Published in final edited form as:

Int J Geogr Inf Sci. 2018 ; 32(7): 1485–1504. doi:10.1080/13658816.2018.1434888.

How many days are enough?: capturing routine human mobility

Kevin Stanley^a, Eun-Hye Yoo^b, Tuhin Paul^a, Scott Bell^c

^aComputer Science, University of Saskatchewan, Saskatoon, SK, Canada;

^bDepartment of Geography, University at Buffalo, Buffalo, NY, USA;

^cGeography and Planning, University of Saskatchewan, Saskatoon, SK, Canada

Abstract

Wedding mobile phone sensor technology and human spatial behaviour has great potential. The ubiquity of Global Positioning Systems (GPS) technology has made gathering data about human mobility simpler, more precise, and with higher fidelity, providing minute-by-minute records of the locations of cohorts from dozens of participants. While this data provides a strong basis for Geographic Information Science research, it also constitutes an invasion of the participants' privacy and can provide more information than researchers require to answer their questions. As an ethical and practical consideration, researchers should gather only as much data as they need. In this paper, we take three weeks of GPS traces from over a hundred student participants in mobile phone-based tracking studies and show that fewer than 14 days of data is necessary to establish complete activity spaces. We define 'complete' as the point at which marginal information gains become negligible according to a pairwise temporal analysis of the Kullback–Leibler (KL) divergence of the spatial (bivariate) histogram through time. For the fixed level of information difference, observable in the data, impacts due to individual variability, population composition, and spatial resolution are evident. However, all populations at each level of resolution examined in the paper demonstrated convergence to low divergence levels occurred within a matter of days, and to negligible information gain in less than two weeks. The methods described in the paper represent a novel metric useful to understand the interaction between measurements and information in human mobility.

Keywords

Activity spaces; Kullback–Leibler (KL) divergence; Global Positioning Systems (GPS)

1. Introduction

Traditionally, research on human space-time activity has relied on travel diaries, questionnaires, or interviews (Hägerstrand 1975, Griffin and Bell 2009). While these methods provide a systematic record of activities and movements, they are cost-, time-, and labour intensive (Janelle *et al.* 1988, Crosbie 2006). In addition, these survey data are often

CONTACT Scott Bell scott.bell@usask.ca.

Disclosure statement

No potential conflict of interest was reported by the authors.

subject to recall bias and questionable reliability (Kitterød 2001, Steinle *et al.* 2013). The increased availability of location-aware technologies such as mobile phones, Global Positioning Systems (GPS)-enabled devices, wireless local area networks and earlier personal digital assistants have enabled researchers to collect information on human spatial behaviour at high spatial and temporal resolution. Mobile devices also have changed the way researchers across multiple disciplines, including geography, public health, planning, traffic engineering, and economic forecasting, analyse data on human spatial behaviours, as they recognized that statistical patterns that characterize human mobility, such as size, shape, and randomness, pose (Gonzalez *et al.* 2008, Song *et al.* 2010a, Rhee *et al.* 2011, Gong *et al.* 2012, Poletto *et al.* 2012, Wesolowski *et al.* 2012, Blumenstock *et al.* 2015, Yuan and Raubal 2016).

The emergence of massive datasets of location data on human mobility has improved our understanding of the impact of human movement on critical outcomes such as infectious disease transmission (Vazquez-Prokopec *et al.* 2013). Wesolowski *et al.* (2012) have shown that changes in human mobility contribute to increased malaria transmission in Kenya. These patterns were identified by examining spatial associations between the high-risk regions of malaria incidence and the regional travel patterns based on the mobile phone uses of 15 million individuals. Similarly, GPS receivers have been widely used for collecting supplementary contextual information in behavioural health-related research (Kerr *et al.* 2011, Schipperijn *et al.* 2014), monitoring personal exposure to air pollutants directly while people are moving through their activity spaces (De Nazelle *et al.* 2013, Steinle *et al.* 2013, Glasgow *et al.* 2014, Yoo *et al.* 2015), and replacing traditional paper-based travel diaries in travel surveys (Shen and Stopher 2014, Paul *et al.* 2016). Tracking individual movement, particularly in health-related studies, is coupled with other measurements, such as motion sensors (accelerometers or pedometers), travel diaries, and post-processing and spatial analyses in Geographic Information Science (GI Science), to elucidate the environmental context of individual health behaviours. Almanza *et al.* (2012) showed the positive association between greenness exposure and free-living physical activity behaviour by children using the accelerometer and GPS data points collected from 386 children.

However, these studies based on individuals' location data also raise methodological, technical, and ethical challenges (Kerr *et al.* 2011, Vazquez-Prokopec *et al.* 2013). While GPS receivers enable the efficient collection of time-space data, infringement on personal information is an emergent concern. Thus, researchers, guided by ethical considerations, may not be able to collect sufficient data to truly represent individuals' routine spatial behaviours. It is important to minimize participant burden and the violation of participants' privacy, but also to collect sufficient data to maximize model reliability and validity. This can be expressed as the research question: how many days are enough to capture individual and group daily activity patterns? The answer to this question may be derived from empirical studies that analysed mobility trace data and have shown that human mobility and activity patterns are recursive and dominated by reproducible scaling laws (Brockmann *et al.* 2006, Song *et al.* 2010a). The goal here would be to establish the point at which recorded spatial behaviours are fully characterized in a given environment with only marginal gains in knowledge arriving with additional observations. This would improve participant privacy and address-related ethical concerns. Intuitively, researchers should collect data until

participants have been sufficiently localized to address their research questions, and no more.

In practice, however, few techniques exist for determining when a participant has been sufficiently localized. The lack of methods available is understandable, because continuous real-time data is required for some applications, such as search and rescue, while coarser spatial fidelity and limited temporal information may be sufficient for other applications, such as characterizing participants' routine and semi-routine daily activity space. While limited work has been done on geographic aspects of regular search patterns of spatial behaviours, Tudor-Locke *et al.* (2005) investigated the minimum number of days of pedometer monitoring required to predict weekly physical activity in adults using classic statistical methods, such as analysis of variance, intra-class correlations (ICC), and regression with the goal of understanding the source of intra- and inter-individual variability. They concluded any 3-day period can provide a sufficient estimate. Similarly, Hart *et al.* (2011) estimated the minimum number of days of monitoring to predict physical activity (PA) and sedentary behaviour in older adults. They examined the number of days needed to predict habitual PA and sedentary behaviour across pedometer, accelerometer, and physical activity log data in older adults using ICC. They concluded that 3 days of accelerometer data, 4 days of pedometer data, or 4 days of completing PA logs are needed to accurately predict PA levels in older adults. These statistics summarize the consistency of measures across days by comparing within subject variability with between subject variability.

We propose in this paper a spatially and temporally explicit metric that quantifies the extent to which data have captured the regularity in routine time-activities. Activity space by definition refers to the local areas within which people travel during their daily activities (Hägerstrand 1975, Mazey 1981, Paul *et al.* 2016), and is considered a defining concept of human spatial behaviours in geography (Golledge 1997, Paul *et al.* 2016). Our proposed metric is grounded in this well-defined concept and the assumption that human mobility is regular and predictable (Song *et al.* 2010a, 2010b). This metric measure spatial variation within participant, not between participants. As predicted by Anchor Point hypothesis (Couclelis *et al.* 1987, Golledge *et al.* 1995), there are a relatively small but stable number of anchors in individuals' activity spaces, but; these places will vary across individuals. Common anchor point types include home, work, recreation, and shopping. Unlike other statistics that primarily focus on the regularity of health indicators, such as physical activity levels, exposure to air pollution concentrations, the proposed metric summarizes the completeness of regular human activity patterns using the Kullback–Leibler (KL) divergence. The metric indicates the degree and rate to which the cumulative-visit-frequency-spatial histogram, also known as bivariate histogram (Tobler 1979) or heatmap (Wilkinson and Friendly 2009), stabilizes over time. It estimates how much uncertainty remains after a given period of observation or data collection, allowing decisions on whether to continue, suspend, or analyse measurement. We demonstrate the application of KL divergence to real-world GPS data collected for different health studies, and estimate a minimum time required to capture individuals' routine time activities.

2. Methods

2.1. Algorithm description

To determine the duration of time necessary to establish an activity space for a participant using location sensing technology, we evaluate the relative impact of an additional unit (e.g. day) of data on the change of spatial histogram using a non-parametric statistic. Here, we refer to the spatial histogram as the frequency of an individual visiting a specific location during a study period. This is often also termed as a bivariate histogram or heatmap, but in other academic parlance those terms denote any two dimensional discrete distribution. For cross-disciplinary clarity, we use spatial histogram to denote the time spent visiting a particular geographic location. Specifically, we impose a regular grid with a specified cell size and calculate the density at each cell based on the total number of time steps an individual spends in to the cell. A stylized example showing the evolution of a spatial histogram through time is provided in Figure 1. The initial location data at time (t), illustrated by dots with light grey shade, impose a discrete distribution of their spatial pattern (bottom left). Additional measurements on the time ($t + 1$), denoted as dots with darker grey shade, skew the distribution (bottom right). This skewing is encoded in the mutual information or KL divergence between the two time points.

For each day, we update a spatial histogram by incorporating additional data points observed. If, after adding new data, the spatial histogram is essentially identical, then the impact of adding that data is marginal. KL divergence calculates the relative information between two distributions, in essence answering the question: ‘Given the distribution of X, how much additional new information do I need to create the distribution of Y?’ (Cover and Thomas 2012). If the distributions are identical, this value is zero, but as they begin to differ, the KL divergence increases by the amount of new information present in the distribution of Y. The classic representation of KL divergence is given in Equation (1).

$$D(Y|X) = \int Y \log\left(\frac{Y}{X}\right) \quad (1)$$

It should be noted that spatial configuration (or arrangement) of each participant’s location data has no effects on the calculation of KL divergence as long as the same ordering is applied to the consecutive spatial histograms. That is, to calculate the KL divergence, the two-dimensional histogram needs to be unwound into a single-dimensional representation. Depending on the method of unwinding, adjacent cells might be adjacent along lines of latitude but latitude information would be ambiguous. Similarly, the ordering of cells in the histogram could be randomly assigned, but as long as the same ordering is maintained for both X and Y, the KL divergence will return the same value. Locations in this treatment of grid cells cover the study area with spatial resolutions of 100, 400, and 1600 m. As described above, if, over time, the geographic arrangement of novel places visited declines, the histogram and KL divergence will approach zero. It is worth noting that the KL divergence is distinct from, but closely related to, the concept of mobility entropy (Song *et al.* 2010b). Mobility entropy attempts to quantify the amount of information present in a sequence of location visits. KL divergence attempts to quantify the difference in information

between two histograms, accumulated from two different sequences of location visits. While mathematically related, these two techniques answer different questions of the data, and are therefore useful to divergent and distinct applications within geographic data analysis. However, KL divergence only encodes one type of ‘situatedness’; that the probabilities associated with the spatial histogram have stabilized over time. Crucially this eliminates temporal variation present at less than the aggregation interval. For example, if an individual has shift work, their temporal behaviour might be highly variable as they leave for or return from work, but their spatial behaviour would be stable as workplace and home. The technique described here captures variability in the latter but not the former.

Because KL divergence encodes the incremental information needed to produce one distribution from another, we expect the divergence to decrease over time regardless of the patterns of mobility inherent in the data. This is the logical result of there being a finite number of ‘places’ (grid cells) that can be visited. Once visited, a cell contains information as part of the spatial histogram. However, we also expect that people with more mobile life style, who visit a larger number of places (and by extension to how we partition space, cells) will yield a slower decline in KL divergence values than participants with static spatial behaviours. The behaviour of KL divergence is better illustrated by examining hypothetical cases of extreme spatial behaviour. Because our approach for employing KL divergence compares individual participants through time, it is less sensitive to the number of anchor points than the consistency of those anchor points through time. Someone with many anchor points will have their KL divergence converge nearly as quickly as someone with fewer anchor points if those points are consistent through time because the information difference between consistent histograms is consistent.

The behaviour of KL divergence as a metric could be illustrated by considering extreme situations. Consider a case of an individual who spends entire time in a single cell during the study period. This could be due to one’s social circumstance (e.g. a prisoner confined to a single building) or the spatial scale of analysis (e.g. a grid cell with the size of the study area). In this case, all probability mass is held within a single cell, and the histogram does not change no matter how long we measure their mobility, meaning that the divergence is always zero because all subsequent histograms are identical. For other extreme case, consider an individual who never visits the same place twice during the study period. After any length of time, the histogram is characterized by a uniform distribution across all visited cells (as every cell is only visited once), where the probability of visiting each cell amounts to $\frac{1}{N_{vc}}$, where N_{vc} is the number of visited cells. An additional unit of time will yield the same distribution over more cells, but less information is gained over time because, for a constant increase in the number of cells, a smaller proportionate amount of information is gained with respect to all previously visited cells. Therefore, information gain diminishes. Given that the KL divergence of the spatial histogram will generally decline as duration goes to infinity, the utility of this technique is to quantify how fast it declines, or when it has declined sufficiently to consider a potential study participant sufficiently localized.

The KL divergence is not free of limitations as documented elsewhere (Cover and Thomas 2012). The first is asymmetry, that is, the amount of additional information required to build

X from Y may not be the same as that required to build Y from X. A trivial example of this phenomenon is when Y is a subset of X: no additional information is required to build Y from X since it is completely contained, but a substantial amount may be required to build X from Y. Fortunately in the case examined here, this distinction is irrelevant because histograms are compared through time and, time's arrow forces a single direction of comparison upon the analysis. In fact, the latter histogram contains the data in the preceding histogram, making such data logical for analysis using KL divergence.

A second shortcoming of the KL divergence is the absence of a significant test unlike well-studied ANOVA analysis. This is a typical drawback for general non-parametric methods, as no assumptions about the underlying structure of the distributions can be made allowing the derivation of significance or performance metrics. As an initial estimate of when an individual or population are sufficiently localized, we denote the time when the divergence has stabilized to less than 10% of the maximum divergence as sufficient localization. However, this threshold assignment must be grounded in the data and its context, as with other non-parametric analysis techniques. This value was chosen as 0.1 in the analysis, as a divergence of 1 typically encompassed all but the most extreme outliers.

Because we limit the direction of comparison based on time, the algorithm unfolds in a straightforward manner. For all points in time, subject to a temporal unit of analysis, we calculate the spatial histograms (probability of visiting a place located within a given grid cell) from the location data measured up to the time (t) for each participant. From these histograms, we compute the KL divergence between consecutive time instants, i.e. the t -th and $(t+1)$ -th samples. For individuals who are appropriately situated, this value should trend towards zero; for populations, the mean should also trend towards zero.

We quantified other two related measures on individuals' or population's spatial behaviours: the temporal derivative of the divergence, which indicates the rate of information gain for subsequent measurements, and the integral of the divergence which indicates the cumulative information gain for each stage of measurement. The derivative should trend towards zero as participants are increasingly localized, and provides potentially more intuitive rate-based representation: when the value reaches zero, information gain has stalled. The integral should provide an indication of the declining returns based on the amount of information gained per unit measure. The shape of both the derivative and integral will depend on the form of the divergence itself.

2.2. Experimental setup

Data for the analysis were cultivated from three studies run between July and December of 2016. These studies are the 7th, 8th, and 9th iterations of the ongoing Saskatchewan Human Ethology Dataset (SHED) collection (Knowles *et al.* 2014, Stanley *et al.* 2016). Participants in all three studies were members of the University of Saskatchewan community who lived in or near the city of Saskatoon. Data was collected using the Ethica Data Systems app, a commercial product based on earlier iEpi work in the department of Computer Science at the University of Saskatchewan Ethica (2017). Data collection ranged between 4 and 7 weeks for the datasets in question, but we considered only 21 consecutive days from each study for presentation consistency. SHED datasets typically collect a wide variety of survey,

sensor, and environmental data, but again we focused on only location data in this work. All location data were collected using Android smartphones, in which data were provided by the abstract Android location service. This service provides the most accurate location inferred from a combination of WiFi routers, cell towers and A-GPS satellites, as computed by Android. Data were collected at approximately 5-min intervals (guaranteed at 10 min intervals, with potentially every second interval moved slightly from 5 min to align with other Android processes' schedules for power savings), for a period of a minute to preserve participants' smartphone's batteries. Participants were monitored as long as the phone had power. Participants were instructed to plug in their phones at night. All data were collected in accordance with institutional ethics approval and with informed consent of participants. As all participants were recruited from the same university pool, some participants are common across datasets.

SHED7 was conceived as a study of built infrastructure walkability and propensity to walk, but inadvertently coincided with the Canadian release of Pokemon Go. SHED8 was conceived as a study of a transit system labour disruption, which was settled before job action took place and a follow up of Paul *et al.* (2016). SHED9 was a study of the impact of the closure of a covered skywalk for students on-campus commuting patterns. Each of these studies contains extensive GPS records, which can be leveraged for research on fundamental mobility metrics. A summary of the type and number of records in each dataset is provided in Table 1 for reference.

Figure 2 presents the demographic information collected from a questionnaire for each dataset. Each row represents a unique dataset, whereas each column represents the genders, age ranges, and primary roles of participants on campus. As is apparent from the demographics, our data are skewed towards participants who are undergraduate or graduate students in their twenties, as would be expected from a university-centred cohort. We also have a skew towards female participants, which again is not unusual in university cohorts. The following two factors should be noted regarding our data collection method and participants: First, our participants include students who live on campus; and second, the unit of analysis ranges from 100 to 1600 m. Due to a limitation in the Ethica data collection framework, a 50 m bin was the absolute smallest unit we could accept as indivisible. While this shortcoming has been addressed in Ethica software, it still limited lower bound of our analysis. Thus, having the smallest bin set at 100 m increases the likelihood that movement from a cell represents actual movements and not perturbations of position caused by uncertainty. The 400 and 1600 m are, respectively, used to test the veracity of KL divergence for characterizing the bulk of a person's spatial behaviours. We anticipated that the confluence of these two factors might reveal spatial histograms that move to zero quickly do the proximity of anchor points for some students. For example, 'home' (dorm) and 'work' (class rooms) could be potentially mapped to a single spatial bin.

Data were transferred from Ethica Data Services Cassandra servers to a MySQL database hosted at the University of Saskatchewan. Data analysis was performed using the scipy family of scientific computing libraries as curated in the Anaconda distribution (v1.4.3) using the Python language (v.3.6.0). We filtered data for accuracy, temporal and spatial extent. Only data with an Android-reported 100 m or better accuracy were retained. Data

analysis was constrained to the first 21 days of each study for temporal consistency. As we are primarily interested in urban movement patterns, we limited geographic extent to Saskatoon and surrounding regions, incorporating a range of locations from WGS 1984 (52.058367, -106.7649138128) to (52.214608, -106.52225318). To ensure metric performance was the result of human mobility and not a measurement artefact, participants who did not report any data for a full day during the 21 days in question were discarded. KL divergence is robust to missing days, simply returning a divergence of zero denoting identical cumulative histograms on the (t)-th and ($t + 1$)-th days. However, this artefact drives the population mean divergence down artificially, so these data were removed. Table 2 contains information about the original number of participants (Part.) and those are excluded due to the imposed constraints, and the impact of filtering on the total number of location (Loc.) records analysed after all filtering steps were complete.

Because KL divergence requires discrete histograms for comparison, continuous location information was transformed into a discrete form over a regular grid with a cell size (spatial resolution) of 100, 400, and 1600 m, respectively. Data were transformed into UTM coordinates, then a set of bins (or tiles) was imposed over the area of interest. Frequency was calculated as the number of GPS samples observed in the bin, echoing the approach taken in Paul *et al.* (2016). We chose the multiple cell sizes to assess the impact of changing the finite number of possible locations. The cell size of 1600 m is large enough to cover an area that are not generally associated with walking distance, and the cell size of 100 m is small enough to decrease the likelihood of trips on adjacent blocks from falling in the same bin. Histograms for each participant were generated by aggregating all GPS records for the entire study period for a total of maximum 21 days. The KL divergence was computed from a pair of consecutive days (t , $t + 1$) for all 21 days ($t = 0, \dots, 20$) for each participant. The simple discrete KL divergence derivative was also calculated by computing the daily change in divergence.

3. Results

To illustrate the dynamics evident in the spatial histograms, we have plotted heatmaps of 10 randomly chosen participants on two arbitrarily chosen days from the SHED9 dataset in Figure 3. To comply with ethics requirements which do not allow participant data to be presented in an identifiable form (e.g. showing home locations that could be used to cross-reference with identifiable material), the image has been cropped to the area of the university campus which all participants attended. Additionally, this data has been presented as an aggregated form over ten participants to obscure individuals' spatial behaviours. However, in the actual calculation of KL divergence described above, comparisons were performed per participant on an unaggregated form.

Two effects are clearly shown in the histograms of Figure 3. First, the impact of the duty cycle-based sampling employed by Ethica Software was illustrated. On the other hand, holes in the commute patterns due to off duty cycles have been filled in on the second day ($t + 1$), where transit along College Drive at the southern extent of the figure is more complete. Second, behaviours consistent with the differences between Monday, Wednesday, Friday classes and all classes are apparent. Where the day t (Monday) covers a subset of buildings

on campus used for undergraduate instruction, both days t and $t + 1$ (Monday and Tuesday) cover the extent of buildings commonly used for undergraduate instruction on campus. This demonstrates the variability in the records due to both sensed locations and individual habits.

To provide an example of the aggregate behaviour captured in the datasets, we mapped the spatial histograms aggregating all participants for the first 2 days with a 400 m bin size and the results are shown in Figure 4. These aggregate maps are not representative of what is being measured with KLD, but provide a spatial distribution of movement of our sample.

Even after only 2 days of observations the differences are difficult to spot. This is partially due to the fact that most activities of participants occurred at the university campus (gold and turquoise) compared to the rest of the study region. The purple background is a uniform value of zero, but other locations are highly improbable and fan out over typical student residential and shopping districts.

Results of the KL divergence calculated over time for each dataset are shown in Figures 5–7. In each Figure, box plots on the left represent the KL divergence and on right the discrete derivative of the KL divergence, across participants for a given day over 21 days of the study. Boxes represent the inner quartiles, whiskers the outer quartiles, circles outliers, and the central line for the median. One boxplot is shown for each day after the second day of the experiment, as our algorithm proceeds using pairwise comparisons. KL divergence is calculated for each participant for each day; the boxplots are then

As expected, the KL divergence tended towards zero as more data were added to the participant's spatial histograms. Mean divergence falls faster with a lower starting value at coarser spatial resolution than finer spatial resolution. For example, divergence reaches a value of less than 0.1 within 2 weeks at 100 m, within 12 days at 400 m and within 8 days for 1600 m. Our findings on the extent to which there is little variability as bins get larger concur with the anchor point hypothesis that most of our movement is confined to a relatively small number of stable places (anchors). The variance across the study populations follows a similar trajectory for each dataset, narrowing to a fraction of their initial span by the midpoint of the study period. For the populations under study, the spatial histograms have stabilized within a period of 2 weeks. Some heterogeneity in divergence is noted among participants, as expected. Everyone has differing complexities in spatial behaviour and differing time complexity to how those patterns are expressed. However, this heterogeneity is overwhelmed by the convergence of the divergence.

Additionally, we found some heterogeneity between datasets exists. In particular, the initial variance and mean of the divergence and derivative are substantially lower in SHED9 than SHED8 or SHED7. Given that all the three datasets are collected from members of the same university community, the differences tend to be marginal and limited to the impact of outliers. The primary effect of converging divergence is maintained across datasets, and the time to nominally situated populations is largely the same.

Notably, both the mean and variance of the divergence and derivative are driven to zero, implying that across-participant information gain is increasingly limited. Negative values in the derivative of the KL divergence curve correspond to decreases in the daily rate of

information gain. The results suggest that participants are sufficiently situated after 1 week at least for the populations studied and almost fully situated after two. Additional measurements have little impact on the net spatial histogram which underlies their activity space. It is worth noting that this might not be universally true, and it is highly anticipated that KL divergence curves of other populations may be characterized by faster or slower convergence depending on their spatial behaviour.

We found that the trend is broadly and consistently downward in all three datasets, while variability exists between individuals in a population. We also found that individual divergences experience a degree of day-to-day instability. To explore how the divergence decays towards zero as the day of observation increases, we calculated the mean of KL divergence values across participants each day and fit decay functions using the non-linear regression system Eureqa from Nutonian Inc. Dubáková (2011). The results indicate that the decay properties are remarkably simple and consistent across all three datasets as the central tendency of the population could be represented as a simple inverse proportion. That is, the mean divergence could be well described by the simple relationship as

$$\bar{x}(t) = At^{-1} \quad (2)$$

where $\bar{x}(t)$ is the mean divergence value across participants on a given day t , $t = 1, \dots, T$. We assume that the parameter of interest of Equation (2), denoted as A , is constant and unique to each population and resolution, and estimated it through a non-linear regression. Table 3 summarizes the estimated value ($\hat{A}$) for this constant parameter, and a measure of model fit (R^2) for each dataset across different spatial resolutions. Because the fit is non-linear, R^2 values must be employed with caution, but remain a useful signpost of quality of fit.

As noted in Table 3, goodness of fit was exceptional for such a simple relationship describing a complex phenomenon. This result lends confidence to our ability to make conclusions about the impact of additional days worth of measurement in these types of studies.

4. Discussion

In this work, we have introduced a novel method and metric for determining the degree to which an individual or population has been localized or situated with respect to their regular spatial behaviours. We summarized individuals' spatial movements by their cumulative spatial histogram. The proposed metric depends on relative information characterized by KL divergence of the spatial histogram through time of observation. As additional data are incorporated into the spatial histogram, we can quantify the relative information gained from the newly observed data and determine if the histogram has become more complex or remained the same. Our method explores the reliability or precision of an activity space acquired over time, but does not pertain to the accuracy of that representation or its applicability to a given scenario. It tells the researcher how complete the summary of a person's spatial behaviour is up to the day of observation, not whether that complete representation is sufficient for their needs. For example, the number of an individual's

activity space may converge over a two week period to a precise value while he or she is on vacation in Hawaii, it would not be an accurate representation of one's activity space at home in Delaware. While this method can help quantify the precision of measurements, and the rate at which maximum precision is reached, it does not guarantee that the measurements are representative of the human behaviour of interest, just that no additional information is available during the period of data collection. It is left to researchers to ensure that the precision estimates are useful for their study by taking into account the context of interest (e.g. vacation or daily life, Hawaii or Delaware).

In all three datasets, both the KL divergence and its derivative tended towards zero as time progressed, implying diminishing incremental information was provided by additional measurements. The divergence reached a stable value after approximately 2 weeks, but was characterized by a small variance about that mean, implying that participants' habits were not fixed within that time frame, but rather that the larger activity space was captured. In summary, while additional information might be added beyond a temporal threshold (such as 14 days), the incremental data did little to alter the spatial histogram.

Three related measures from the KL divergence were identified, which could be useful in future research when researchers attempt to quantify the degree to which a participant or population has been situated. First, the rate of KL divergence converging towards its final mean value (approaching zero, or no change in the locations/bins visited during daily activity) contains information about the regularity of habits, the timescale over which this variability manifests. Little variability was noted in our analysis. The final divergence to which the population mean converges expresses the degree to which drift occurs in spatial occupancy patterns through time. As expected, little divergence is noted after convergence in all experiments, as student life is relatively routine. Finally, the variance around the final value provides a measure of the minimum precision possible. Similarly, the variance in the divergence indicates how stable the final information differential is. Small and near constant variances, as observed in our study, may indicate the normal variation in routine. While these metrics are all promising, substantial future experimental and analytic work would be required to fully describe and validate them.

We have analysed the profile of the divergence decay and found it to be remarkably consistent across different datasets and spatial scales of analysis. Furthermore, we found this decay was best described as a simple inverse proportion. If one considers the integral of this curve, representing the cumulative information gain for each day of measurement, a logarithmic curve is the most appropriate. This implies, at least for the populations of interest, that there is a clear law of diminishing returns for additional measurement. It would take twice as long to achieve the same information gain due to the logarithmic cumulative return on measurement investment. However, more datasets, covering broader demographics should be considered before our findings are generalized to represent human spatial behaviour.

The process described here is applicable to any evolving spatial histogram through time. However, the finding that it took approximately 2 weeks to localize a population is only applicable to the specific demographic considered in this study. We feel most strongly that

the size and extent of the larger space within which the population's spatial behaviour is generally bounded (its action space) is critical. In the case of Saskatoon, we did not observe meaningful activity outside the city's urban extent. This leaves us with an interesting geographic conundrum: what role does geography (extent, shape, population density, infrastructure, etc.) have on the predictability of human spatial behaviour? It would be reasonable to expect that a similar convergence would be evident for student populations in other mid-sized college towns in North America; however, a more general conclusion is not warranted from the data analysed. Our findings were surprisingly robust to spatial resolution. While variability in convergence rate and final variance was observed between resolutions, this was significantly less than observed in other information theoretic measures (Song *et al.* 2010b). A smaller divergence evident between spatial resolutions from initial measurements is mathematically sensible, because there is less information contained in a spatial histogram with larger bins, as there are fewer possible states (locations) that a participant could be in. We suspect that the spatial robustness is primarily due to the fact that we are studying a static histogram through time, and that large changes in relative information corresponding to unusual or diverse participants' behaviours will manifest over similar time scales. For example, the occurrence of the weekend is independent of spatial scale, and therefore capturing the weekend behaviour for the first time will not depend on the size of cell. However, at critical cell sizes we expect phase changes in the divergence as the majority of behaviours become subsumed into a small number of cells. The rate at which convergence occurs might vary with scale, but because we selected a day as the unit of our temporal resolution, finer grained behaviour may have been aggregated out.

While this study presents a novel method for evaluating the reliability and precision of a GPS-derived spatial histogram, it does suffer from a number of limitations. First, as noted above, the skewed demographic under consideration does not allow us to broadly generalize our findings. This is easily addressed through future work analysing datasets with more diverse behaviours and population in gender, age, ethnicity, occupation, and geography. However, the skewed dataset demonstrates the efficacy of the technique in general, at least for datasets characterized by city-scales, and regular mobility. Validation and extension of this method against a myriad of datasets, but human and non-human is a major component of our future work. Second, while the new derivative metrics (time to converge, mean resting divergence, divergence variance) provide some insight into the level of situatedness achieved, they have not been validated, and their absolute interpretation has not been established. Third, the KL divergence only provides a measure of the precision of the cumulative spatial histogram through time, and can not provide immediate insight into whether specific geographic questions related, for example, to the spatial impacts on mobility, exposure to pollution and pathogens or access to facilities are affected. It instead addresses questions of whether incremental information benefit can accrue from further measurement. Other metrics must be used to determine if the situated spatial histogram is sufficient for specific applications. Fourth, the KL divergence is not the only non-parametric test which could be applied to spatial histograms in this manner. Alternative examples such as the Kolmogorov–Smirnov test, simple squared histogram distance and others could be examined for consistency of results. Finally, we used the raw spatial histograms for comparison. Some may argue that spatial smoothing should be applied on the histograms

prior to divergence calculation to disperse the impact of specific GPS measurements over large spatial extents. This additional analysis is an intriguing target for future research.

5. Conclusions

Human spatial behaviour is largely routine and limited by the spatial and social dictates of geography. Intuitively, this should mean that an activity space for an individual can be inferred within a finite amount of time. Quantifying that amount of time is still an open question in GI Science. In this paper, we have made an important methodological contribution, by proposing to use the KL divergence of the spatial histogram through time as a method of quantifying the declining returns of additional measurement. We found that, for the university student-oriented datasets we analysed, little additional information or within population information variance was observed after 2 weeks. We expect that similar populations will have similar convergence rates, but that convergence rates for more mobile entities, such as taxis would converge more slowly, and exhibit more baseline variability, but this is left for future work. Using empirical regression fits, we found that the decay could be represented by a simple inverse proportional model, implying a logarithmic increase in total information through time. These findings constitute an important step in establishing theoretical and empirical methods for determining required data volumes for geographic pattern analysis/modelling. This work provides several interesting opportunities for future work in validation, application, and theoretical development.

Acknowledgements

This research was also supported by a Social Science and Humanities Research Council of Canada Partnership grant, GeoThink. We appreciate the anonymous reviewers who helped improve the quality of this paper greatly.

Funding

This research is partially supported by grant [Grant Number 5R01GM108731] from the National Institute Of General Medical Sciences of the National Institutes of Health (NIH/NIGMS) awarded to the second author; Natural Science and Engineering Research Council of Canada [Grant Number RGPIN-2015-06318]; Social Sciences and Humanities Research Council of Canada [Grant Number 895-2012-1023]; NSERC Discovery [Grant Number #356043-2008].

References

- Almanza E, et al. 2012 A study of community design, greenness, and physical activity in children using satellite, GPS and accelerometer data. *Health & Place*, 18 (1), 46–54. doi:10.1016/j.healthplace.2011.09.003 [PubMed: 22243906]
- Blumenstock J, Cadamuro G, and On R, 2015 Predicting poverty and wealth from mobile phone metadata. *Science*, 350 (6264), 1073–1076. doi:10.1126/science.aac4420 [PubMed: 26612950]
- Brockmann D, Hufnagel L, and Geisel T, 2006 The scaling laws of human travel. *Nature*, 439 (7075), 462–465. doi:10.1038/nature04292 [PubMed: 16437114]
- Couclelis H, et al. 1987 Exploring the anchor-point hypothesis of spatial cognition. *Journal of Environmental Psychology*, 7 (2), 99–122. doi:10.1016/S0272-4944(87)80020-8
- Cover TM and Thomas JA, 2012 *Elements of information theory*. Hoboken, NJ: John Wiley & Sons.
- Crosbie T, 2006 Using activity diaries: some methodological lessons. *Journal of Research Practice*, 2 (1), 1.
- De Nazelle A, et al., 2013 Improving estimates of air pollution exposure through ubiquitous sensing technologies. *Environmental Pollution*, 176, 92–99. doi:10.1016/j.envpol.2012.12.032 [PubMed: 23416743]

- Dub áková R, 2011 Eureka: software review. *Genetic Programming and Evolvable Machines*, 12 (2), 173–178. doi:10.1007/s10710-010-9124-z
- Ethica, 2017 Ethica data. [online]. Available from: <https://community.ethicadata.com/hc/en-us> [Accessed 17 March 2017].
- Glasgow M, et al. 2014 Using smartphones to collect time–activity data for long-term personal-level air pollution exposure assessment. *Journal of Exposure Science and Environmental Epidemiology*, 26 (4), 356. [PubMed: 25425137]
- Golledge RG, 1997 *Spatial behavior: a geographic perspective*. New York, NY: Guilford Press.
- Golledge RG, Dougherty V, and Bell S, 1995 Acquiring spatial knowledge: survey versus route-based knowledge in unfamiliar environments. *Annals of the Association of American Geographers*, 85 (1), 134–158.
- Gong H, et al. 2012 A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems*, 36 (2), 131–139. doi:10.1016/j.compenvurbsys.2011.05.003
- Gonzalez MC, Hidalgo CA, and Barabasi AL, 2008 Understanding individual human mobility patterns. *Nature*, 453 (7196), 779–782. doi:10.1038/nature06958 [PubMed: 18528393]
- Griffin AL and Bell S, 2009 Applications of signal detection theory to geographic information science. *Cartographica: the International Journal for Geographic Information and Geovisualization*, 44 (3), 145–158. doi:10.3138/carto.44.3.145
- Hägerstrand T, 1975 *Space, time and human conditions*.
- Hart TL, et al. 2011 How many days of monitoring predict physical activity and sedentary behaviour in older adults? *International Journal of Behavioral Nutrition and Physical Activity*, 8 (1), 62. doi: 10.1186/1479-5868-8-62 [PubMed: 21679426]
- Janelle DG, Goodchild MF, and Klinkenberg B, 1988 Space-time diaries and travel characteristics for different levels of respondent aggregation. *Environment and Planning A*, 20 (7), 891–906.
- Kerr J, Duncan S, and Schipperijn J, 2011 Using global positioning systems in health research: a practical approach to data collection and processing. *American Journal of Preventive Medicine*, 41 (5), 532–540. doi:10.1016/j.amepre.2011.07.017 [PubMed: 22011426]
- Kitterød RH, 2001 Does the recording of parallel activities in time use diaries affect the way people report their main activities? *Social Indicators Research*, 56 (2), 145–178. doi:10.1023/A: 1012289811886
- Knowles DL, Stanley KG, and Osgood ND, 2014 A field-validated architecture for the collection of health-relevant behavioural data. In: 2014 IEEE International Conference on Healthcare Informatics (ICHI), September 2014, Verona, Italy, 79–88.
- Mazey ME, 1981 The effect of a physio-political barrier upon urban activity space. *The Ohio Journal of Science*, 81 (56), 212–217.
- Paul T, et al., 2016 Scaling behavior of human mobility distributions. In: 9th International Conference, GIScience 2016, Montreal, QC, Canada, 27–30 9, 2016 Springer, 145–159.
- Poletto C, Tizzoni M, and Colizza V, 2012 Heterogeneous length of stay of hosts' movements and spatial epidemic spread. arXiv preprint arXiv:1207.4746.
- Rhee I, et al. 2011 On the levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking (TON)*, 19 (3), 630–643. doi:10.1109/TNET.2011.2120618
- Schipperijn J, et al., 2014 Dynamic accuracy of GPS receivers for use in health research: a novel method to assess GPS accuracy in real-world settings. *Frontiers in public health*, 2, 21. [PubMed: 24653984]
- Shen L and Stopher PR, 2014 Review of GPS travel survey and GPS data-processing methods. *Transport Reviews*, 34 (3), 316–334. doi:10.1080/01441647.2014.903530
- Song C, et al. 2010a Modelling the scaling properties of human mobility. *Nature Physics*, 6 (10), 818–823. doi:10.1038/nphys1760
- Song C, et al. 2010b Limits of predictability in human mobility. *Science*, 327 (5968), 1018–1021. doi: 10.1126/science.1177170 [PubMed: 20167789]
- Stanley K, et al., 2016 Opportunistic natural experiments using digital telemetry: a transit disruption case study. *International Journal of Geographical Information Science*, 30 (9), 1853–1872.

- Steinle S, Reis S, and Sabel CE, 2013 Quantifying human exposure to air pollution—moving from static monitoring to spatio-temporally resolved personal exposure assessment. *Science of the Total Environment*, 443, 184–193. doi:10.1016/j.scitotenv.2012.10.098 [PubMed: 23183229]
- Tobler WR, 1979 Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74 (367), 519–530. doi:10.1080/01621459.1979.10481647 [PubMed: 12310706]
- Tudor-Locke C, et al. 2005 How many days of pedometer monitoring predict weekly physical activity in adults? *Preventive Medicine*, 40 (3), 293–298. doi:10.1016/j.ypmed.2004.06.003 [PubMed: 15533542]
- Vazquez-Prokopec GM, et al. 2013 Using GPS technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment. *PloS One*, 8 (4), e58802. doi:10.1371/journal.pone.0058802 [PubMed: 23577059]
- Wesolowski A, et al. 2012 Quantifying the impact of human mobility on malaria. *Science*, 338 (6104), 267–270. doi:10.1126/science.1223467 [PubMed: 23066082]
- Wilkinson L and Friendly M, 2009 The history of the cluster heat map. *The American Statistician*, 63 (2), 179–184. doi:10.1198/tas.2009.0033
- Yoo E, et al. 2015 Geospatial estimation of individual exposure to air pollutants: moving from static monitoring to activity-based dynamic exposure assessment. *Annals of the Association of American Geographers*, 105 (5), 915–926. doi:10.1080/00045608.2015.1054253
- Yuan Y and Raubal M, 2016 Analyzing the distribution of human activity space from mobile phone usage: an individual and urban-oriented study. *International Journal of Geographical Information Science*, 30 (8), 1594–1621. doi:10.1080/13658816.2016.1143555

Figure 1.
Schematic of a spatial histogram at two time periods. The addition of observations at a later time (right) alter the cumulative frequency of location visits.

Figure 2.
Demographic data of participants.

Figure 3. Spatial histograms for ten participants as heatmaps on two consecutive days for the first 2 days in the area of the university campus.

Figure 4. Spatial histogram of first day and first 2 days for all participants for each experiment.

Figure 5. KL divergence (right) and its discrete derivative (left) for the SHED7 dataset at grid resolutions of 100 m (top), 400 m (mid), and 1600 (bottom) m.

Figure 6. KL divergence (right) and its discrete derivative (left) for the SHED8 dataset at grid resolutions of 100 m (top), 400 m (mid), and 1600 (bottom) m.

Figure 7. KL divergence (right) and its discrete derivative (left) for the SHED9 dataset at grid resolutions of 100 m (top), 400 m (mid), and 1600 (bottom) m.

Table 1.

Total number of primary sensor data records in each dataset.

Dataset	Accelerometer	Battery	Bluetooth	GPS	WiFi
SHED7	98,560,884	376,813	292,020	11,141,544	5,489,682
SHED8	73,510,000	510,863	849,561	12,646,521	7,781,138
SHED9	826,505,638	644,366	780,362	15,878,570	10,620,481

Table 2.

The number of participants and GPS location records in each dataset before and after filtering.

Dataset	Before		After	
	Part.	Loc.	Part.	Loc.
SHED7	61	11,141,544	35	4,765,907
SHED8	74	12,646,521	42	5,204,079
SHED9	88	15,878,570	42	7,001,289

Table 3.

Fit parameters and quality for inverse proportionality.

Dataset	Parameters	100 m	400 m	1600 m
SHED7	$\hat{A}$	0.307	0.206	0.144
	R^2	0.955	0.873	0.879
SHED8	$\hat{A}$	0.253	0.154	0.0738
	R^2	0.943	0.833	0.941
SHED9	$\hat{A}$	0.215	0.128	0.0738
	R^2	0.976	0.960	0.898