i-1} = G_j)}{\mathcal{T}}, \\
&\to \frac{\sum_{k=1}^{K}(b_k - a_k)}{\mathcal{T}}, \\
&= \pi_j.
\end{aligned}$$