Modeling Human Mobility Entropy as a Function of Spatial and Temporal Quantizations

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the degree of Doctor of Philosophy
in the Department of Computer Science
University of Saskatchewan
Saskatoon, SK, Canada

By

Tuhin Paul

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis. Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
110 Science Place
176 Thorvaldson Building
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5C9
Canada

Abstract

The knowledge of human mobility is an integral component of several different branches of research and planning, including delay tolerant network routing, cellular network planning, disease prevention, and urban planning. The uncertainty associated with a person's movement plays a central role in movement predictability studies. The uncertainty can be quantified in a succinct manner using entropy rate, which is based on the information theoretic entropy. The entropy rate is usually calculated from past mobility traces. While the uncertainty, and therefore, the entropy rate depend on the human behavior, the entropy rate is not invariant to spatial resolution and sampling interval employed to collect mobility traces. The entropy rate of a person is a manifestation of the observable features in the person's mobility traces. Like entropy rate, these features are also dependent on spatio-temporal quantization. Different mobility studies are carried out using different spatio-temporal quantization, which can obscure the behavioral differences of the study populations. But these behavioral differences are important for population-specific planning. The goal of dissertation is to develop a theoretical model that will address this shortcoming of mobility studies by separating parameters pertaining to human behavior from the spatial and temporal parameters.

Acknowledgment

I am sincerely grateful to my supervisor Dr. Kevin Gordon Stanley for his erudite supervision, contributions, and support to complete the PhD degree program. I especially thank Dr. Nathaniel Daniel Osgood, who extended his helping hand to complete this thesis. I am grateful to Dr. Scott McKinley Bell, Dr. Derek Eager, and Dr. Dwight Makaroff for their insightful feedback and support. I am thankful to Dr. Nazeem Muhajarine, examining committee chair Dr. Adelaine Leung, and graduate chair Dr. Julita Vassileva for their help and support. I thank external examiner Dr. Alain Barrat for his supportive and encouraging evaluation of the thesis. I thank graduate program assistant Gwen Lancaster for her help and encouragement.

I have always found great inspiration and encouragement in my family. My parents Naresh Chandra Paul and Basanti Rani Paul, wife Mossammat Shakila Akter, sister Pratima Paul, and brother Tushar Paul kept me mentally strong as usual. The birth of our first child Shankar Narayan Paul made the completion of the degree more joyful. I thank my relatives, friends, and labmates for their encouragement.

I acknowledge the Natural Sciences and Engineering Research Council of Canada for providing funding.

Table of Contents

List of Tables

List of Figures

List of Abbreviations & Acronyms

AMD	Advanced Micro Devices
BT	Bluetooth
CCDF	Complementary Cumulative Distribution Function
CDR	Call Detail Record
CLT	Central Limit Theory
CTRW	Continuous Time Random Walk
DTN	Delay/Disruption Tolerant Network
EBR	Encounter Based Routing
GB	Gigabyte
GHz	Gigahertz
GIS	geographic information system
GPS	Global Positioning System
GSM	Global System for Mobile communication
HMM	Hidden Markov Model
ICT	Inter Contact Time
ID	Identifier
KL	Kullback-Leibler
K-NN	K-Nearest Neighbors
LZ	Lempel-Ziv
MANET	Mobile Ad Hoc Network
MGF	Moment Generating Function
MSD	Mean Squared Displacement
MSE	Mean Squared Error
NP-hard	Non-deterministic Polynomial-time Hard

OPF	Optimal Probabilistic Forwarding
PCA	Principal Component Analysis
PDF	Probability Density Function
PER	Predict and Relay
PMF	Probability Mass Function
PROPHET	Probabilistic ROuting Protocol using History of Encounters and Transitivity
RAM	Random Access Memory
RFID	Radio-Frequency IDentification
RoG	Radius of Gyration
RWP	Random Way Point
RW	Random Walk
SHED	Saskatchewan Human Ethology Datasets
SMS	Short Message Service
VANET	Vehicular Ad Hoc Network
WLAN	Wireless Local Area Network
WSN	Wireless Sensor Network

Chapter 1

Introduction

Decision making in many disciplines including urban planning and disease prevention relies on the study of human mobility [1,2]. Such studies often use large volumes of data collected by tracking locations of volunteers, either periodically such as with GPS data loggers, or asynchronously driven by events like diary entries, or phone calls contacting cell towers. Although the voluminous mobility data collected may contain actionable information, those insights are often not easily available upon casual inspection. The study of human mobility aims to uncover inferrable and subtle patterns or characteristics entwined with human mobility data, drive models of human mobility based on these observations, and apply those to real life applications such as modeling the spread of contagion, next location prediction of a person, and urban planning.

Different applications focus on different aspects of human mobility. For example, inter-contact time is an important metric in mobile or delay tolerant network analysis, whereas the span of mobility area and connectedness to other individuals are important elements in monitoring and controlling spread of infectious pathogens [2,3,4]. Different studies have examined human mobility with a focus primarily on particular measures of interest, and proposed mobility models reflecting those observations. Most currently employed metrics of mobility represent people's behavior as distributions over a spatial variable of interest (e.g., trip length). Researchers have defined few characteristics and patterns of human mobility (e.g., distribution of quantifiable aspects like inter-contact time or dwell time) [3,5]. There is a high degree of correlation among these features, confounding possible conclusions. Such distribution parameters can vary due to the behavioral heterogeneity in people, which may arise because of factors like age, gender, or cultural background. However, these parameters

also vary with the underlying spatio-temporal quantization, making it difficult to compare behaviors based on these metrics.

Mobility data sets feature large volumes of data for each user [6]. Many applications do not require fine-grained detail at the individual level; rather they use aggregate metrics. Entropy rate computed from past mobility traces is such a metric, and is widely used for appraising the predictability of a person's movement [7,8]. Entropy rate succinctly represents the extent of periodic behavior in the daily life of an individual, and is particularly important in human mobility because it helps establish an upper bound on movement predictability. Similar to the previously discussed metrics of human mobility, entropy rate is also dependent on the underlying spatio-temporal quantization.

1.1 Motivation

Given the importance of human mobility in many disciplines including computer science, public health, and urban planning, researchers have studied mobility traces to reveal inherent regularities that widen our understanding of human mobility and are useful in practical applications.

Distributions of aspects like travel length, pause duration, and span of travel form the foundation of aforementioned human mobility applications. For example, knowledge of travel span and travel duration may inspire an urban planner to reshape transportation infrastructure to lower the traveling times among popular locations. Quantification and representation of mobility features are, therefore, pivotal in making sense of high fidelity mobility traces.

A typical prior condition to quantifying the distribution parameters of many human mobility features, such as visit frequency and dwell time, is to quantize space and time to describe location. This quantization can be explicit, determined by the experimenter, or implicit, determined by the sensitivity of the apparatus or techniques employed. The underlying spatio-temporal quantization influences the resulting metrics.

Different research studies have explored distributions of these features to represent how people move through and consume space. In the literature, it is common to find different values of distribution parameters for similar studies. While there is a behavioral element in-

fluencing the parameters, the spatio-temporal resolutions also play a key role in determining the parameter values, making it difficult to compare different studies recorded with different methods. To compare two studies on two different populations, researchers need a method to understand the impacts of spatio-temporal quantization and social or human factors on the distribution parameters. Researchers also need to understand the degree to which metrics respond to changes in spatio-temporal quantizations, and the extents and underlying conditions of these dependencies.

The distributions of different mobility features underlie the higher level concepts used in practical applications. One example of this is the predictability of a person. Predictability is an important measure in mobile computer networks, urban planning, and marketing. The metric of predictability, used in the literature, is called the *entropy rate*, which is based on the information theoretic entropy. In information theory, the entropy of a variable specifies the uncertainty associated with the variable. In the study of human mobility traces, each sample is represented as a discrete variable. As the number of variables increase, so does the the associated uncertainty. The entropy rate is dependent on the underlying spatio-temporal resolution used in sampling mobility traces. A model relating entropy rate to the spatio-temporal quantization is yet to be explored. Such a model would help researchers cross-reference studies on human mobility entropy, and allow for principled comparison of predictability absent quantization effects.

While a mobility feature may follow some specific distribution, different mobility features may show different amounts of sensitivity to spatial and temporal scaling. Knowledge of their comparative sensitivity is important to researchers and planners when comparing results of different studies. The human mobility entropy rate, like other mobility features, varies as the underlying spatial quantization and sampling interval change. An entropy rate scaling model would help researchers and planners evaluate predictability studies across populations.

1.2 Problem Statement

It is important to know the ranges within which the mobility features respond to changes in spatial or temporal quantization and whether they do so in a well defined manner. Location-

based technologies have made location-sensing capabilities an integral part of many commodity devices (e.g., smart phones) and such location-enabled devices are now easily accessible to the research community and the general population. Data collected at regular intervals using these devices can enable researchers to measure different aspects of human behavior, such as the distributions of travel and resting times with a high degree of accuracy and precision. Data collection with these devices is generally carried out periodically, where choosing an optimal period and frequency is a research design decision.

Although the data collected with GPS-enabled hand-held devices are normally precise up to a scale of meters, GPS-based mobility studies have limitations. The cost of the study and management complexity go up with the number of people whose mobility traces are collected. Therefore, such GPS devices have not been commonly used to conduct studies on a large scale.As a result, these data do not necessarily represent the overall population. On the other hand, data generated by mobile phone usage logs, as recorded at all phone towers, contain a broader cross-section of the overall population-wide data. However, such data are not as precise as the GPS data, and they are generated according to user schedules, when calls or texts are made or received, and not at regular intervals. As both sources have pros and cons, one important research question is how the location precision and periodicity of data collection impact the interpretation of human behaviors observed in the data, and how different distributions respond to changes in spatial and temporal quantization.

Researchers use aggregate metrics to represent the uncertainty or predictability of the movement related behaviors of a population. In the literature on human mobility studies, the concept of entropy from information theory is widely used as a metric to ascertain predictability or uncertainty of a group of people. However, we should know which mobility behaviors effect entropy-based metrics and to what extent these behaviors shape the metric. Because of the dependence of observable mobility behaviors with spatial and temporal scales, it is expected that the entropy-based metric will depend on them as well. The first step to establish a relationship between the entropy-based metric and the spatial/temporal resolutions is to derive a theoretical model, which may then be evaluated against empirical data.

Differences in spatio-temporal quantization obscure the behavioral differences between

two populations. Systematic spatio-temporal effects might be corrected for with the right theoretical formulation, which may provide a first level of understanding of the results from different studies. However, theoretical models are often based on simplifying assumptions to make the mathematical derivation tractable. While the theoretical models are generally based on researcher intuition and mathematical tractability, empirical data are required to validate a model in practice.

1.3 Contributions

The goal of this dissertation is to determine if a general theoretical model of mobility entropy rate can be derived and whether it can be applied successfully in realistic studies utilizing human mobility traces. Under specific assumptions, this dissertation provides an initial theoretical model, which provides a closed-form expression of the mobility entropy rate as a function of spatio-temporal scaling. Apart from spatio-temporal scaling, the model depends on two path parameters: distance travelled and average speed. Only four terms are required for the model: the length of the path, the average velocity of the agent, the period of the sampling rate, and the width of the square spatial bins. The derivation is based on the Lempel-Ziv (LZ78) compression method [7, 9, 10, 11]. The model provided excellent fits for stylized results within a range of spatial resolutions. The scaling formulation encodes the mobile agent's behavior (through the average velocity and path length) besides the effect of spatio-temporal quantization.

The initial theoretical model of entropy rate scaling demonstrates a sampling rate for the maximal entropy rate. This implies a preferred sampling rate for a given spatial quantization and velocity. This maximal entropy rate may serve as a common comparator between datasets. Researchers designing data collection studies may benefit from this model to identify a preferred sampling period from anticipated average velocity, trip length, and spatial bin size.

The dissertation improves upon the initial theoretical model by analyzing the effects of spatio-temporal quantization on mobility parameters like dwell time and average velocity, which influence the mobility entropy rate. The extended scaling model of mobility entropy

rate relaxes some key assumptions of the initial model, and is applicable to complex physical paths, and the movements of actors having some degree of agency. The final model depends on spatio-temporal quantization and mobility parameters of the actors. The model can provide researchers with insight into how data from two different studies, measured using two different spatio-temporal resolutions, could be meaningfully compared. The model will also help them predict the effects of changing spatio-temporal quantization in their mobility studies, and reach conclusions regarding the relative mobility behaviors of different populations or objects.

With this model, aggregation behavior in sample datasets is modeled, and the following conclusions can be made based on the aggregate metrics.

- Although the type of the distribution was consistent across datasets and resolutions, the parameters describing these distributions varied with spatio-temporal resolution.
- For two datasets, changing resolutions generally does not change the ordering of metrics although the effects of varying spatio-temporal resolution are different on different datasets. Different populations and environments have greater or lesser sensitivity to resampling than others.

The observed value of a predictability metric would be correlated with the underlying mobility features such as dwell time or trip duration distribution. This dissertation shows that mobility entropy rate depends on spatio-temporal scaling as well as path parameters (e.g., movement speed).

1.4 Definitions

The extracted knowledge regarding mobility features from voluminous mobility traces is represented in mathematical form along with corresponding analyses. The data considered in this dissertation are limited to GPS traces that are collected with GPS loggers or smart phones with location sensors. The location data are quantized before analysis, meaning space and time are represented as discrete samples.

To comprehend human mobility as discrete samples, or predictability metrics (e.g., entropy rate), it is important to understand the properties of the canonical density functions representing these behaviors. Many mobility related features are known to follow variants of the *power law distribution*. Common properties of power law distributions, found in the literature, are described in this section.

The relationship between *entropy*, *predictability*, and the *properties of random variables* is central to this thesis and necessary to review. Therefore, entropy and its properties are also discussed in detail. Estimation of entropy rates from human mobility traces is based on the assumption that the trace samples can be represented as a *stationary ergodic process*. Lempel-Ziv algorithm-based methods to estimate entropy rate are also presented. The definitions that follow in this section provide the mathematical background, within the scope of this dissertation, to understand the interaction between human mobility traces and the statistical models used to represent them.

1.4.1 Power-Law Distribution

Many natural and man-made phenomena exhibit well defined regularities in the form of power law distributions. The power law distribution has been used to represent many natural phenomenon - from the frequency of the n^{th} most common word in English, to the frequency of Pluto's crater sizes [12]. Many researchers have established the fact that the power law is prevalent in some features of human mobility traces (e.g., inter-contact time and dwell time distribution [5, 13]).

The probability density function (PDF) of a continuous random variable, x, that follows a power law distribution is defined in (1.1), where $\alpha > 1$ [14]:

$$\begin{aligned} p(x) &= Cx^{-\alpha}\left[x \geq x_{min}\right] \\ &= \frac{\alpha - 1}{x_{min}}\left(\frac{x}{x_{min}}\right)^{-\alpha}\left[x \geq x_{min}\right], \end{aligned} \tag{1.1}$$

$$C = (\alpha - 1)x_{min}^{\alpha-1}. \tag{1.2}$$

In (1.1), C is the normalizing constant, which is defined in (1.2). Truncation of (1.1) at

$x = x_{min}$ is required to make the power law function a probability distribution because the integral of $p(x)$ would be infinite over the range of x without this constraint, where $x \in \mathbb{R}^+$. Therefore, the distribution of (1.2) is also called a truncated power law distribution.

The m^{th} moment, $\langle x^m \rangle$, as defined in (1.3), where the angle brackets indicate the expected value of the variables of the type of the enclosed quantity, of a power law is well defined if $m < (\alpha - 1)$. Therefore, the first $\lfloor \alpha - 1 \rfloor$ moments of a power law distribution exist (i.e., are finite). The practical implication of an nonexistent moment is a growing estimate of the moment, computed over the sample, as the sample size increases. If $\alpha > m+1$, then the m^{th} moment exists, but the $(m+1)^{\text{th}}$ moment does not. If α is slightly larger than $m+1$, the m^{th} moment may converge slowly. A small sample size may not provide a faithful estimate of a moment because convergence on the true value may be slow [15].

$$\begin{aligned} \langle x_m \rangle &= \int_{x_{min}}^{\infty} x^m p(x) dx \\ &= x_{min}^m \left(\frac{\alpha - 1}{\alpha - 1 - m} \right) \qquad for \;\; \alpha > (m+1) \end{aligned} \tag{1.3}$$

The *complementary cumulative distribution function* (CCDF) of $p(x)$, in the form of $Pr[X \geq x]$, is given in (1.4). The exponent term α in (1.1) is required to be > 1 so that (1.4) is a probability distribution function.

$$\bar{P}(x) = Pr[X \geq x] = \left(\frac{x}{x_{min}} \right)^{-(\alpha - 1)} \tag{1.4}$$

Some power law distributed features of human mobility exhibit exponential cutoff at large values of the independent variable (e.g., inter-contact time of people). A power law distribution with exponential cutoff is defined as in (1.5) for $\lambda \geq 0$. At small values of x, (1.5) behaves like a power law but at large values of x, the exponential decay term $e^{-\lambda x}$ overwhelms $p(x)$, and, therefore, $p'(x)$ drops exponentially. For $\lambda = 0$, (1.5) transforms into a pure power law. A distribution $f(x)$ is scale free if $f(bx) = g(b)f(x)$ for any scaling constant b [14]. The power-law distribution is scale free: from (1.1), we find that $p(bx) = b^{-\alpha} C x^{-\alpha} = b^{-\alpha} p(x)$. In a strict sense, unlike the power law distribution, the distribution in (1.5) does not scale, and is not a power law asymptotically. However it captures the power-law effects within a finite range before cutoff, which corresponds to many phenomena in the nature.

$$p'(x) = p(x)e^{-\lambda x} \tag{1.5}$$

Texts that deal with power law distributed phenomena often mention the Pareto distribution and Zipf's Law to describe the distributions of such phenomena. Both the function associated with the PDF of the Pareto Distribution and Zipf's Law are power laws. The plots describing Zipf's Law are also called *Rank/Frequency Plots* [14]. Zipf's Law is described in the scenario when the elements of a sample space are ranked according to their frequencies in a descending order. The frequencies are plotted against the associated ranks [14]. An example of where Zipf's law applies is the ranking of words according to their frequencies in English texts. The following two propositions are equivalent:

- The r^{th} most frequent word has n occurrences. Using Zipf's Law, r is plotted along the X axis, and n is plotted along the Y axis, and $n \sim r^{-b}$.
- There are r words with n or more occurrences. In the context of Pareto distribution, n is plotted on the X axis and r is plotted on the Y axis. The relationship between r and n is presented in this case as (1.6). Because a Pareto distribution is also a power law distribution, comparing (1.6) with (1.4) gives (1.7) [16].

$$r \sim n^{-\frac{1}{b}} \tag{1.6}$$

$$\alpha = 1 + \frac{1}{b} \tag{1.7}$$

Some features of human mobility traces follow fat-tailed or heavy-tailed distributions in the literature [5]. Heavy and fat tails correspond to asymptotic behaviors of the underlying distributions.

A heavy tail of a heavy-tailed distribution is heavier than the exponential distribution (i.e., not exponentially bounded) [17, 18]. Although the right tail is normally of interest, either or both tails may be heavy. The distribution of a random variable X exhibits a heavy right tail if (1.8) is satisfied, where $\bar{F}(x)$ is the CCDF of X [17]. As shown in Fig. 1.1, the

tails of the exponential distributions fall off quickly compared to log normal and power law distributions, which have heavy tails.

$$\lim_{x\to\infty} e^{\lambda x}\bar{F}(x) = \infty \qquad \text{where } \lambda > 0 \tag{1.8}$$

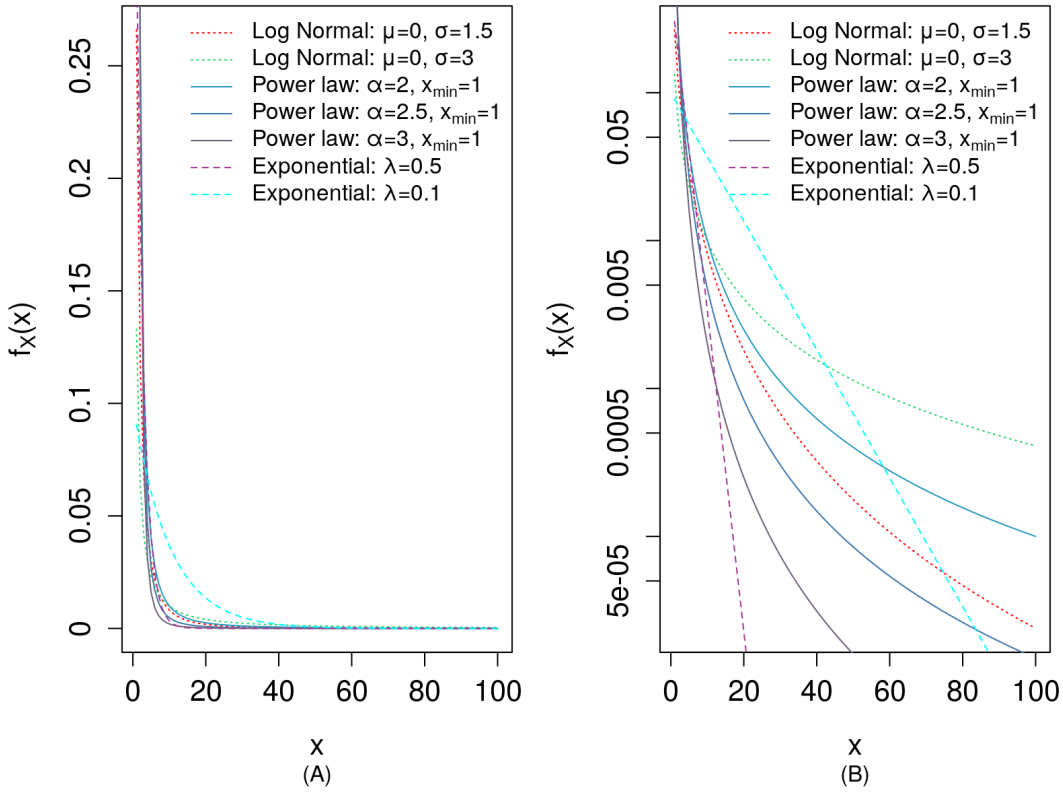


Fig 1.1: Tails of exponential and heavy-tail distributions. The Y-axis has linear and log scale in **(A)** and **(B)** respectively. Y axis limits are different in **(A)** and **(B)** to make tail differences visible.

A heavy-tailed distribution is called a fat-tailed distribution if the CCDF decays as x^{-a} as $x \to \infty$ (e.g., a power-law distribution) [19]. Formally, a fat-tailed distribution is defined as in (1.9), which is equivalent to (1.10).

$$\bar{F}(X) \sim x^{-\alpha} \text{ as } x \to \infty \qquad \text{where } \alpha > 0 \tag{1.9}$$

$$f_X(x) \sim x^{-(1+\alpha)} \text{ as } x \to \infty \qquad \text{where } \alpha > 0 \tag{1.10}$$

From the definition of (1.10), we can see that the power law distribution of (1.1) is an example of the fat tailed distribution.

1.4.2 Stationary Ergodic Process

Researchers have widely used the *entropy rate* metric to study and quantify the predictability of humans from past mobility traces [7, 8, 20, 21, 22]. The formulation primarily used for entropy rate calculation in mobility studies assumes that the underlying traces manifest two properties: stationarity and ergodicity.

Stationary Process A stochastic process, denoted as $\{X_t\}_{t=-\infty}^{t=\infty}$ or $X(t)$ or $\{X_t\}$, generates a sequence of random variables, where the t^{th} variable is represented as X_t, where t is the time index. The process is called stationary if the joint distribution of a sub-sequence $\{X_t\}_{t=t_1+\tau}^{t=t_\lambda+\tau}$ is invariant to the shift (τ) in the time index. That means that the joint distribution of $(X_{t_1}, X_{t_2}, \ldots, X_{t_\lambda})$ will be the same as that of $(X_{t_1+\tau}, X_{t_2+\tau}, \ldots, X_{t_\lambda+\tau})$ for any λ and τ [23, 24].

The stationary stochastic process defined above is also called a *strictly sensed* or *strictly stationary* process. A weaker form of stationary process, called *second-order* or *covariance* or *weak-sense* or *wide-sense* stationary stochastic process is defined as a stochastic process whose first moment and auto-covariance are invariant of the time index t. Several other types of stationarity (e.g., first-order, nth-order) are also similarly defined [25].

Ergodic Process A stochastic process is called ergodic with respect to a statistical property if the time average estimate of the property of a sample or realization, which is a sequence of random variables generated by the process, is the same as the ensemble average, which is the expected value of the random variable generating each sample [26]. A process $X(t)$ may exhibit ergodicity in different statistical properties. For example, if $\hat{\mu}_X$, the estimate of mean from the realization, converges in squared mean to the ensemble average μ_X then $X(t)$ is called *mean ergodic*:

$$\lim_{T\to\infty} \hat{\mu}_X = \lim_{T\to\infty} \frac{1}{T} \int_0^T X(t)dt = \mu_X \tag{1.11}$$

Similarly, $X(t)$ is autocovariance ergodic if $\hat{r}_X(\tau)$, the time average estimate of the autocovariance of a realization converges in squared mean to the ensemble average $r_X(\tau)$:

$$\lim_{T\to\infty} \hat{r}_X(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_0^T [X(t) - \mu_X]\,[X(t+\tau) - \mu_X]\,dt = r_X(\tau) \tag{1.12}$$

Therefore, ergodicity of $X(t)$ implies that the ensemble mean or variance of an ergodic process can be deduced from a sufficiently long sample or realization of the process.

1.5 Entropy

The concepts of probability distribution and uncertainty are closely related. Without delving into the depth of astronomy, the probability that the sun will rise in the east tomorrow morning is 1 from the perspective of an average individual. Therefore, there is no uncertainty in the event. If we toss a fair coin once, there is some uncertainty about what will show up: the probability of either the head or the tail appearing is $\frac{1}{2}$. Now if we consider tossing this coin twice, the sample space is $\Omega = \{\text{HH, HT, TH, TT}\}$. All elements in the sample space are equally likely. Regarding the toss results, it appears that there is twice as much uncertainty associated with two tosses as the uncertainty of just one toss. As the number of outcomes of a probability distribution increases and the probability tends more towards a uniform distribution, the associated uncertainty goes up. The following background study on entropy is primarily adapted from Cover *et al.* [23].

The random variables considered in this thesis are discrete in nature. The probability distribution of a discrete random variable is represented by the *probability mass function* (PMF), which is a function defined on the sample space, and which assigns to each outcome its probability. Given a random variable X with alphabet $\mathcal{X}$ and PMF $p(x)$, the uncertainty associated with it is represented by the metric entropy, as defined in (1.13) [23].

$$\begin{aligned} H(X) &= \sum_{x\in\mathcal{X}} p(x)\log_2\left(\frac{1}{p(x)}\right) \\ &= -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) \end{aligned} \tag{1.13}$$

The unit of entropy depends on the base of the log-term in (1.13). When the base is 2, the unit is *bit*; whereas it is the *nat* for the natural logarithm. The entropy of a variable represents the average uncertainty associated with it. For an impossible event, the convention is to consider $0\log 0 = 0$ because $x\log x \to 0$ as $x \to 0$ [23]. The entropy of X is upper

bounded by $\log_2 |\mathcal{X}|$, as shown in (1.14), which is achieved when the distribution of X is uniform.

$$H(X) \leq |\log_2 \mathcal{X}| \tag{1.14}$$

The entropy of a variable is a lower bound of the number of bits needed to represent the possible values of that variable. As an example, for the single toss of a fair coin, the uncertainty or the entropy is $\left(\frac{1}{2}\log_2 2 + \frac{1}{2}\log_2 2\right) = 1$ bit. If the variable represents the outcome of tosses of two fair coins, the entropy is $4\left(\frac{1}{4}\log_2 4\right) = 2$ bits, which is the minimum number of bits needed to represent four different values. If the coin is biased (e.g., $P(H) = 0.6$), the head is more likely (i.e., uncertainty reduces) and the entropy is $\left(0.4\log_2 \frac{1}{0.4} + 0.6\log_2 \frac{1}{0.6}\right) = 0.97$, which is lower than the maximal entropy achieved by a fair coin, in agreement with (1.14). As an analogy in data compression or network transmission, if all the bits or symbols are the same (e.g., 'X'), the amount of information needed to convey the message is small; for example, 'X, 200 times'. However, for a more diverse string, the amount of information required is more because different symbols and their run-lengths have to be encoded; for example, 'X 50 times, Y 70 times, X 30 times, Z 50 times'.

As we get to know a random variable X (i.e., its distribution), the uncertainty is eliminated or reduced. In other words, after we get to know the previously unknown information (e.g., about a variable), the uncertainty is eliminated/reduced. Information, therefore, is defined as the change in uncertainty as follows.

$$\text{Information} = \text{Previous Uncertainty} - \text{Current Uncertainty} \tag{1.15}$$

Predictability and uncertainty are closely related. The concept of entropy is widely used as a metric of predictability in human mobility studies to represent the predictability in mobility traces precisely [7, 8, 21].

The *conditional entropy* of a discrete random varaible Y, given another discrete random variable X, is given in (1.16) [23]. Conditioning reduces entropy, as shown in (1.17) - equality holds if and only if X and Y are independent.

$$
\begin{aligned}
H(Y|X) &= \sum_{x\in\mathcal{X}} p(x) H(Y|X=x) \\
&= \sum_{x\in\mathcal{X}} p(x) \left[\sum_{y\in\mathcal{Y}} p(y|x) \log_2 \left(\frac{1}{p(y|x)} \right) \right] \\
&= -\sum_{x,y} p(x,y) \log_2 p(y|x)
\end{aligned}
\tag{1.16}
$$

$$
H(Y|X) \leq H(Y) \tag{1.17}
$$

The *joint entropy* of two discrete random variables is shown in (1.18) [23]. The relationship between joint entropy, conditional entropy and individual entropies is shown in (1.19) [23]. The relationship is shown pictorially in Fig. 1.2.

$$
H(X,Y) = -\sum_{x,y} p(x,y) \log_2 p(x,y) \tag{1.18}
$$

$$
\begin{aligned}
H(X,Y) &= H(X) + H(Y|X) \\
&= H(Y) + H(X|Y)
\end{aligned}
\tag{1.19}
$$

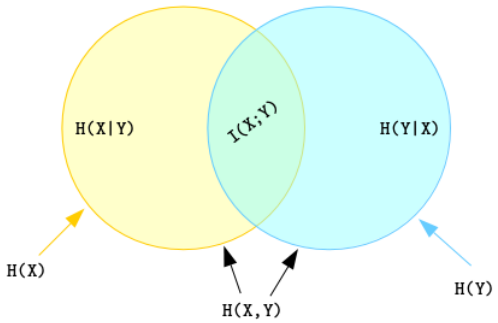


Fig 1.2: Relationship between Individual, Joint, and Conditional Entropies

Relative entropy or Kullback-Leibler Divergence is a measure of distance between two probability distributions, as defined in (1.20). KL divergence measures the error, in terms of entropy, of assuming that the PMF is $q(x)$ when the true PMF is $p(x)$. For brevity,

$D(p(x)||q(x))$ is referred to as $D(p||q)$ [23]. Relative entropy is a non-negative quantity (1.21).

$$\begin{aligned} D(p(x)||q(x)) &= \sum_{x\in\mathcal{X}} p(x)\log_2\left(\frac{1}{q(x)}\right) - \sum_{x\in\mathcal{X}} p(x)\log_2\left(\frac{1}{p(x)}\right) \\ &= \sum_{x\in\mathcal{X}} p(x)\log_2\left(\frac{p(x)}{q(x)}\right) \end{aligned} \tag{1.20}$$

$$D(p||q) \geq 0 \tag{1.21}$$

The mutual information between two random variables X and Y, denoted as $I(X;Y)$, is defined in (1.22) [23]. From Fig. 1.2, we see that conditioning reduces entropy and the reduction is equal to the mutual information of the two random variables. Mutual information $I(X;Y)$ is the amount of reduction in the uncertainty of X because of knowledge about Y, or reduction in the uncertainty of Y because of the knowledge about X. Mutual information is commutative, as shown in (1.23) and mutual information of a variable with itself is the entropy of that variable, as shown in (1.24). Mutual information is also a non-negative quantity (1.25).

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)} \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{1.22}$$

$$I(X;Y) = I(Y;X) \tag{1.23}$$

$$I(X;X) = H(X) \tag{1.24}$$

$$I(X;X) \geq 0 \tag{1.25}$$

The conditional mutual information of X and Y given a third random variable Z is defined in (1.26) and (1.27) [23]. Like mutual information, conditional mutual information is also a non-negative quantity, as shown in (1.28); the equality holds when X and Y are independent given Z.

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= H(Y|Z) - H(Y|X,Z) \end{aligned} \tag{1.26}$$

$$\begin{aligned} I(X;Y|Z) &= \sum_{z} p(z) \left(\sum_{x,y} p(x,y|z) I(X;Y|z) \right) \\ &= \sum_{x,y,z} p(x,y,z) I(X;Y|z) \\ &= \sum_{x,y,z} p(x,y,z) \log_2 \left(\frac{p(x,y|z)}{p(x|z)p(y|z)} \right) \end{aligned} \tag{1.27}$$

$$I(X;Y|Z) \geq 0 \tag{1.28}$$

The joint entropy of more than one random variable is given by (1.29), which is the chain rule for joint entropy [23]. The chain rule for relative entropy is defined in (1.30) [23]. The chain rule for mutual information is defined in (1.31) [23].

$$\begin{aligned} H(X_1, X_2, X_3, \ldots, X_n) &= H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_{n-1}, \ldots, X_1) \\ &= H(X_1) + \sum_{i=2}^{n} H(X_i|X_{i-1}, \ldots, X_1) \end{aligned} \tag{1.29}$$

$$\begin{aligned} D(p(x,y)||q(x,y)) &= D(p(x)||q(x)) + D(p(y|x)||q(y|x)) \\ &= D(p(y)||q(y)) + D(p(x|y)||q(x|y)) \end{aligned} \tag{1.30}$$

$$\begin{aligned} I(X_1, X_2, X_3, \ldots, X_n; Y) &= I(X_1; Y) + I(X_2; Y|X_1) + \ldots + I(X_n; Y|X_{n-1}, \ldots, X_1) \\ &= I(X_1; Y) + \sum_{i=2}^{n} I(X_i; Y|X_{i-1}, \ldots, X_1) \end{aligned} \tag{1.31}$$

Given that n random variables, $X_1, X_2, \ldots, X_n$, have joint PMF $p(x_1, x_2, \ldots, x_n)$, the upper bound on their joint entropy is given in (1.32). This equality holds if and only if the variables are independent of one another.

$$H(X_1, X_2, X_3, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i) \tag{1.32}$$

Entropy quantifies the uncertainty of a single variable. A stochastic process, on the other hand, is represented by a time series of random variables [23, 24]. As the number of random variables n in a sequence (e.g., stochastic process) grows, the *entropy rate* or *per symbol entropy* gives an estimate of how the entropy of the sequence, $H(X_1, X_2, \ldots, X_n)$, grows with n. Entropy rate is defined in (1.33), given the limit exists and all random variables have the same alphabet $\mathcal{X}$ [21, 23].

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) \tag{1.33}$$

If the variables in a growing sequence are independent and identically distributed, then entropy rate is the same as the entropy of any individual variable:

$$\begin{aligned} H(\mathcal{X}) &= \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) \\ &= \lim_{n\to\infty} \frac{nH(X_1)}{n} = H(X_1) \end{aligned}$$

If, however, the variables are independent but not identical, the entropy rate is as follows:

$$\begin{aligned} H(\mathcal{X}) &= \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) \\ &= \lim_{n\to\infty} \frac{\sum_i H(X_i)}{n}. \end{aligned}$$

Since conditioning reduces entropy, $H(X_n|X_{n-1}, \ldots, X_1) \leq H(X_n|X_{n-1}, \ldots, X_2)$. If $\{X_i\}$ is a stationary process, $H(X_n|X_{n-1}, \ldots, X_2) = H(X_{n-1}|X_{n-2}, \ldots, X_1)$. Therefore, for a stationary stochastic process, $H(X_n|X_{n-1}, \ldots, X_1) \leq H(X_{n-1}|X_{n-2}, \ldots, X_1)$, which means that $H(X_n|X_{n-1}, \ldots, X_1)$ is a non-negative decreasing quantity, Thus, it has a limit:

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, \ldots, X_2, X_1). \tag{1.34}$$

The entropy rate for a stationary stochastic process is related to the conditional entropy of the last variable given the earlier ones. Entropy rate can be expanded using the chain rule for joint entropy as shown in (1.35). The conditional entropies in 1.35 have a limit of $H'(\mathcal{X})$ as shown in (1.34). Therefore, we can apply the limit of *Cesàro mean* [27], which states that $\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} a_i\right) \to a$ if $a_n \to a$, to the summation in (1.35), to find that $H(\mathcal{X}) = H'(\mathcal{X})$:

$$\begin{aligned} H(\mathcal{X}) &= \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) \\ &= H'(\mathcal{X}). \end{aligned} \tag{1.35}$$

$H(X_n|X_{n-1}, \ldots, X_2, X_1)$, therefore, converges monotonically to $H(\mathcal{X})$ from above.

1.6 Entropy Rate Estimate

Given a stochastic process, as the number of time steps grows, the number of possible combinations of outputs from those time steps grow exponentially. But in practical applications, the available data (e.g., human mobility traces) are insufficient to precisely calculate the probabilities of these combinations. Therefore, approximation methods are applied in practice to find an estimate of the entropy rate.

Given the joint PMF distribution of the realizations $\{X_t\}_{\tau+1}^{\tau+n}$ of a stochastic process, the *block entropy* is defined as:

$$H_{\tau,n} = -\sum_{x_{\tau+1}, x_{\tau+2}, \ldots, x_{\tau+n}} p(x_{\tau+1}, x_{\tau+2}, \ldots, x_{\tau+n}) \log p(x_{\tau+1}, x_{\tau+2}, \ldots, x_{\tau+n}). \tag{1.36}$$

As the blocks can be of arbitrary length, concepts like differential entropy or entropy rate become more relevant than Shannon's entropy. The differential entropies [9] for the process used in (1.36) are the following:

$$h_{\tau,n} = H_{\tau,n} - H_{\tau,n-1}. \tag{1.37}$$

Differential entropy measures the new information introduced by the n^{th} outcome, having known the preceding $(n-1)$ outcomes. If we consider a stationary process, we can ignore

τ to make the mathematical expressions easier to follow. Schürmann *et al.* [9] shows that differential entropy can also be expressed as in the following equation:

$$\begin{aligned} h_n &= -\sum_{x_1,x_2,\ldots,x_n} p(x_1, x_2, \ldots, x_n) \log p(x_n|x_1, x_2, \ldots, x_{n-1}) \\ &= H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1). \end{aligned} \tag{1.38}$$

To consider all correlations and constraints in the realizations of the process, the average amount of information per symbol (entropy rate) is defined in the limiting case as follows:

$$h = \lim_{n\to\infty} h_n. \tag{1.39}$$

In (1.39), h_n converges monotonically to h from above for a stationary stochastic process. For a stationary stochastic process, $p(x_1, x_2, \ldots, x_n)$ can be computed from a finite sequence of length N, where $N > n$ [9]. However, with increasing n, the number of combinations of $(x_1, x_2, \ldots, x_n)$ in (1.38) increases exponentially and so does the minimum value of N to faithfully compute probabilities $p(x_1, x_2, \ldots, x_n)$. This renders (1.38) unrealistic as a practical means for computing the metric. Therefore, researchers have proposed alternative methods to approximate the entropy rate from finite length symbol sequences. The estimator of (1.40), which is based on the LZ compression algorithm, converges faster, and has been widely used to approximate the entropy rate of human mobility traces [7,8,28]. The estimator (1.40) gives the entropy rate of a string S of length L as $L \to \infty$, where i is the index of a character in the string (with the first character being at $i = 0$), and Λ_i is the length of the minimum substring beginning at i such that this substring has not previously been observed in the prefix of S terminating at position $(i - 1)$.

$$H = \left(\frac{1}{L}\sum_{i=0}^{L-1} \Lambda_i\right)^{-1} \log(L) \tag{1.40}$$

The LZ compression algorithm parses a symbol sequence $\{x_k\}_{k=1}^{n}$ into words such that the next word to parse is the shortest sequence that was not seen before. The words are encoded as (j, c) pairs, where c is the last character of the next word to encode and j is the codeword index of the word that corresponds to all but the last character of the next word to encode. The first symbol constitutes the first word and is encoded as $(0, x_1)$.

As an example of the calculation of entropy rate using (1.40), consider a string 'ABABAAAC', where each character represents an unique location sample. Table 1.1 shows the Λ_is calculated at each zero-based index i. The estimate of entropy rate for the string, according to (1.40) and the findings in Table 1.1, is $\left(\frac{15}{8}\right)^{-1} \log_2(8) = 1.6$ bits for base-2 logarithm.

Table 1.1: Demonstration of entropy rate estimation

Index (i)	Shortest Sequence (Highlighted)	Λ_i	$\sum \Lambda_i$
0	A	1	1
1	AB	1	2
2	ABABA	3	5
3	ABABAA	3	8
4	ABABAAA	2	10
5	ABABAAAC	2	12
6	ABABAAAC	2	14
7	ABABAAAC	1	15

1.7 Dissertation Overview

This thesis is manuscript-styled. The three papers related to this thesis comprise the main thesis body (Chapter 3 through Chapter 5). A summary of the chapter contents follows:

Chapter 1: Introduction

Introduces the field, frames the problem, and provides an overview of the contributions. Mathematical background and definitions required to understand the dissertation are provided.

Chapter 2: Literature Review

This chapter discusses the developments in the areas related to the problem defined in the Problem Statement (Section 1.2). The importance of human mobility disciplines like computer science, geography, and public health are discussed. Meaningful utilization of human

mobility in these disciplines depends on properly understanding mobility features, which are discussed in this chapter. Observable mobility features depend on quantization parameters and algorithmic mediation. With this in mind, typical mobility models and features in the literature are discussed. Predicting the next location of a person or vehicle, and overall predictability are important research areas in computer science. These are discussed in light of the available literature.

Chapter 3: Manuscript 1

Citation: Paul T, Stanley K, Osgood N, Bell S, Muhajarine N. Scaling Behavior of Human Mobility Distributions. In: International Conference on Geographic Information Science; 2016. p. 145–159. Springer International Publishing.

Large data sets comprising GPS traces and periodic or event-driven samples of the activities of tens to hundreds of thousands of people are comprised of millions of records, whose interpretation requires data analysis expertise and tools. Because of the scope of the data, it is common to express analytical results as aggregate distributions of parameters of interest. This paper shows the effects of spatio-temporal resolution of data collection on some canonical features. The distributions of these features were found to respond to rebinning. This paper presents what types of relationships we can expect from these features, and gives a comparative presentation of the suitability of these features to compare human behavior across data sets. In discussion with my supervisor, I came up with the set of parameters to analyze. I contributed to writing and synthesis.

Chapter 4: Manuscript 2

Citation: Osgood ND, Paul T, Stanley KG, Qian W. A Theoretical Basis for Entropy-Scaling Effects in Human Mobility Patterns. PLoS ONE. 2016;11(8):1–21.

The entropy rate of the string representing visited locations is widely accepted as a succinct metric to represent periodicity of movement patterns. However, mobility entropy rate is not invariant under changes in spatial or temporal scale. This limits the utility of this metric by confounding inter-experimental comparisons. This paper leverages the Lempel

Ziv (LZ78) compression process to derive a scaling relationship for mobility entropy rate of non-repeating straight line paths. The mobility entropy rate attains its maximum value at a particular sampling rate, indicating the existence of an optimal sampling rate for particular movement patterns. Under certain conditions, the formulation, presented in this paper, can predict the scaling behavior of simulated mobility traces indicating its potential utility as a tool on empirical traces. Dr. Osgood came up with the key insight that the model was synthesizable. We all contributed to developing the model in different ways. I did the simulations. We all contributed to analysis and writing.

Chapter 5: Manuscript 3

Citation: Paul T, Stanley, K, Osgood, N. Multiscale Entropy Rate Analysis of Complex Mobile Agents. Submitted to Science. January, 2017.

Technological innovations have allowed researchers to probe human spatio-temporal behavior at unprecedented scales. It is now straightforward to obtain detailed mobility traces of individuals on a meter-by-meter and second-by-second basis. However, the diversity of information generated makes direct comparisons between individuals or populations difficult. Metrics which preserve characteristics of mobility in a more tractable form are desirable. One such metric, which has received significant attention, is mobility entropy rate or the average information content in an person's path through space. Although mobility entropy rate provides a compact representation of spatio-temporal data, the metric is sensitive to the spatial and temporal scales at which the data were acquired. The paper provides a general mobility model that is applicable to movements of objects having some degree of agency. The model depends on spatio-temporal scales and mobility parameters of the actors. The paper shows that analyzing the effects of spatio-temporal scaling on mobility entropy rate can reveal interesting features of population behaviors. The model is validated with six datasets containing movement traces from a variety of sources: humans, moose, seabirds, and ocean drifters. This scaling relationship can be used to compare mobility behaviors of populations or actors observed at different spatio-temporal resolutions, and to provide insight into the desing of mobility studies. I did the analysis along with Dr. Stanley, and derived the model

equation. Dr. Stanley and I contributed to the writing of the main paper. I performed data processing, and wrote the supplementary material with edits from Dr. Stanley and Dr. Osgood. We all contributed to the analysis.

Chapter 6: Discussion and Conclusion

The contributions are summarized in light of the *Problem Statement*. The chapter includes a discussion on potential future work, and our overall conclusions.

1.8 Summary

Reasoning about human mobility is an important part of decision making processes in many disciplines. Such disciplines would benefit from the ability to properly interpret mobility features along with the underlying patterns, and envision how to apply the findings of a study to different environments. The dissertation identifies the features which demonstrate orderly response to sampling conditions. Distribution parameters of these features may be used to make sense of a study when the findings are projected to a different dataset or environment. Mobility entropy rate is a potential tool to quantify predictability/uncertainty associated with human movements. The initial entropy rate scaling model developed in this dissertation provides a theoretical background to understand how the metric varies with spatio-temporal quantization and path-specific mobility parameters. The primary contribution of this dissertation is to develop a general scaling model, which makes it feasible to compare independent studies on human mobility. The model is applicable not only to human movement, but also to that of any object demonstrating a degree of agency. The model was validated with empirical mobility datasets of a variety of actors.

Chapter 2

Related Literature

There is some degree of uncertainty in every person's mobility. Information theoretic entropy rate is a widely used metric to quantify uncertainty associated with stochastic variables [7, 8, 21, 29, 30]. As human mobility history is a sequence with regular repetitive behavior, it can be represented as a stochastic process, and entropy rate could be a suitable metric to express the uncertainty or predictability associated with the movement [8, 30]. Researchers have proposed different techniques to quickly estimate the entropy rate from human mobility traces [8,28]. This metric has been used to measure the upper limit of predictability of human mobility [7,21]. However, the calculated entropy depends on the underlying spatio-temporal resolution of sampling the locations of the person [21, 22].

The study of human mobility has flourished as an independent research area because it is required to correctly frame solutions to research problems in various disciplines including computer science, epidemiology, public health, urban planning, and geography. Interactions between people and locations or transportation networks play an important role in urban planning. Knowledge of human mobility helps determine the spatial distributions of locations of activities: a fundamental problem in spatial economics and geography [31]. Spatial models also help describe traffic flow within and among transport networks, and predict the future location of a vehicle or individual. Mobility studies also facilitate the implementation of recommender systems in geomarketing by analyzing the dominant flows found in individual mobility traces for the region of interest [31, 32, 33]. Human mobility has important implications in modelling the spread of infectious diseases by analyzing the structure of contact patterns. Understanding these contact patterns may help control or prevent the spread of infectious diseases [4, 34], and build better models of contagious diseases [35, 36].

Given the importance of human mobility, identifying typical characteristics of mobility traces is the key to making human mobility useful. Researchers have methodically proven that human movements are predictable up to reasonable spatial and temporal granularity [3, 7, 13, 21, 22].

The availability of inexpensive motion sensors and GPS trackers in devices like smartphones has greatly facilitated collecting large datasets of human movement traces and associated data (e.g., speed, accelerometry data) [6, 31]. These datasets may be enriched by including the information of visited WiFi routers, or Bluetooth devices by the smart phone carriers, that help approximate indoor locations [35,36]. Whereas data collection with smart phones requires some degree of direct interaction with the participants, location data generated by mobile phone base stations, from call or SMS records, have been used by many researchers [5,6,7,37,38,39,40]. Although the locations in cell tower data are imprecise, they are voluminous and provide the mobility information of a large number of people. However, data from proprietary sources such as social media organizations and cellular network operators are subject to proprietary limits and undisclosed mediation [6], which may effect the conclusions made from the analysis of the data.

Given a dataset, the next important step is to analyze the statistical properties therein as they correlate with the regularity and predictability of movements. Research on human mobility analysis has shown that human mobility shows statistical regularities in terms of some well defined properties such as dwell time distribution [5, 13], or temporal patterns in cell phone usage behaviors [41]. Previous studies on mobility data reveal interesting information about general activities carried out by different groups of people (e.g., younger people and teens use SMS more than voice calls, and commuters text/talk on phone more while on commute) [7, 41]. Differences in movement patterns of different groups may be explained with a mobility model.

Although there are many different features showing statistical regularities in human mobility traces, some of them exhibit strong correlations with others [40]. Using *principal component analysis* (PCA), Csaji *et el.* achieved significant dimensionality reduction without substantial loss of information because of the correlation of some location and movement related measurements with others [40]. Features extracted from analyzing mobility traces

are taken into account to build mobility models for the applications of interest. These synthetic models emulate the statistical characteristics observed in real human trajectories. As underlying datasets and environments of different models vary, so do model parameters. The models do not generally apply to movements outside the environments considered to develop them [31, 42].

Research studies conducted on datasets are affected by the spatio-temporal resolution of data collection, making inter-study or inter-dataset comparison implausible. Moreover, as device or system capabilities vary, the minimum resolution at which data can be analyzed may be limited. As an example, the communication ranges of different location tracking systems are shown in Table 2.1 [43, 44]. As the signal transmission range increases, the precision of collected data drops. Cell towers are more densely located in urban areas than in the countryside. The precision of cell tower location records was reported to be about one square mile [7,45]. Average position errors of $2-15\ m$ have been reported [46]. The significant variation in the precisions of different data sources may result in different conclusions from similar studies.

Table 2.1: Communication ranges of location tracking technologies

Communication System	**Range**
Cell Tower	up to 35 Km [43]
GPS	N/A
WiFi	100m
Bluetooth	10m

2.1 Applications of Human Mobility Modelling in Computer Science: MANETs, VANETs, and DTNs

Mobility entropy rate is a measure of predictability, which has implications in designing routing protocols for networks where configurations change with human movements. Human mobility has noteworthy importance in computer science, especially networking. Hu-

man mobility models have been studied for routing algorithms of Mobile Ad-hoc Networks (MANETs), Vehicular Ad-hoc Networks (VANETs), and Delay Tolerant Networks (DTNs).

Considering human mobility in MANETs is important in disaster management scenarios after natural catastrophes like hurricanes, forest fires, or earthquakes. Such catastrophes are common and they are known to have struck down mobile communication networks for days to weeks even in technologically advanced countries [47, 48]. The 140 mph winds[1] and torrential rain caused great devastation including power outrage and collapse of mobile networks in September 2005 in the USA. Survivors who trying to call for help and medical assistance found themselves disconnected from the cellular services.[2] Even three weeks after the catastrophe, over 60% of the networks were still non-functional [47]. Similarly, after the Chi-Chi earthquake in Taiwan in 1999, it took Chunghwa Telecom, the largest telecom operator in Taiwan at that time, 15 days of 24x7 work to restore its networks [48].

There are many cases of such catastrophes. The disaster area may be large, encompassing the incident site, transport zone, casualty treatment area and hospitals, and it requires a large rescue team (e.g., 150-200 rescue units in the cases of train accident and roller coaster fire to care for affected lives and restore the infrastructure [49]. There are several reasons for communication failure after a disaster. Some reasons are ruptured power lines by broken bridges and roads, failure of backup power generators, base station failure, unavailability of electricity to charge cell phones, and to operate cooling systems for critical equipment, and communications traffic jams [48]. Ad-hoc networks, therefore, have been studied for use in disaster management [47, 48, 49, 50]. Human mobility models are employed to make routing decisions. Movement predictability may be useful in coordinating rescue tasks in the disaster sites. Better mobility models would enhance routing performance in MANETs [48,49,51,52].

Vehicular Ad hoc Networks (VANETs), which are communication networks among vehicles on the road, are envisioned to facilitate trip optimization, deliver advertisements, or communicate safety information and entertainment data [53,54,55]. VANETs have been proposed to provide information on traffic conditions, bad weather, and emergency situations (e.g., traffic accidents) while on the road [56]. VANETs are implemented using vehicle-to-

[1] http://www.discovery.org/a/2881

[2] http://www.washingtonpost.com/wp-dyn/content/article/2005/09/14/AR2005091402262.html

vehicle and vehicle-to-infrastructure communication. Like in MANETs, mobility features and models (especially, vehicular) play an important role in evaluation and determination of communication protocols [56]. For example, *Hou et al.* showed that beyond a threshold speed, the decline in connectivity in VANETs follow a power law for varying speed [57].

The salient features of human mobility are useful in modeling co-locations of humans, and have been used in Delay Tolerant Network (DTN) routing. A DTN is characterized by the possible lack of a complete end-to-end to path at any moment between the source and the destination. The communicating entities are normally mobile and sparsely distributed. There may be long delays before a pair of entities comes into contact and within two consecutive contact periods, an entity may be completely isolated for a long time. Therefore, conventional routing algorithms of MANETs or WSNs are not useful in DTNs. Routing in DTNs relies on greedy techniques of forwarding messages opportunisticly as entities encounter one another. An important aspect of DTN routing is to predict future contact patterns. Therefore, for DTN networks having humans in the loop, incorporating innate human movement patterns seems an attractive source of routing heuristics.

Routing in DTNs has progressed from earlier flooding-based techniques to more sophisticated methods that include other aspects of mobility. Earlier algorithms used flooding or variants (e.g., *Epidemic Routing* [58], *Spray and Wait* [59]). Some algorithms use different measures and metrics to determine which nodes are more likely to send the data to the destination and forward to them (e.g., *FRESH* [60]). A vast majority of DTN routing protocols apply heuristics because optimal routing in the general DTN case is NP-hard [61, 62].

Some DTN routing algorithms use the history of encounters between nodes for decisions on routing and buffer management [63, 64]. To eliminate replications that are unlikely to reach the destination, the Probabilistic Routing Protocol using History of Encounters and Transitivity (PROPHET) [65] protocol exploits the non-randomness of real-world encounters whereas Epidemic routing implicitly assumes that encounters were random [58]. Some DTN routing algorithms proposed mobile devices that form a moving backbone to facilitate data transmission [66, 67, 68, 69]. Optimal probabilistic forwarding (OPF) [70] assumes long term regular patterns of node mobility and that each node knows the inter-contact times of all pairs of nodes in the network. The forwarding decision depends on whether or not forwarding

will increase the overall delivery probability.

Forwarding decisions in *predict and relay* (PER) [71] depend on the probability distribution of future contact times of humans and aim at improving the delivery probability. Forwarding decisions leverage the deterministic nature of human trajectories, based on mobility history, and landmarks commonly encountered by humans. Similarly, Encounter Based Routing (EBR) [72] prioritizes nodes that encountered a lot of nodes in the past because such a node is more likely to pass the message to the final destination. By analyzing human mobility traces, *Rasul et al.* leveraged the social diversity of people to improve message delivery latency in a network of intermittently connected low-power devices [73, 74].

2.2 Data Collection Strategies

As mobility data is fundamental to all the research work in modeling human movement, the quality and quantity of such data play important roles in advancing human mobility research. Before the advent of the technologies that made collecting big data a reality, researchers used traditional travel diary surveys or using custom surveying instruments for mobility studies [6]. The data sets were small to moderate and it was costly and time-consuming to collect and process survey questionnaires [6]. The questions were designed to answer specific questions about human activities and the patterns that resulted from subconscious behaviors were not reliably reflected in those data sets. However, in recent years, location-aware technologies (e.g., smart phones), cell network communication data, public transit data, and social networks have made massive data collection possible from tens to millions of people.

GPS records are a popular source of data. Some common sources of GPS records are GPS on vehicles [75], public transit buses equipped with smart card readers and GPS [76,77], GPS-enabled mobile phones, and GPS data loggers [35,36,78]. Research studies are often based on data collected from university students and staff who are given GPS-enabled mobile phones [35,79]. The mobile phones may also collect other types data that may help improve location accuracy,and may provide insight into social life and patterns of the study participants. Although these data contained fine grained, and frequently sampled (in the order of a second

to few minutes) location information, they lack demographic diversity because the studies are generally conducted on a small to medium sample of a particular group (e.g., students).

Location data from a large number of users can be collected from mobile cell tower records [5, 7, 37]. These data pertain to the communication details (e.g., mobile phone call and SMS) of the cell phone carriers and may accommodate such data from millions of users randomly distributed in the population. Although these data sets may represent the population better, they suffer from the lack of precision in both spatial and temporal dimensions. Mobile cells are large, a reflection of the long range of cell towers, as shown in Table 2.1. Therefore, movements inside a cell may not be differentiated. Moreover, ping-pong effects between adjacent cell towers introduce impurities in the locations and travel-related data [6]. The data are recorded only when the user is active on the mobile phone (e.g., calling or texting or using mobile data). Therefore, such data may under-represent activities of some people.

In recent years, social media data have been used as another source of human mobility traces [6,80,81,82]. Some social networking web applications let users share their current location by virtual checking-in to the place they visit. The data generated from such check-ins are location specific and sporadic in nature [81]. Although sporadic, such data may complement other more detailed data in human mobility studies. The check-ins have implications in understanding traveling behaviors and relationship between travelling and friendship or social ties.

The quality of data effects the results of mobility studies [6]. Data collection in this dissertation is limited to GPS traces collected using GPS loggers and mobile phones because they allow to evaluate the impacts of spatial and temporal scales on the findings on human mobility behaviors, without the sampling confounds found in all tower or check-in-based data sets.

2.3 Statistical Properties of Human Mobility

The mobility entropy rate model developed in this dissertation is closely related to mobility predictability, which again is correlated with the statical properties of mobility traces. Many

researchers noticed regular patterns in human mobility [5, 7, 13, 30]. They looked at the statistical properties present in the mobility traces, and how they correspond to individual mobility [83,84,85], spread of daily movements of different individuals, geographic influences on movements due to factors like transport infrastructure, residential and work locations [31, 86, 87], predictability of future locations [5, 21], and location prediction algorithms [31, 84,88,89,90]. Although mobility patterns vary from person to person, mobility traces exhibit regularity in some properties [13, 31, 91, 92]. Some features with well defined statistical regularities [13, 91] are the following:

Flight length: A flight is defined as a straight line movement without pause or change of direction [13, 84, 92, 93]. It has been shown that flight lengths follow truncated power law distrubutions [13, 92, 93]. Brockmann *et al.* observed the scale free distribution given in (2.1) [93]. Song *et al.* observed cutoff after a characteristic length, as shown in Table 2.2, and reported $\alpha = .55 \pm .05$ (mean $\pm$ standard deviation) [5, 94]. Gonzalez *et al.* gave an alternative form of flight length distribution (2.2), by incorporating the cutoff into the heavy-tailed distribution function [84]. In (2.2), $\alpha = .75 \pm .15$ (mean $\pm$ standard deviation), $\Delta r_0 = 1.5$ Km, and κ takes values of 400 Km and 80 Km for two different datasets.

$$p(\Delta r) \sim |\Delta r|^{-(1+\alpha)} \quad \text{where } \alpha < 2 \tag{2.1}$$

$$p(\Delta r) = (\Delta r + \Delta r_0)^{-(1+\alpha)} \, e^{\left(\frac{-\Delta r}{\kappa}\right)} \tag{2.2}$$

Pause time: Pause time refers to the interval between two flights. Pause times (also called dwell times) are known to follow truncated power law distrubutions [5, 13, 92, 95] as shown in (2.3). *Song et al.* reported $\beta = .8 \pm .1$ and cutoff at 17 hours, as shown in Table 2.2 [5].

$$p(\Delta t) \sim |\Delta t|^{-(1+\beta)} \tag{2.3}$$

Inter-contact time: Inter contact time (ICT) refers to the interval between two successive contacts of two persons. ICTs follow a truncated power law up to a threshold time after which it shows exponential decay [13, 91, 96, 97, 98].

Bounds of mobility area: People tend to travel to nearby locations and movements of most people are confined to short distances [3, 13, 31, 99]. However, long trips occasionally occur and some people regularly move long distances [5,31]. Although the ranges of mobility area differ greatly from person to person, the spatial distribution of travel patterns are found to follow reproducible patterns [83,84]. Song *et al.* proposed that *Radius of Gyration (RoG)* represents spans of individual mobility [5, 7, 20]. RoG, expressed as r_g, is defined in (2.4), where c is the center of the polygon bound by spatial resolution-dependent coordinates $\{r_i : i \in \mathcal{N}^+ \wedge i \leq N\}$ of samples from past location traces. RoG distribution describes how compact the areas traversed by participants are. Song *et al.* show that *RoG* follows a fat-tailed distribution [5]. Gonzalez *et al.* approximated RoG with the truncated power law equation in (2.5) [84], where $r_g^0 = 5.8$ km, $\beta_r = .65 \pm .15$ (mean $\pm$ standard deviation), and $\kappa = 350$ km.

$$r_g = \sqrt{\frac{1}{N}\sum_{i}^{N}(r_i - c)^2} \tag{2.4}$$

$$P(r_g) = \left(r_g + r_g^0\right)^{-(1+\beta_r)} e^{\left(\frac{-r_g}{\kappa}\right)} \tag{2.5}$$

Propensity to visited popular places: People visit a few popular places more often than other places [13, 91]. Waypoints of human trajectories may be modelled as fractal points, indicative of social context or common gathering places of people of shared interests [13, 99, 100, 101]. The distribution of visit frequencies of different places has been shown to follow Zipf's law [5, 84].

2.3.1 Mobility Models and Random Walks

The mobility models in the literature are based on simplistic assumptions or the analysis results of their foundational datasets, which vary widely [40, 102, 103, 104, 105]. Some studies on wireless ad-hoc network routing assumed exponential distribution of *inter contact times (ICT)* of human encounters [102, 102, 103]. Simplistic mobility models such as *Random Way Point (RWP)* or *Random Walk (RW)* models produced exponentially distributed ICTs

[103,106]. In a *Random Walk (RW)* model that has finite variance of flight lengths and finite average step time (flight time + pause time), the displacement from the origin after time t follows a normal distribution, in accordance with the *central limit theory (CLT)*. The width of this distribution varies as $\sqrt{t}$ [3] and the *mean squared displacement (MSD)* from the initial position of the walker grows linearly with time [3]. Human walking patterns are not random in nature [3,99]; it was later found that ICTs in human movement follow a truncated power law distribution with exponential decay after a characteristic time [13, 96, 97].

Moreover, the statistical features inherent in human mobility show that flight lengths follow a truncated power law distribution $p(l) \sim l^{-(1+\alpha)}$ with $0 < \alpha < 2$ [3, 92, 99, 103]. This distribution does not have a finite second moment and, therefore, CLT can not be applied. Such heavy tailed step length distribution is known as a *Levy Flight* [3, 31, 107, 108]. MSD in Levy walk varies as t^{γ} where $\gamma > 1$ (super-diffusion). However, introducing power law pause times to Levy walk (*Levy walk with trapping*) makes the random walk either super-diffusive ($\gamma > 1$) or sub-diffusive ($\gamma < 1$). Rhee *et al.* found that $MSD(t)$, which is MSD after time t, changes from being super-diffusive ($\gamma > 1.2$) to sub-diffusive ($\gamma < .9$) around 30 minutes [3]. They confined location records within 10 km of study areas [3]. This truncation impacts dispersion as time grows and MSD becomes normal or sub-diffusive [3, 92, 109, 110]. Another reason for sub-diffusion is the homecoming tendency of humans [3, 92]. Rhee *et al.* also reported high correlation between speed and flight length: high velocities are associated with longer flight lengths [3].

Although researchers had earlier concluded similarity between human mobility and *Continuous Time Random Walk* (CTRW) [111,112,113], and then *Levy walk* based on the heavy tailed distriburions of flight lengths and pause times [3, 92, 93, 99, 114], it was later shown that many other characteristics of human trajectories are in contradiction with random walk, Levy flight or truncated Levy flight models [5, 115]. Song *et al.* used mobile phone call data of three million users in one study and location records of 1000 users in another study where participants' locations were recorded every hour for two weeks [5]. Their findings on hourly displacement and Dwell Time are shown in Table 2.2. The cutoff for trip length corresponds to a distance humans can reasonably travel in an hour (location samples were taken every hour). The dwell time cutoff generally represents the awake time of an adult human.

Therefore, Levy Walk or Random Walk models of human mobility should not be used for the following reasons:

- As time passes, the number of distinct places visited in human trajectories, whose distribution is shown to be t^{μ} grows slower ($\mu = .6 \pm .2$) than Levy Walk ($\mu = 1$) or CTRW ($\mu = \beta$) models, [114, 116, 117].
- Whereas the CTRW model features that *mean squared displacement* (MSD) from the initial location grows as t^{ν} where $\nu = 2\beta/\alpha$, human trajectories display a slower than logarithmic growth in MSD,
- Humans return home on a regular basis, and
- Levy Walk or CTRW models have a uniform distribution of visit frequencies of locations visited by humans as $t \to \infty$, however, the distribution has been empirically shown to follow Zipf's law [5, 84].

Table 2.2: Distribution of flight length and dwell time of human trajectories [5]

Feature	Distribution	Exponent	Cutoff
Flight Length	$p(\Delta r) \sim \lvert\Delta r\rvert^{-1-\alpha}$	$.55 \pm .05$	100 Km
Dwell Time	$p(\Delta t) \sim \lvert\Delta t\rvert^{-1-\beta}$	$.8 \pm .1$	17 Hours

Lee *et al.* attributed the heavy tailed distribution of flight lengths to the burstiness of locations or the points that humans visit [99]. They grouped the points into clusters or hot spots using the transitive closure of the connected relation. They found that distributions of the sizes of the clusters (up to about 10^4 visit points in a cluster) and inter-cluster distance are heavy-tailed [99]. They concluded that the burstiness of visit points contribute to the heavy tailed distribution of flight lengths in human movements because people tend to visit neighboring visit points before moving to a distant cluster.

Statistical and scaling properties of human mobility from spatial, temporal and contextual or social aspects have been studied at different levels of the spread of human mobility - from global, continental, or country-wide movement to urban or university campus-centric

movement [31, 39, 40, 90, 118]. However, Asgari *et al.* mentions that no comprehensive study on spatial scale incorporated traffic of all the different modes of transportation, which would be a daunting task [31], and would require data collection at international level. Researchers have investigated periodic (from long-term such as monthly or weekly to short-term such as less than hourly) patterns in human mobility traces, and addressed what might be considered as the optimal scale for mobility studies [31, 40, 119, 120, 121]. As temporal scales are made coarser, it should be noted that enough events should be captured within the sampling interval to make the study credible [31, 121]. Moreover, if the sampling interval is not chosen judiciously, strong bias may surface in the study results and conclusion [31].

The existing mobility models are based on the analytical results of the underlying mobility datasets. Datasets vary widely, and the models emulate the distributions found in their foundational datasets or environments [40, 104, 105, 122]. Therefore, these models may not be considered as generic mobility models [31, 42]. Moreover, mobility patterns may vary in different groups of people, where groups are defined based on different aspects (e.g., profession or gender) [31, 45, 86, 115]. Because different patterns exist over the wide population, mobility models try to give a generalized concept of mobility by representing global aggregate behaviors or presenting a collection of patterns [31, 104].

Some researchers have drawn an analogy of Newton's gravity law to the number of individuals moving between two locations [2, 31, 83, 86]. In this gravity model, the number of individuals (T_{ij}) moving from location i to location j per unit time, is expressed in (2.6), where m_i and m_j are the populations of the locations, $f(r_{ij})$ is a deterrence function depending on the distance r_{ij}, and α and β are adjustable exponents:

$$T_{ij} = \frac{m_i^{\alpha} m_j^{\beta}}{f(r_{ij})}. \tag{2.6}$$

The reception range of a mobile phone tower may vary from a few hundred meters in a metro to a few kilometers in rural areas [5].

Becker *et al.* clustered people into groups from anonymized *call detail records* (CDRs) generated from mobile phone communication (e.g., voice calls or SMS). Although the locations are not precise because the location uncertainty is about a square mile [7, 45], these location records may provide valuable insight into aggregate behaviors of people [5, 45]. The loca-

tion uncertainty from CDR varies with tower height, terrain and radio power. The authors looked into the hourly intensities of voice calls and SMS of only heavy users, who exceeded the thresholds in their study. By analyzing the locations of towers that handled voice calls and texts, the patterns of usage (e.g., intensities at different times in weekdays/weekends), and the percent of mobile calls, they found strong agreement between the distinctive features of the clusters and the group of people represented by them.

2.4 Next Location Prediction

Entropy rate, based on information theoretic entropy, is an well established metric of the predictability of human mobility when mobility history is considered a time series [7]. Song *et al.* provided theoretical background for an upper bound of correctly predicting the next location based on past location history, and established relationship between observed high upper bound of predictability and features such as visit frequency or radius of gyration [7].

The best that a predictive algorithm can perform when predicting the next location based on a person's past history is to choose the most likely location [20]. Alternatively, it can be said that given the past history, the probability of being in the most likely location is an upper limit of predictability for human mobility [20]. Song *et al.* defined predictability, Π, as (2.7), where h_{j-1} is the time series of past history between time intervals 1 and $(j-1)$. Given h_{j-1}, the probability that the person will be in the most likely location at time interval j is given as $\pi(h_{j-1})$. $\Pi(j)$ is the predictability or best success rate of prediction at time interval j based on time series of length $(j-1)$. Given h_{n-1} and $\pi(h_{n-1})$, if there are N candidates for x_n. Song *et al.* [20] assigns $p = \pi(h_{n-1})$ to the most likely location and $\frac{1-p}{N-1}$ to each remaining location. An upper limit of overall predictability, Π^{max}, is given as the solution to (2.8), where $H(X_n|h_{n-1})$ is the conditional entropy of X_n given h_{n-1}. The dataset used consisted of phone call records of three months, collected for billing purpose, of 50,000 high-use users [5, 7]. Mobile calls of the chosen users were handled by more than two cell towers during the study period, and on average, the chosen users had made at least one call every two hours.

$$\begin{aligned}\Pi &= \lim_{n\to\infty} \frac{1}{n} \sum_{j}^{n} \Pi(j) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{j}^{n} \left(\sum_{h_{j-1}} P(h_{j-1})\pi(h_{j-1}) \right)\end{aligned} \tag{2.7}$$

$$\begin{aligned}&\lim_{n\to\infty} \frac{1}{n} \sum_{j}^{n} \left(\sum_{h_{n-1}} P(h_{n-1}) H(X_n|h_{n-1}) \right) = \\ &- [\Pi^{max} \log_2 \Pi^{max} + (1 - \Pi^{max}) \log_2 (1 - \Pi^{max})] + (1 - \Pi^{max}) \log_2 (N - 1)\end{aligned} \tag{2.8}$$

Song *et al.* [7, 20] defined three types of entropy: random, temporal-uncorrelated, and true. Random entropy (H^{rand}) is calculated by considering each location equally likely to be selected as the next location by a location prediction algorithm. Temporal-uncorrelated entropy (H^{unc}) is calculated by assigning probabilities to locations according to the respective frequencies in the past traces, but the correlation in temporal pattern is ignored. Calculation of the true entropy (H) considers the temporal and spatial patterns.

The three types of entropy are related as $0 \leq H \leq H^{unc} \leq H^{rand} < \infty$ [94]. The entropy distribution of the users peaked at 0.8 which means that the uncertainty about a user's location is $2^{0.8} = 1.74$ locations. Π^{max} distribution peaked at 0.93 for true entropy distribution contrary to 0.3 for uncorrelated entropy, which indicates that the temporal order of location visits hold a significant amount of information about the next location to visit. The upper limit of predictability was shown to saturate at 0.93 for users with large radius of gyration ($\geq 10km$ or $\geq 100km$). For smaller radius of gyration, the maximum predicatbility was higher. A person spends most of his/her time in a few locations and even two top locations account for 60% predictability, which increases as more top locations are considered.

Qian *et al.* pointed out the dependence of entropy on spatial quantization [22]. Smith *et al.* introduced constraints to the model put forward by Song *et al.* [7] and spatial quantization finer than that of GPS data, to lower the upper bound of predictability [21].

For the purpose of predictability analysis, the entropy rate is generally calculated off line from the entire trajectory [7,21,22]. Researchers have also considered the case of real time or

local estimates of entropy rate [8, 29, 123]. This may be useful to detect any deviation from regular behavior (e.g., arrival at a new place or an unconscious patient). Carrion *et al.* [8] proposed an estimator to overcome the computational intensiveness of the instantaneous entropy estimation [123] to make the computation feasible on devices with limited processing power (e.g., mobile phones).

Apart from the limit of predictability, researchers have proposed various algorithms for predicting human behaviors and predict mobility patterns into the future. To predict the next location of a person in a cellular network, Anagnostopoulos *et al.* classified the observed trajectories of that user [124]. The short term movement up to the point of prediction is then matched against the historical observations to perform the prediction, taking into consideration the surroundings or neighboring network cells at the time of prediction. Jeong *et al.* clustered the mobility data of the users and applied Bayesian models to predict the next location from the mobility information in similar user groups [125]. Bohnert *et al.* uses state based predictors and transition frequencies between discrete locations from the data of similar users to estimate transition probabilities [126].

The Hidden Markov Model (HMM) is a popular tool for next location prediction [127, 128, 129, 130, 131]. It is common to combine trajectory clustering using algorithms such as K-Nearest Neighbors (K-NN) and G-means clustering alongside HMM for location prediction [128,129,130,132]. Mathew *et al.* used temporal periods of location visits for clustering [131]. Ying *et al.* leveraged semantic data in the clustering process to mine significant locations which are compared with the current trajectory [132].

Markov models of prediction take decisions based on the current state of the model and observed data, but don't account for behavioral correlations that span longer periods (e.g., evening activity and waking up late in the following morning). Eagle *et al.* used eigen-decomposition of behavior data (hourly classification of locations into home, office, elsewhere, no signal coverage, and phone turned off) from human mobility data sets to predict daily behaviors of people and demographic groups from the weighted sums of a metric called an eigenbehavior [133]. The eigenbehaviors are defined as the eigenvectors of the covariance matrix of the behavior data. The eigenbehaviors with highest eigenvalues are considered primary or top eigenbehaviors, and account for most of the variance. They found that the

deviation from a set of top eigenbehaviors is limited. They achieved 79% accuracy with prediction after calculating the weights for the eigenbehaviors halfway through a day. By using six primary eigenbehaviors, they also achieved up to 96% accuracy in clustering people into demographic or behavioral homophily-based groups.

2.5 Data Mediation

Geographic knowledge is produced from data that are normally mediated by algorithms, which can be either computer-based programs or manually completed methodologies. Geographic knowledge, therefore, is not independent of the mediating methodologies, data model/structure, and computational platforms [6]. Before the advent of big data borne by location-aware technologies, Internet search engines, and social media, researchers used specific questionnaire-based surveys or activity and travel diaries to study human mobility. Those methods were limited to small to moderate data sets. Due to the small size it was feasible to study the effects of different algorithms on the data, and to correct errors. Big data warrants the need to automate and expedite the tasks of researchers, associated with some perils:

- it becomes prohibitively costly to analyze the effects of different algorithms as the data size scales up,
- Identifying errors is difficult,
- Automation is needed to analyze and model the data. Errors that are difficult to identify may magnify as they propagate across processes, and
- Different algorithms may result in different end results.

The algorithms used to mediate big data may alter the actual data when collected from a commercial organization [6]. Unlike traditionally collected data, automated data do not normally contain qualitative information including socio-demographic data. Some of these are home and work locations, gender, race, travel route and mode, and incomes. These data play important role in geographic decision making, and are normally inferred by algorithms,

or complemented by traditionally collected data from the participants. Moreover, the algorithms used to mediate big data may evolve over a short period, and thereby, introduce uncertainty. This mediation may become an obstacle if the data are provided by a business organization.

Many researchers have used data from mobile phone call records in human mobility studies [5, 6, 7, 21, 38, 39, 40] because they provide data on a large number of people, who can be considered representative of the overall population, over long periods. Location data from mobile phone call or SMS records do not provide the actual positions of phone users, but those of the towers that handle the phone calls. This limitation introduces a spatial constraint (the distance within which a user's movement can not be distinguished) and a temporal constraint (a user needs to spend a minimum amount of time at a place to consider it an activity location) in the mediated data. As a result, accommodating fine-scaled movements and activities is infeasible in these datasets. Short trip data are unavailable in these datasets. The available trajectories may differ from the actual ones by a large margin, and can account for things like ping-pong effects, where tower hand offs between large cells are regarded as significant trips.

Many different technologies have been brought together to capture human mobility data at a small scale. Although some algorithms may infer activities from space-time characteristics of the activity or land use data, the results may not be verifiable [6]. Street networks have been used to approximate movement but this may not work well when the street network is dense [6, 134]. Kwan cautions that mediating algorithms may play an important role in the research findings, and algorithmic internals becomes increasingly invisible to the researchers with the increase in the algorithmic mediation of geographic knowledge [6].

2.6 Summary

Despite the problem of data mediation and proprietary algorithmic processing, the availability of big data has driven novel research not possible with previous data collection strategies. The many facets of these research works addressed different avenues to explore in-depth knowledge of human mobility and its applications.

Many simplified mobility models that assumed random movement or exponential distribution of different mobility features, were proposed as early endeavours of demonstrating human movements in applications like wireless ad-hoc network models [52, 95, 106, 106]. Further explorations with the advent of location-based technologies have mostly found the prevalence of power-law distributions in different mobility behaviors including trip length, trip duration, dwell time or inter-contact time [3, 5, 13, 39, 96]. Pioneering work by Barbasi's group has paved the foundation of estimating limits of predictability by location prediction algorithms [5, 7].

Different algorithms have been proposed to predict the next location of a person and achieved satisfactory levels of accuracy, an indication that humans are generally predictable. Some algorithms also address how to handle newly encountered locations in real-time next-place prediction. Such predictive algorithms have been used in diverse application areas including traffic monitoring and the spread of infectious diseases. A metric of predictability is required in many such applications. Entropy rate is an accepted metric for predictability or uncertainty with human movements. The predictability is calculated empirically from past mobility traces using the LZ-compression algorithm.

Researchers have shown the presence of a high degree of correlation among different mobility properties. Different studies generally report different distribution parameters for the same mobility features. This occurs because of different underlying spatio-temporal resolutions and that the populations studied are different. As a metric of uncertainty, entropy rate is a reflection of the underlying mobility features (e.g., dwell time distribution) in the location traces. Therefore, the entropy rate depends on the spatial and temporal scales used in the study as well as the studied community.

The intended mobility entropy rate model of this dissertation is developed in three steps. As the first step, this dissertation shows in Chapter 3 that mobility metrics depend on spatio-temporal resolution at which mobility traces are captured. Different populations have different degrees of sensitivity to changes in spatio-temporal resolutions. This shows that population-specific factors effect the mobility predictability beyond the spatio-temporal resolution. Secondly, the model in Chapter 4 makes simplistic assumptions to explain how spatio-temporal resolution effects the mobility entropy rate. The third manuscript in Chapter 5

exploits the correlation among mobility metrics, and extend the simple model of Chapter 4 to represent the mobility entropy rate as a function of five population specific parameters in addition to the spatio-temporal resolution. The development of the final model depends on the GPS traces of the participants. The dissertation, therefore, builds upon the topics like applications of mobility predictability, distributions exhibited by different mobility metrics, mobility models, and data collection and mediation effects.

Chapter 3

Manuscript 1

Title: Scaling Behavior of Human Mobility Distributions

Citation: Paul T, Stanley K, Osgood N, Bell S, Muhajarine N. Scaling Behavior of Human Mobility Distributions. In: International Conference on Geographic Information Science; 2016. p. 145–159. Springer International Publishing.

Abstract Recent technical advances have made high-fidelity tracking of populations possible. However, these datasets, such as GPS traces, can be comprised of millions of records, well beyond what even a skilled analyst can digest. To facilitate human analysis, these records are often expressed as aggregate distributions capturing behaviors of interest. While these aggregate distributions can provide substantial insight, the spatio-temporal resolution at which they are captured can impact the shape of the resulting distribution. We present an analysis of five spatial datasets, and codify the impact of rebinning the data at different spatio-temporal resolutions. We find that all aggregate metrics considered are affected by rebinning, but that some distributions do so regularly and predictably, while others do not. This work provides important insight into which metrics can be used to compare human behavior across datasets and the kinds of relationships between that can be expected.

Relationship to this Thesis The ultimate goal of this thesis is to provide a scale-invariant metric of entropy rate for studying human mobility. To establish the degree to which entropy rate depends on spatio-temporal measurement resolution, we must first empirically establish the resolution dependence of the aggregate distributions on which the entropy rate indirectly depends. I show here, that although different human mobility features are sensitive to spatial

and temporal scales, their degrees of sensitivity vary. The parameters of the models, which change with varying resolutions can be tracked well within the ranges of spatial and temporal resolutions on the scale of human movement. Mobility features are related to mobility entropy rate and predictability. This gives the intuition for the following work that entropy rate should also be defined as a function of spatial and temporal resolutions over the range dictated by the scale and frequency of human movement.

3.1 Introduction

Human spatial behavior underlies many disciplines, including geography, sociology, architecture, and many forms of engineering. Research effort has been invested attempting to describe how people move through and use space. Through studies conducted with pen and paper through diaries, surveys, or ethnographies, researchers have made significant strides in codifying how people move through and utilize space. With the advent of mobile communications and location sensing technology, vast new repositories of spatio-temporal information on human mobility have become available. Voronoi diagram-based spatial decomposition of location data from cell towers or WiFi router access logs, trajectory data from GPS logs, or interaction level data from RFID and Bluetooth (BT) beacons all provide previously unprecedented representations of a person's spatial trajectories [7,32,92,135,136]. However, all of these data sources have different characteristics: cell record and WiFi data are characterized by irregular spatial distributions contingent on inter-device spacing and only generate records when people connect, GPS logs are only reliable outdoors, and BT and RFID devices provide reliable measures of proximity but only in controlled settings. Even reliable measurements via GPS have variable accuracy depending on the device, atmospheric conditions, and built environment.

To cope with the large amounts of data generated by these new measurement techniques, researchers often employ aggregate metrics, which can be characterized as distributions over a single variable such as trip length, to help describe the data. The model parameters (e.g., mean and variance of a Gaussian) corresponding to these distributions can be used to describe the data concisely. For example, many human-centric statistics, such as visit frequency or interpersonal contact duration, are characterized by truncated power law distributions [13]. The power coefficient describing that distribution can provide insight into the relative behavior of two populations. However, because changing the spatial extent over which these data are collected can change the shape of the distribution, studies of the same populations at different spatio-temporal scales will be described by different model parameters, and by extension may generate erroneous conclusions. This hearkens back to the Modifiable Areal Unit Problem, a recurring challenge when working with data and variables that can

be aggregated to different units of analysis [137, 138]. Understanding to what extent these distributions are susceptible to the spatial and temporal resolution of collection, and to what extent these sizing and sampling effects are predictable based on underlying mathematical processes, should help human behavioral researchers make meaningful comparisons across datasets and between populations.

Employing five mobility datasets, recorded from either smartphone GPS or GPS logging devices, we analyzed sampling effects. To model spatial binning, an area of interest was binned into square sections of varying sizes. To model temporal granularity, we downsampled the mobility traces at regular intervals. This selective and regular resampling allows us to examine the impact of spatio-temporal resolution on the resulting aggregate distributions. We find that some distributions have definitive scaling behaviors, indicating the possibility of meaningfully comparing datasets across resolutions. Other metrics do not vary as regularly under resampling, indicating that caution should be exercised when comparing results from different data sources using these techniques.

3.2 Related Literature

Human mobility is not random, but follows well defined patterns [5, 13, 92, 94], sometimes characterized by aggregate statistics like: 1) inter-contact time, 2) visit frequencies, 3) dwell time, 4) radius of gyration, 5) trip length, and 6) trip duration [5, 13]. Because mobility is continuous in space and time, quantization (binning) is often applied [5, 21, 22]. While the transmission range of a GSM (Global System for Mobile Communications) base station is normally up to 35 km [43], Bluetooth and WLAN (Wireless Local Area Network) transmission ranges are limited to tens of meters to a few hundred meters [44]. A study found position errors of $2m$ - $15m$, on average, using GPS [46].

The examination of different units of analysis in geography has a long tradition [137,138]. Persuasive arguments for considering these effects in GIScience are also well documented [139], including in work on the convergence of GIScience and Social Media [82]. Bell *et al.* examined similar sized units of different types (census vs neighbourhoods) and found similar patterns [140].

Eagle *et al.* used mobile phones to collect location history and behavioral data of people from multiple sources [79]. For better granularity of spatial data indoors, and social context, locations of surrounding Wi-Fi access points have been used [35, 36, 141]. The resultant datasets provide valuable insight into the interwoven patterns in human movements. Research on intertwined patterns in human mobility led the development of synthetic human mobility models, which emulated observed patterns in human mobility traces [13,92,95,142].

Data-driven geographic inquiry or algorithmic geographies [6], are increasingly important to understanding human behavior, complex systems, and our environment-behavior interactions [143]. Urban geographers, demographers, and behavioral geographers are using open, big, and real-time (or streamed) data in new ways. This includes health [144], communication networks [30], transportation [141], and behavior modelling [145]. In GIScience and its cognate geographic disciplines, the application of grid cells and varying spatial resolutions has primarily been in remote sensing and elevation modelling [146, 147].

3.3 Experimental Setup

We used five data sets: the Saskatchewan Human Ethology Datasets (SHED) 1, 2, and 5 [35, 36, 141], the open source dataset GeoLife [75], and GPS traces from the 'Seasonality and Active Saskatoon Kids' dataset (hereafter, the 'Kids' dataset) [144]. The SHED datasets are technical pilots for the ongoing development of iEpi [135], and contain detailed mobility, activity, and contact traces from graduate students and staff (SHED1 and SHED2) or undergraduate students (SHED5). GeoLife is an open source collection of mobility traces collected using GPS loggers by Microsoft Research [75] in China. The GPS traces in GeoLife correspond to self-identified trips taken by participants, and do not include stationary periods. The Kids study [144] used GPS loggers and wearable accelerometers to study a large number of elementary students from low income neighbourhoods over a week, to determine activity and mobility patterns.

Software glitches, hardware failure, and participant non-compliance led to significant variance within the number of available records in each of the databases. Individual participants returned anywhere from negligible fractions, to almost complete records of possible

Table 3.1: Dataset properties

	SHED1	SHED2	SHED5	Kids	GeoLife
Study duration	4 Weeks	4 Weeks	4 Weeks	1 Week	5+ years
#(participants)	38	37	29	745	182
#(used participants)	34	27	24	722	33
#(GPS records)	1.35e6	3.41e7	279,298	1.54e8	2.5e7
#(used records)	107,409	101,746	80,998	1.42e8	1.86e7

data, but only a portion of the total number of records included GPS data (e.g., while at school or university, SHED or Kids participants might report accelerometer but not GPS records due to poor GPS reception indoors). Participants were classified into two groups, responders (at least 20% of possible time slots or samples with GPS data over the data collection period) and non-responders, for all but Geolife, where compliance was difficult to assess because data corresponds to participant-identified trips. The threshold of 20% was chosen arbitrarily based on inspection of trajectories. Participants whose GPS records were available for less than 20% of the possible time slots were removed. GeoLife data were sampled at $1-5$ s intervals [75], and participants were included in the analysis if they had recorded trips spanning at least two weeks. The number of participants and records before and after filtering are presented in Table 3.1.

To determine the impact of the temporal sampling rate, we down-sampled the data (expressed by T), between subsequent measurements. A down-sampling period (T) is an integer multiple of the base period (T_0) at which GPS data are collected. Down-sampling at T is performed by taking every $(\frac{T}{T_0})^{th}$ sample from the base data. Because each dataset has a different minimum sampling time (between 1 second and 8 minutes), we standardized the minimum sampling duration to be 8 minutes for SHED5, and 10 minutes for others. For SHED5, $T \in \{8mins \times (1, 5, 10, 15, 30, 60)\}$ and for others, $T \in \{10mins \times (1, 3, 6, 12, 24, 48)\}$. The fastest sampling rate was chosen for consistency with SHED5, which had the slowest base rate; the slowest sampling rate was chosen to be 3 times per day, consistent with the minimum number of daily cellphone records required by Song *et al.* [7]. This downsampled sequence

was then sampled spatially using a regular square grid. If no location record existed at the downsampled timestep, then a special symbol for unknown location was used for the location of that participant at that timestep. These special symbols were ignored when creating the aggregate distributions, but they broke trips during trip length and duration calculations. Both choices were intentionally conservative; we assign no location if the location is unknown, and do not assume that a trip continues if data during a trip is missing. This will tend to make trips shorter, as potentially longer trips may be broken into a number of shorter sub-trips.

Location was binned with a maximum granularity of 4 km, consistent with a suburban cell tower area, with that granularity successively reduced by factors of 2 to a minimum of 15.625 m, consistent with the nominal accuracy of commodity GPS receivers common in typical smartphones. The spatial resolution is reported as the length of the square bins or grid cells, and given the symbol d. The coverage area of a dataset was gridded at the coarsest resolution (4 km edged squares), and increasingly finer resolution cells were created by subdividing these larger cells into 4, halving the edge dimension while conserving the topology of the spatial binning, until the finest resolution of 15.625 m was reached. Locations were taken to be the centers of the grid cell in subsequent calculations. Over short time scales and at fine resolutions, there was strong agreement between the recorded position and the binned locations; as temporal and spatial scales expanded, agreement between computed location and measured location began to diverge, as expected. Intra-step shifts in time or space (e.g., changing the start time or base grid locations) was not investigated, but would also be expected to have an impact, particularly at coarse spatial or temporal scales.

We computed five previously employed aggregate metrics [5, 13] for each dataset at each spatio-temporal resolution: visit frequency, dwell time, trip length, trip duration, and radius of gyration (RoG). All empirical distributions are aggregated across locations and participants through time.

Visit Frequency: The distribution of the count of participant samples in a given location. Remaining in a cell increases the count for that cell. This metric indicates overall place popularity.

Dwell Time: The distribution of the number of time steps participants spent in a cell without changing cells. This metric distinguishes between places visited often, for short duration, versus those visited occasionally for longer.

Trip Length: The distribution of contiguous trips, where a trip is defined as changing locations for at least three consecutive downsampled time steps. Distance is calculated as the Euclidean distance, which is an integer multiple of d, between cell centers for each stage of the trip. If a trip spans l cells, the trip length is ld. The trip length distribution specifies the probability of traveling a certain distance.

Trip Duration: The distribution of time spent in a trip (as defined above), with a resolution of the current sampling period. Trip duration describes how long participants are likely to remain in transit.

Radius of Gyration: This metric, represented as r_g, is defined in (3.1), where c is the center of the polygon bound by spatial resolution-dependent coordinates $\{r_i : i \in \mathcal{N}^+ \wedge i \leq N\}$ of trip samples. The RoG distribution describes how compact the areas traversed by participants are. We computed c as the centroid of the convex hull of the polygon defined by trip samples.

$$r_g = \sqrt{\frac{1}{N}\sum_{i}^{N}(r_i - c)^2} \tag{3.1}$$

Given the distributions of the above metrics at chosen spatio-temporal resolutions, we used regression for power-law-based fits of the distributions because the metrics have been reported to follow truncated power law distributions [5, 13]. Under the power law model, each distribution has two parameters, a constant term and an exponent term, encoded as α and k, as shown in (3.2).

$$f(x) = \alpha x^k \;\; [x \geq x_0] \tag{3.2}$$

After determining the model parameters α and k of (3.2) from the distributions of each of the five metrics at different spatio-temporal resolutions, we determined how these model

parameters varied with d and T using the following models on the basis of R^2-based goodness of fit:

Linear: $f(x) = c_1 + c_2 x$,

Logarithmic: $f(x) = c_1 + c_2 \log x$,

Exponential: $f(x) = c_1 c_2^x$,

Power: $f(x) = \alpha x^k$.

Data were stored as text files. Initial data exploration was done using Eureqa[1] from Nutonian, Inc. Our final fits were done using R statistical software[2] with R^2 as the goodness of fit metric. Calculations were carried out on a computer with 4-Core AMD processor and 8 GB memory running Ubuntu 15.10.

3.4 Results

As we are primarily interested in determining how aggregate distributions of mobility change under different spatio-temporal measurement regimes, we have plotted the distributions of aggregate metrics. Fig. 3.1 and Fig. 3.2 show the distributions of our key metrics at spatial dimensions of 31.25 m and 500 m respectively, where the counts are plotted along the Y-axis. Each curve within each graph denotes a particular (dataset, sampling time) pair.

Several trends are notable within each graph. First, most curves show the characteristic forms for power law distributions, which is consistent with the literature [5, 13]. All curves (with the exception of RoG) are characterized by linear descent on the log-log plots over large portions of their span, indicating heavy tailed power distributions.

Second, not all datasets are equal. The Kids dataset is characterized by longer dwell times than the other datasets. This is likely indicative of the relative difference between elementary school students' and university students' lifestyles. The GeoLife dataset, comprised

[1] http://www.nutonian.com/products/eureqa-server/

[2] https://www.r-project.org/

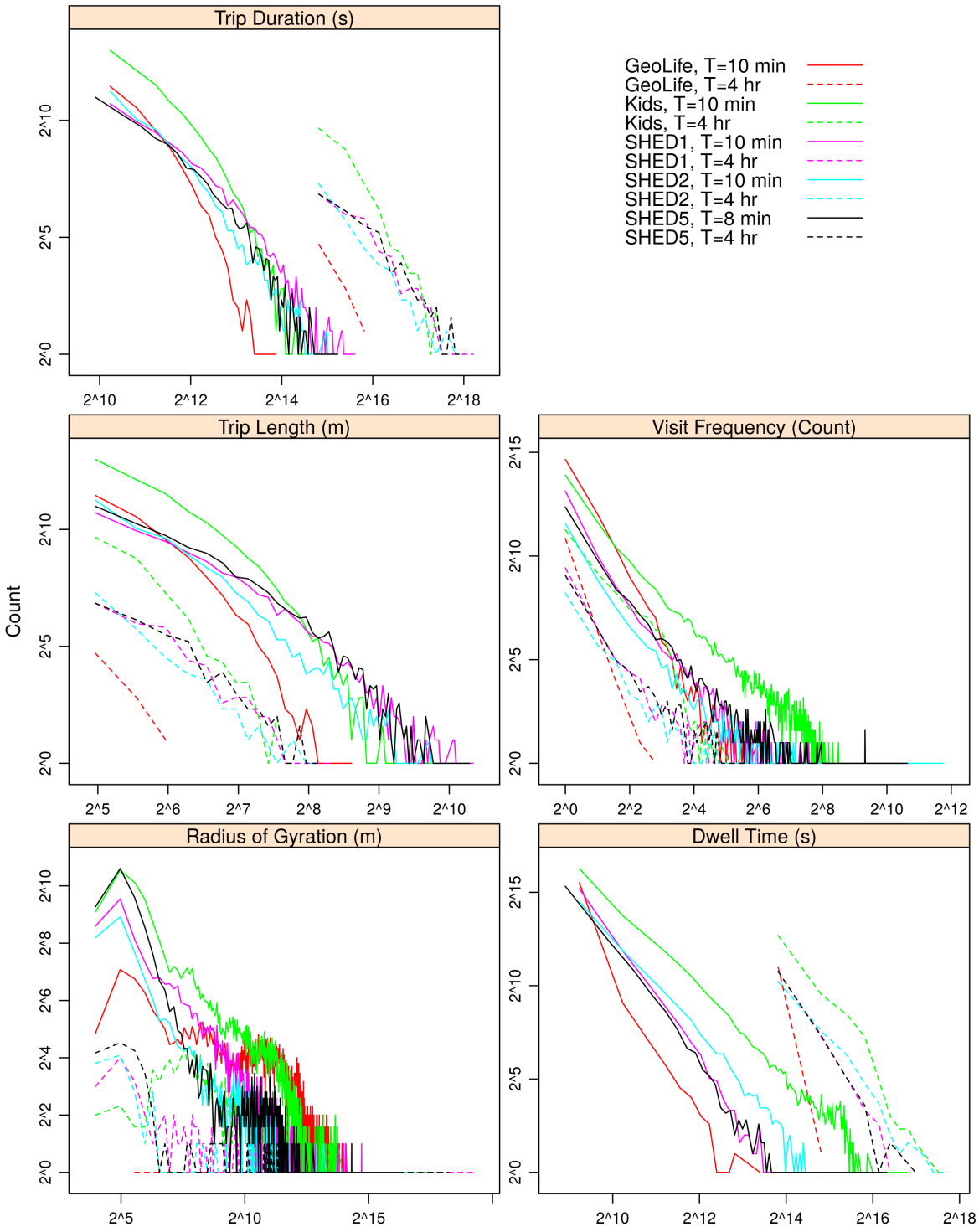


Fig 3.1: Distribution of dataset features at $d = 31.25m$

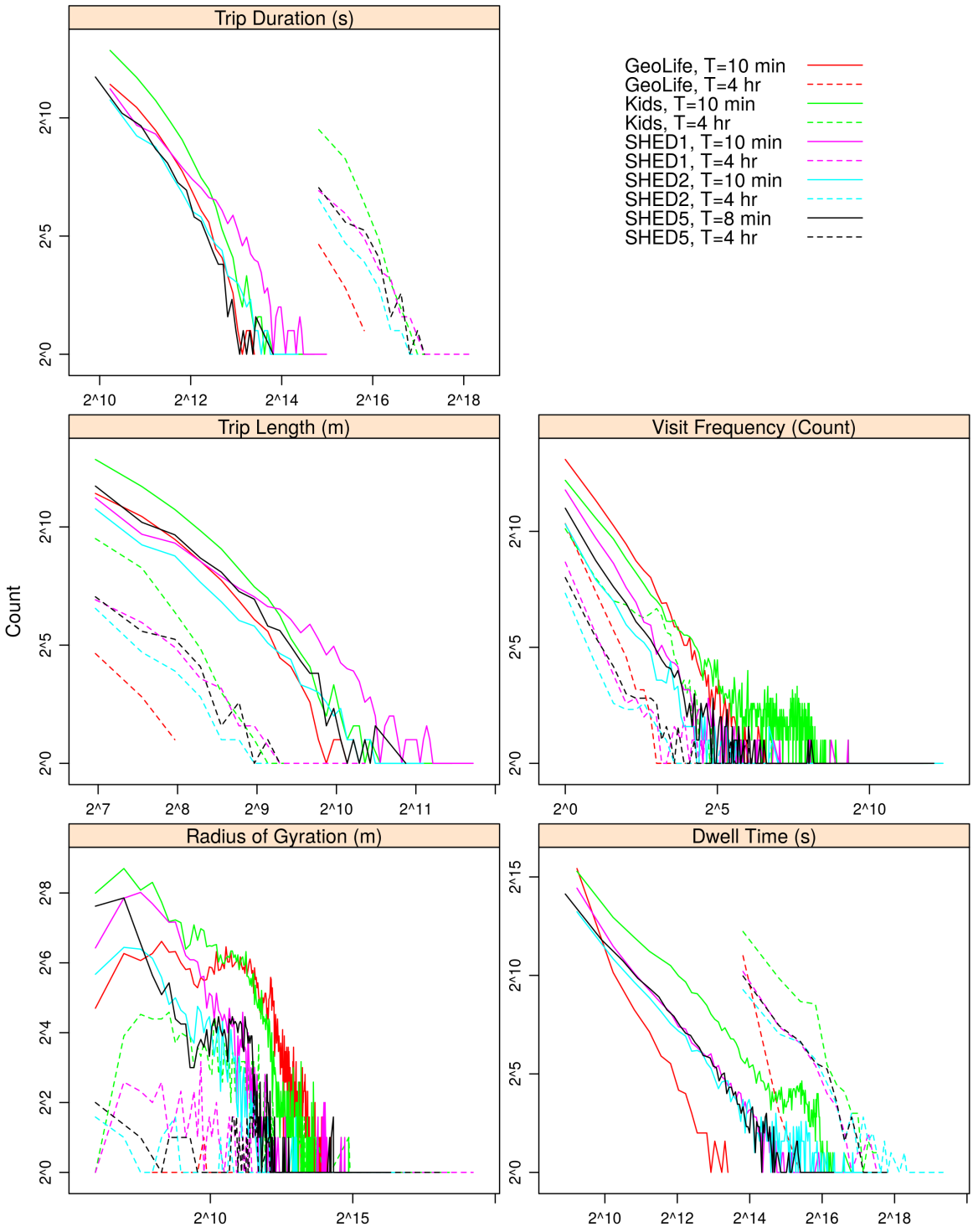


Fig 3.2: Distribution of dataset features at $d = 500m$

exclusively of trips, has a lower dwell time, and higher visit frequency and RoG, which is as expected for participants who are always on the move.

Third, the RoG measure is noisy with respect to sampling regime. Given that the formulation for RoG implicitly depends on the sampling regime, this makes sense, as altering capture resolution alters the parameters of RoG. As a result, we conclude that RoG is a poor measure for inter-experiment meta analysis, as significant variation in computing values will be expected due to the data capture resolution.

Fourth, dwell time, trip length, and trip duration are well characterized by power law distributions, as characterized by Fig. 3.3, where each box plot represents the distribution of R^2 values when fitting a power law to curves aggregated over participants, as seen in Fig. 3.1 and Fig. 3.2, for each d and T pair considered in the expertiment. As expected from the noisy signal, RoG is poorly characterized by a power law. Visit frequency does not appear to be strongly power law distributed, particularly near the tails. The noisy tails also make visit frequency susceptible to changing fit quality with spatio-temporal resolution.

Fifth, there is apparent regularity in much of the variation in both Fig. 3.1 and Fig. 3.2, implying underlying mathematical relationships. To determine the regularity of effect, we further fit curves to the model parameters derived from the regression for power law distributions fits, to determine if the coefficients of the fit equations also vary regularly with resolution. That is, we wished to determine if the model parameters could be expressed as functions of the resolution.

Given the model parameters α and k, derived from power law-based regressions of the distribution of key metrics, we tried to relate them to d and T. Fig. 3.4 presents the R^2-based fit qualities, aggregated over all datasets, of exponential, linear, logarithmic, and power law regression models to establish relationships between model parameters (α and k) and spatio-temporal resolutions (d and T). Overall, power models explain the behavior of α and k with d and T best, exhibiting the largest mean R^2 values and smallest variances. However, values of k showed significant variance, and trip length and RoG had generally poor fits for all models tested.

Fig. 3.5 presents the R^2 fit quality values of regression fits of power law model parameters (α and k), as d or T broken down by dataset. Each value in the boxplot is represented by

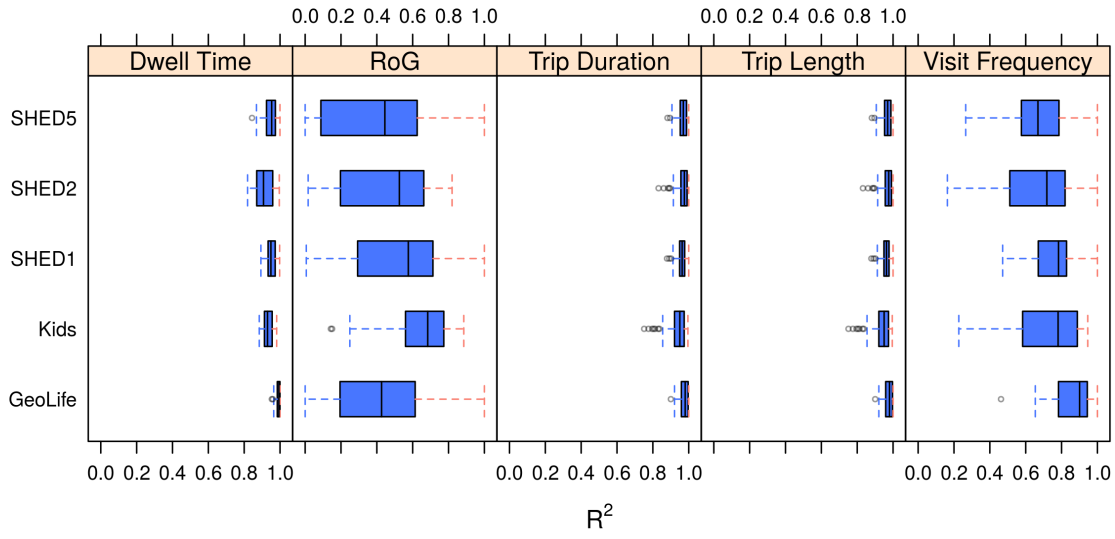


Fig 3.3: R^2-based quality of power law fits of distributions of dataset features

a single spatio-temporal resolution (value of d and T), aggregated over all participants for a single dataset. Much of the variance in these fits can be ascribed to the power law only describing a region of variation, as would be expected from Fig. 3.1 and Fig. 3.2, where, for example, visit frequency becomes quite noisy with large T, or there is limited variation among datasets for trip duration at small T. Visit frequency and dwell time seem to have the strongest power dependence on both d and T. It is interesting to note that, while visit frequency did not consistently hew to a power law distribution, the variation of the model parameters did vary regularly. Trip length model parameters vary somewhat regularly with d, but are inconsistent across datasets with T. RoG and trip duration do not exhibit strong fits. With RoG, this is expected, given the noisiness of the original signal, but with trip duration this is more likely due to the changing definition of a trip, as changing d and T changes possible trip lengths.

3.5 Discussion

Understanding human mobility and its measures is increasingly important for many fields. In this paper, we sought to examine the impact on aggregate metrics of spatial scale and

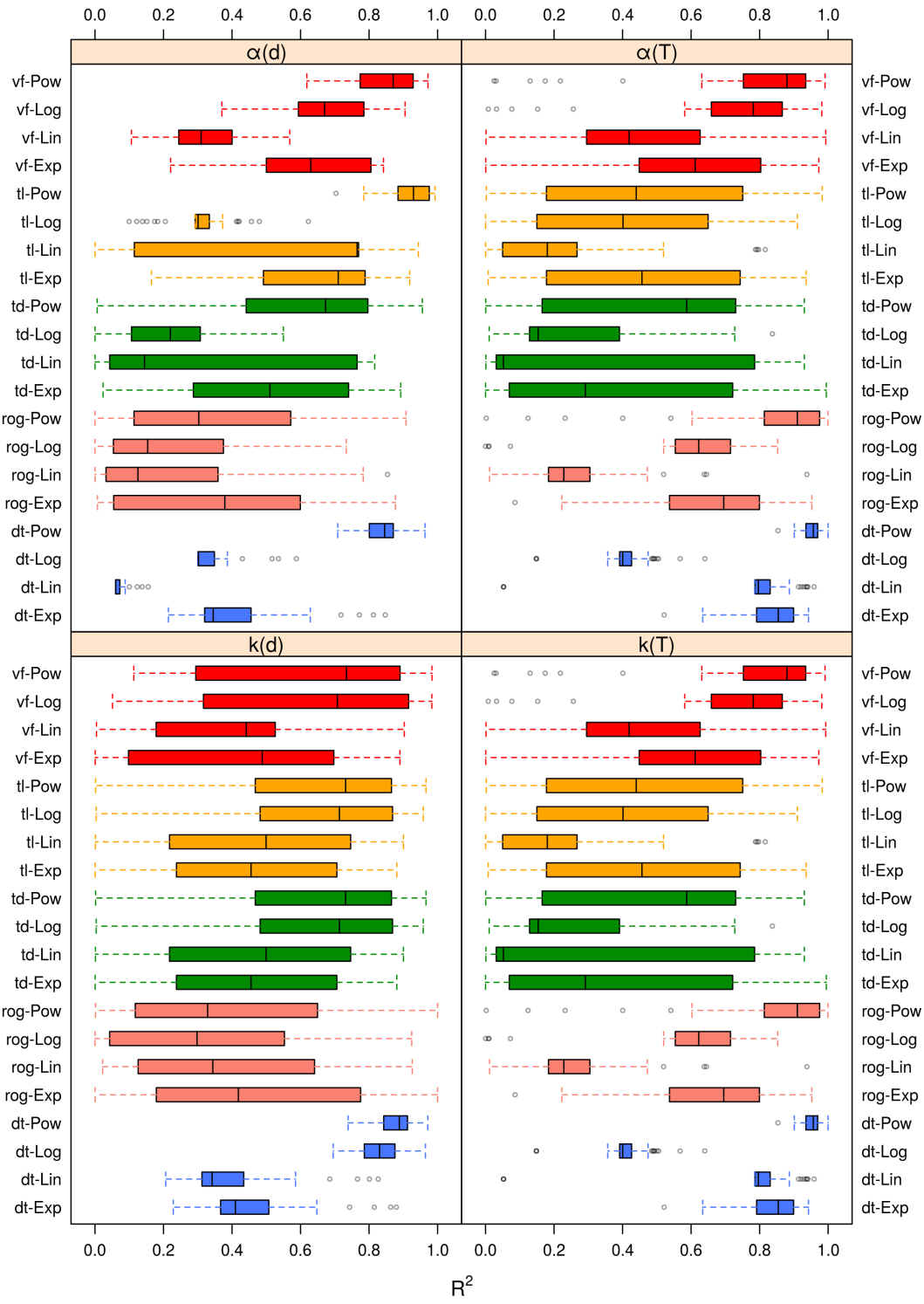


Fig 3.4: Goodness of fit of $\alpha(d)$, $\alpha(T)$, $k(d)$, $k(T)$, for key metrics over all datasets, to exponential (Exp), linear (Lin), logarithmic (Log), and power law (Pow) models

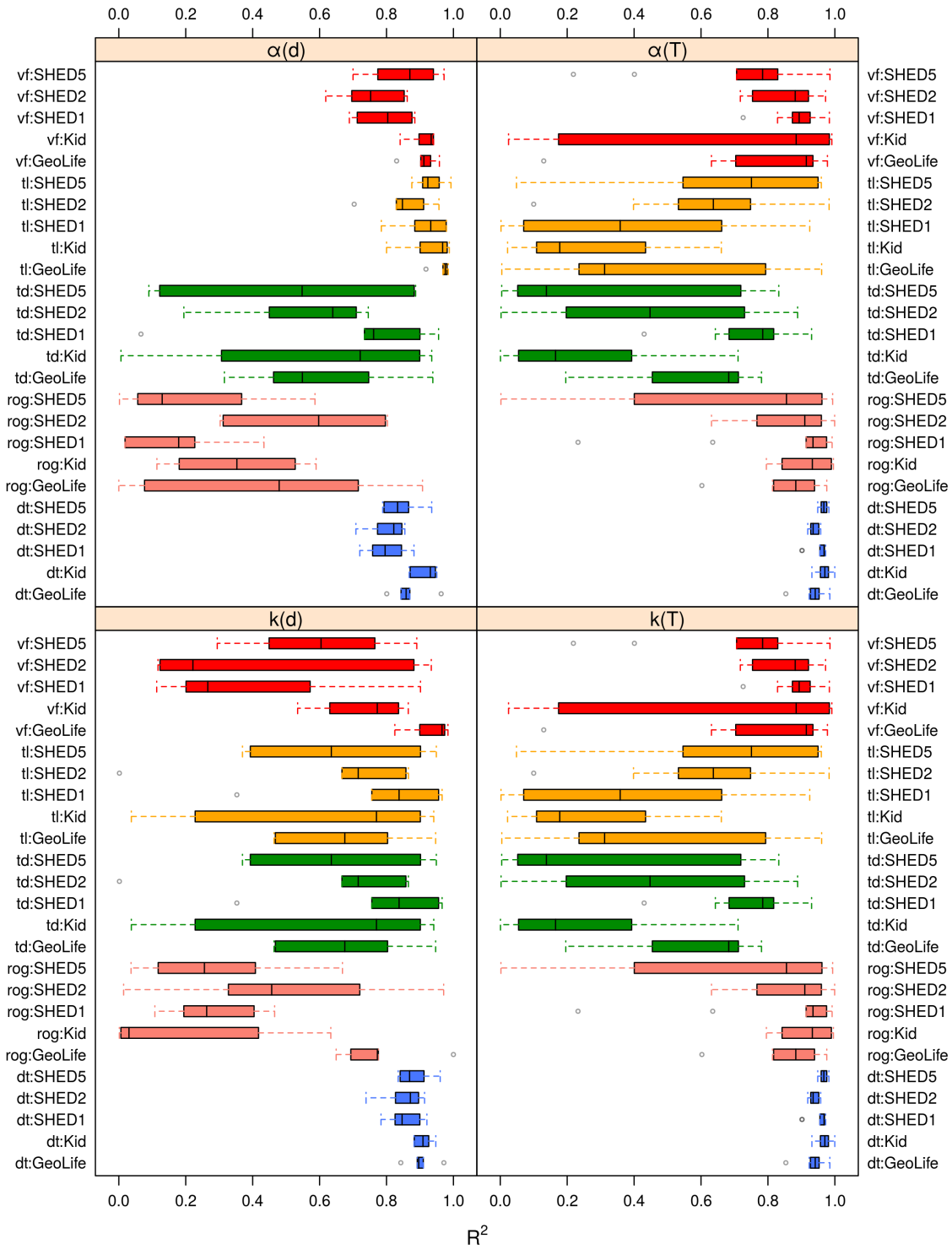


Fig 3.5: Power function-based fit quality dependence of a and k on d and T

temporal sampling period. We analyzed five spatial datasets, which have not been analyzed in this manner before, deriving distributions of previously reported spatial metrics. Through our analysis, we report the following findings:

1. **Metrics had well defined and consistent distributions**. With the exception of RoG, distributions were found to generally be heavy tailed power law distributions, as expected. The form of the distribution was consistent across datasets and resolutions, although parameters describing these distributions varied with spatio-temporal resolution.

2. **Binning changes the data and fit**. For all metrics, changing the spatial bin size or temporal sampling period changed the shape of the resulting distribution. That is, measuring or analyzing the data at different resolutions provides different answers. When employing datasets obtained from empirical data in models, or when comparing two empirical datasets, caution must be exercised to ensure that resolutions match, or the comparison might not be phenomenologically meaningful.

3. **Ordering between metrics over datasets is generally preserved under resampling**. While it could be perilous to compare metrics over distributions captured at different resolutions, changing resolutions generally did not change the ordering of such metrics. For example, the trip duration of the Kids dataset was almost always greater than GeoLife, for each sampling resolution. There were instances at longer T, where points on the SHED5 tail overlapped the Kids that altered slightly due to sampling effects, but the overall shape of the curves was consistent.

4. **The impact depends on the dataset**. Not all datasets were affected equally by the varying resolution, implying that varying resolution impacts datasets from a sampling mathematics viewpoint, through the underlying behaviors of the individuals, and the data collection context. Different populations and environments may have greater or lesser sensitivity to resampling than others.

5. **The sensitivity to resampling can itself be a metric**. While substantial additional research would be required to understand the behavioral drivers which give rise

to the differential impact of scaling, the fact that there is regularity in the behavior of the model parameters and spatio-temporal scale might be diagnostic of different populations, for example the greater sensitivity of SHED5 over GeoLife to dwell time scale indicates that information about population mobility is encoded in the scaling behavior.

These findings have implications for how mobility data should be employed in research and practice. Finding 1 validates work from other researchers with new data [5, 13]. Finding 2 cautions modelers and researchers employing this data. Because the distributions do not generalize across resolution, data from an empirical study conducted at one resolution cannot, with certainty, provide the underlying distributions for models with a different underlying spatial resolution. Finding 3 indicates that derived metrics such as mobility entropy, which exhibit resolution sensitivity [21, 22] may derive their resolution dependence from the variation described here. Finding 3 suggests that resampling within datasets will not compromise conclusions of an ordinal nature. Finding 4 indicates that the scaling effects are not entirely due to the mathematics of sampling: human behavior patterns in the data also contribute. Finding 5 hypothesizes that resampling behavior itself could be used as a metric of human mobility. These scaling metrics could also be used to evaluate agent-based models of human mobility used in simulation. Synthetic mobility models ([13, 92, 95]) should not only reproduce the distributions of key metrics at a given resolution, but the scaling behaviors noted here. Taken together, these findings provide a meaningful contribution to the study of human mobility metrics.

While we have made a significant contribution to the literature, several shortcomings of this study could be addressed in future work. First, while we used five datasets comprising millions of records, these datasets had a relatively small number of participants and durations measured in weeks. Further analysis of larger, longer duration, and more diverse datasets would help validate the work. Second, we employed GPS datasets, downsampled regularly in time and space. While this approach facilitated the analysis, location data sources such as WiFi and cell tower records have irregular shaped cells based on the Voronoi diagram of transmitter locations, and stochastic sampling patterns based on connectivity behavior. Understanding how irregularity in spatial and temporal sampling impacted these distribu-

tions would also be worthwhile. Finally, we have not attempted to employ these insights into building better models of human behavior. Further research into the application of these findings to building higher fidelity models of human behavior for simulation systems could have wide ranging impacts.

3.6 Conclusion

Spatio-temporal resolution changes the shape and model parameters of aggregate distributions used to describe human mobility. This variation appears to conserve, at least in ordering, the differences between datasets, implying that indications of the differences in human behavior being observed are also preserved. Because spatio-temporal resolution matters, making quantitative comparisons between datasets with different resolutions is potentially dangerous and should be avoided, at least until regularities in the scaling relationships can be better characterized. While significant research remains, this work represents an initial step in understanding how to properly employ newly available high-fidelity datasets in human mobility analysis.

3.7 Addendum

The manuscript in this chapter has been reformatted, and some paragraphs/sentences of the published version have been modified/added/deleted, based on edits proposed by the examining committee, for inclusion in the dissertation. No substantial changes to the results/findings were made.

Chapter 4

Manuscript 2

Title: A Theoretical Basis for Entropy-Scaling Effects in Human Mobility Patterns

Citation: Osgood ND, Paul T, Stanley KG, Qian W. A Theoretical Basis for Entropy-Scaling Effects in Human Mobility Patterns. PLoS ONE. 2016;11(8):1–21.

Abstract: Characterizing how people move through space and time has been an important component of many disciplines. With the advent of automated data collection through GPS and other location sensing systems, researchers have the opportunity to examine human mobility at spatio-temporal resolution heretofore impossible. However, the copious amounts and complex characteristics of data collected through these logging systems can be difficult for humans to fully exploit, leading many researchers to propose novel metrics for encapsulating movement patterns in succinct and useful ways. A particularly salient proposed metric is the mobility entropy rate of the string representing the sequence of locations visited by an individual. However, mobility entropy rate is not scale invariant: entropy rate calculations based on measurements of the same trajectory at varying spatial or temporal granularity do not yield the same value, limiting the utility of mobility entropy rate as a metric by confounding inter-experimental comparisons. In this paper, we derive a scaling relationship for mobility entropy rate of non-repeating straight line paths from the definition of Lempel-Ziv compression. We show that the resulting formulation predicts the scaling behavior of simulated mobility traces, and provides an upper bound on mobility entropy rate under certain assumptions. We further show that this formulation has a maximum value for a particular sampling rate, implying that optimal sampling rates for particular movement patterns exist.

Relation to the Thesis: This work forms the theoretical foundation for the primary contribution of the dissertation - a theoretical model expressing entropy rate as a function of spatial and temporal quantization. The model depends on other parameters: distance travelled and movement speed. Given the parameters, the model provides a basis for comparing two different mobility studies conducted under different sampling configurations. While the model forms a theoretically rigorous foundation for describing mobility entropy rate scaling patterns, it only performs well for stylized simulated mobility models. Further extension/generalization of this model, and its validation against empirical data is left for the final paper. Detailed derivation of the scaling law is provided as a supplementary material (Section 4.6) at the end of this chapter.

4.1 Introduction

The importance of understanding how humans move through, consume and interact with the space they inhabit is a central tenet of geography, urban planning, architecture, and many other social sciences [82]. Being able to concisely represent the quality of human movement through space allows practitioners in these disciplines to design better cities, buildings, and policies. Traditionally, human motion was studied using the pen-and-paper tools of the anthropologist, including retrospective surveys, direct observation, ethnography, or self-report through interviews or diaries. While these techniques have provided remarkable insight into human mobility, particularly into its cognitive aspects, they are limited in spatio-temporal resolution, are prone to observer or reporter bias, and can be time consuming. Technological advances in localization have opened new opportunities for analyzing human mobility [148, 149].

Electronically-mediated population tracking is a practical alternative to traditional pen and paper techniques. Inexpensive loggers or smartphone apps can use the Global Positioning System (GPS) to record trajectories through space [35, 36, 81]. While GPS-based systems provide exceptional positioning quality and coverage when outdoors, they can be unreliable in institutional buildings or in terrain where sky views are blocked. GPS-based data acquisition can also be more cumbersome as participants have to be recruited, potentially outfitted with appropriate equipment and debriefed. An alternate approach is to mine cell tower or WiFi router contact traces through time to generate trajectories by representing the locations of the device and, therefore, the person, as the locations of the towers or routers to which the device is connected [79]. In proximity-based representations, space is implicitly represented as a sequence of polygons, derived from the Voronoi diagram of the beacons. While these representations can be easier to obtain, as cell or router contact records are often maintained by telecommunication companies or institutions, they are also often characterized by a heterogeneous spatial decomposition (based on the Voronoi diagram structure) and intermittent sampling, as records are often only generated for active connections (calls, texts, or data transmission).

These technologically-mediated localization systems provide much higher spatial and

temporal fidelity than traditional methods, are less prone to bias, but are divorced from the cognitive processes underlying the decision making. The additional spatio-temporal resolution can be a double edged sword, as traditional statistical analysis techniques suitable for analyzing survey responses are no longer sufficient for characterizing such data. To address the overabundance and complexity of the data, researchers have looked at visualization methods or statistical metrics to represent the important components of the data more concisely. Binned or aggregate statistical representations are popular. Heatmaps, visualizations of the two dimensional frequencies of parameters of interest, are a standard method of aggregating location over time and space (e.g., [38,150]). Space is typically binned at a specific resolution, then location data is accumulated for each bin. Aggregate distributions of secondary measures can also be useful to summarize high fidelity data. Aggregate measures such as visit frequency, trip duration, trip length, and radius of gyration have been previously reported in the literature [5, 7, 13, 92]. In all of these representations, spatio-temporal variation is marginalized over some variable, destroying important information about the structure of the variability. However, several researchers have observed simple and reproducible patterns and a high degree of spatial and temporal regularity in visited locations of humans [84, 151, 152, 153].

In their seminal paper, Song *et al.* [7] proposed the entropy rate of a mobility pattern as a metric of variability or predictability in human behavior. By discretizing the world, and providing a label to each discretized location, a trajectory through space could be converted into a string of location labels or symbols. As a string, this representation could be summarized by the entropy rate, which is closely related to the compressibility of the string. People with a great deal of regularity in their schedules would be represented by a lower entropy rate than people whose spatio-temporal habits were less predictable. This metric had the advantage of providing a measure of the regularity of spatio-temporal habits of a population as a single number. Song *et al.*'s original work has been extended to other aspects of human behavior, including social contact and activity in both complete and moving average implementations [29, 37, 154].

According to Shannon's original definition, entropy is calculated directly from a random variable or distribution [155, 156]. Entropy could be calculated for aggregated distributions

such as trip length or dwell time, but that representation does not capture the empirical entropy rate for the trajectory string. To approximate entropy rate empirically, lossless compression algorithms are generally employed [28]. In particular, the Lempel-Ziv 78 (LZ) algorithm has been shown to provide asymptotic estimates for the entropy rate of a string as the length of the string goes to infinity [7, 8, 28]. Following the example established in Song *et al.*'s original paper, researchers estimate the entropy rate of a mobility string through LZ compression, although shortcomings with this approach have been noted [21].

Employing the methodology originally proposed by Song *et al.*, it is possible to use LZ compression to approximate the entropy rate of a person's trajectory. However, the entropy rate calculated for this path is not universal, as it depends on the spatial and temporal resolution with which the path is sampled. That is, the resolution of binning and the regularity and rate of sampling impact the entropy rate calculated from the LZ compression technique [21, 22, 157]. Meaningful comparisons of entropy rates between different people or populations can only occur if those rates were calculated from strings with identical spatial and temporal resolution. This implies that meaningful comparison of mobility entropy across experiments is not possible in general, as the experimental protocol changes. It further implies that comparing different individuals in the same dataset could be problematic if there is heterogeneity in the geographic bin size or sampling rate; for example, in a study comparing the mobility of rural and urban populations through cell phone records, where the rural Voronoi cells were systemically and significantly larger than their urban counterparts.

Because mobility entropy rate is a useful metric, some researchers have studied or proposed empirical methods of describing variations in spatio-temporal scale [21, 22, 157]. However, empirical models can be difficult to generalize, as specific models may be tightly tied to the datasets from which they were derived. In this chapter, we provide a theoretical derivation of a scaling law for mobility entropy rate calculated through Lempel-Ziv compression. This derivation is theoretically valid for non-overlapping trajectories which can be represented as a series of line segments navigated at constant velocity over a regular four-connected grid. This scaling model shows excellent agreement with simulated trajectories, even when those trajectories violate assumptions underlying the derivation. Analysis of the mathematical properties of the model yields several key findings. First, variation with

spatio-temporal scale is an inevitable consequence of the LZ approximation. Second, mobility entropy rate at any spatio-temporal scale can be represented by four parameters: the length of the trajectory, the velocity of each segment and the spatial and temporal scales. Third, the model has a unique maxima with respect to the temporal sampling rate, implying that there is a natural sampling rate for a given trajectory which maximally captures the information it encodes. Finally, the performance of this model indicates it might be possible to express mobility entropy rates measured with different experimental configurations at common resolutions, allowing comparison between disparate populations and experiments, allowing mobility entropy rate to be employed to its full potential as a metric.

4.2 Analysis

4.2.1 Problem Structure

Our derivation relies upon the performance of Lempel-Ziv (LZ) compression in approximating the mobility entropy rate, the most common method for estimating entropy rate based on the seminal work of Song *et al.* [7]. As many other researchers have noted [8, 28], this approximation makes strong assumptions about the behavior of the string, notably that it represents a stationary ergodic process, and is sufficiently long for the algorithm to converge. While these assumptions may be violated in practice, the approximation is widely used in the literature. Examining the extent to which this approximation scales will provide valuable insight into the interpretation of existing and future results using this approximation, independent of whether the underlying assumptions are correct.

We constrain our derivation to the behavior of the LZ approximation for patterns of movement only, and do not explicitly consider parameters such as location dwell time. That is, our analysis is most suited to datasets concerned with trips or trajectories, and will not necessarily apply to datasets which capture prolonged periods of rest. The derivation problem then becomes examining how LZ compression functions for a set of paths.

The most fundamental assumption required for this examination is the definition of a path. We define a human mobility path as a series of piecewise linear two dimensional

segments, navigated at a constant velocity. We assume that these paths are executed over a discretized space, as is common in the literature. For convenience, non-uniform Voronoi decompositions of the space have been used [5, 7, 22] as these decompositions flow naturally from the cell tower or WiFi router locations. However, these datasets are characterized by irregular boundaries and variable cell sizes, greatly complicating mathematical derivation of scaling properties. Instead, for tractability, we have chosen a regular grid approximation, which is more appropriately used when discretizing higher fidelity tracked datasets obtained through GPS trackers or smartphone locations [35,36,75,153]. Finally, we assume that paths are sampled regularly in time, again consistent with GPS tracking, rather than the stochastic data arrival associated with cellular call records. Because we assume that we are starting with a high-fidelity source like GPS traces, interpolation of locations between timesteps is not required.

As an agent traverses the discretized space, each location sample can be represented by a symbol corresponding to the label of the grid cell at the measured location. The symbols form a single dimensional string representing the agent's trajectory through the two dimensional space. The symbols are represented as letters in the examples for convenience. Because we assume a piecewise linear path through regular grids, sampled at regular intervals, we can begin to analyze how traversing these grids would appear. For a path parallel to either axis of the grid, the agent will emit a sequence of symbols characterized by repetition of the current grid cell. For constant velocity paths through multiple grid cells, this will lead to a uniform repetition of symbols, based on agent speed and cell size (e.g., 'AAAABBBBCCCCDDDD' for one speed and 'AABBCCDDEEFFGGHH' for an agent traveling twice as fast). However, if the path is not parallel to the grid cells' axes, then the agent may clip edges of cell (e.g. 'AAAAABCCCCC') changing the string and the entropy rate. As defining all possible arbitrary paths through cells is not mathematically tractable, we assume that agent must traverse the entire cell. This is the strongest assumption that we make, and the most likely to fail when applied to empirical data. This assumption has the additional impact of forcing paths to be bin-sized aligned; individual line segments must have a length that is an integer multiple of the bin size. Finally, we assume that each line segment traces a unique path through space, and crosses no other segment. While on the surface this seems like a

limiting assumption, made to facilitate derivation, we mean to eliminate strongly repeating trajectories, like orbits, which would significantly depress the entropy rate as calculated from the LZ approximation. We expect that crossing but non-overlapping paths, as investigated by Lee *et al.* [13], would have entropy rate approximations close to the unique path case, because while individual symbols might repeat, we would not expect to observe the repetition blocks of multiple symbols.

We limit the analysis to a sampling regime that will return sensible answers. Specifically, we consider regimes for bin width (resolution) and sample period in which scaling is meaningful.

Our assumptions can be summarized as follows:

1. *Path*: we assume that the path can be sufficiently well approximated as a series of line segments;

2. *Velocity*: we assume a non-zero constant velocity v for each line segment $dv_i/dt = 0$;

3. *Accuracy*: we assume that a given location measurement offers perfect accuracy, but relax this assumption in additional analysis;

4. *Measurement Density*: we assume that measurements are made with sufficiently high resolution devices so as to support a spatial decomposition into square bins of characteristic length W and a regular temporal sampling of period T, with no need for interpolation;

5. *Connectedness*: we assume that agents traverse the square bin or block in a classic four-connected manner, that is that participants only move in the cardinal directions though a block and traverse the entirety of the block, implying that the time to traverse a block is always W/v;

6. *Scale*: we consider a mesoscopic sampling regime with the following characteristics:

   (a) *Spatial*: the bin size is no bigger than the extent of the smallest line segment in the path.

(b) *Temporal*: no cells crossed by the path are skipped due to undersampling: $T \leq W/v$.;

7. *Independence*: we assume that each segment traces a unique and independent path from all previous segments. This assumption is necessary for tractability, but eliminates repetition (and, therefore, reductions in entropy rate) at an inter-path segment level. Repetition would decrease entropy rate, so we expect that this assumption pushes our derivation towards an upper bound;

8. *Termination*: we assume that each sequence of location symbols terminates with a unique symbol.

In the subsequent sections, we derive scaling behavior from the process of Lempel-Ziv compression, under the above assumptions. For readability, derivations are summarized in the main text. For detailed step-by-step derivations, please refer to Section 4.6.

4.2.2 Single Segment Derivation

We begin by considering a single line segment of length x traversed at constant velocity v parallel to one grid axis, then extend this to multiple non-overlapping line segments. The path requires $t = \frac{x}{v}$ time to traverse. Given our assumptions, the traversal of each grid cell will require at least one sampling period T and possibly more, resulting in one of more instances of each cell-symbol being emitted as the agent crosses the cell. Because the agent traverses each cell in its entirety, and in a four-connected manner, it takes the same amount of time to cross each cell. This results in a series of repeated symbols representing each of the cells that the segment passes through, where the number of repeats per cell is given by $L_b = \frac{W}{vT}$ and the total length of the string is $L = \frac{x}{vT}$.

The LZ-derived entropy rate of a string S of length L is given by the function

$$\lim_{L\to\infty} \left[\left(\frac{1}{L} \sum_{i=0}^{L-1} \Lambda_i \right)^{-1} \ln L \right] \tag{4.1}$$

as $L \to \infty$, where i is the index of a character in the string (with the first character being at $i = 0$), and Λ_i is the length of the minimum substring beginning at i such that this

substring has not previously been observed in the prefix of S terminating at position i, and L is the length of the string [7].

When scaling the spatial and temporal resolution for simplicity, we consider inter-sample periods given by $T = T_0 2^{-m} (m \geq 0)$, and bin sizes as $W = W_0 2^n (n \geq 0)$, where W_0 and T_0 are governed by our assumptions bounding the bin size and sampling rate.

The values T_0 and W_0 are not necessarily fixed constants, but instead vary with the parameters and the choice of v, x (for W_0) and T. Practically, there are bounds for each, given the method of localization employed, but in our formulation, W and T are parameters to some degree controlled by the experimenter, while x and v are properties of the observed agents.

Structure of the Sampled Sequence

Both the temporal inter-sampling rate T and the spatial scale W affect the structure of the sampled sequence. The sequence has a total length of $L = \frac{x}{vT}$ symbols, but is composed of $\frac{x}{W}$ blocks each consisting of $L_b = \frac{W}{vT}$ uniform repeating symbols. The number of symbols per block is an interaction between W, T, and v. Larger blocks take longer to traverse, leading to more repeated symbols. For $W = x$, the sampled string consists of a single, homogeneous, block of L symbols. For our lower bound of $W = vT$, this sampled sequence of length L consists of $\frac{L}{L_b}$ blocks, each consisting of a single unique symbol.

Because we assume non-overlapping paths, the binned values associated with different blocks are distinct. Because the sampled values within a given block are homogeneous, and because the sample value within the block is unique, the values of Λ_i all follow a regular pattern, *which depends only on the index within the block, and not on the index within the sampled string as a whole.* That is, we will have $\frac{L}{L_b}$ unique symbols and blocks, with each symbol repeating L_b times within its block. Thus, $\Lambda_i = \Lambda_{i \mod 2^n}$, given the structure of our downsampling.

We can thus decompose

$$\frac{1}{L} \sum_{i=0}^{L-1} \Lambda_i = \frac{1}{L} \sum_{b=1}^{\frac{x}{vT2^n}} \sum_{j=0}^{2^n-1} \Lambda_j \tag{4.2}$$

The terms in the outer sum (over b) correspond to the number of blocks, which is also

the number of unique symbols $\frac{x}{L_b}$. The index terms in the inner sum (over j) correspond to the number of repetitions in a block of length $L_b = 2^n$. To derive this sum, we consider two distinct cases: the positions in the first half of the block, and those in the latter half of the block. The pattern for the Λ_j in the first half of the block is a simple rising sequence. Regardless of the block, the first sample in the block (i.e., $j = 0$) is a unique character not previously seen in the string, and thus $\forall_{j=0}\Lambda_j = 1$. Similarly, for all blocks of length of at least 2, the second sample in the block concatenated with its following symbol (in this or the next block) has not previously been seen in the string, and thus $\forall_{j=1}\Lambda_j = 2$. Using similar reasoning, the lambda values continue to rise within the block up to the index of $j = \frac{2^n}{2}$. Thus $\forall_{j\leq\frac{2^n}{2}}\Lambda_j = j + 1$. That is, for indices up to the halfway point through the string, the substring starting at that point and including j additional subsequent characters (and thus of length $j+1$) consists purely of repetitions of the same character associated with this block, of successively larger lengths, and has not previously been seen. We consider now the cases of the Λ_j in the second half of the block, noting the assumption above of a unique terminating symbol following characters in the final block. For characters at indices just beyond the midpoint of their block (i.e., $j = \frac{2^n}{2} = 2^{n-1}$), there is a minimum unique string consisting of the character at that point, $\frac{2^n}{2} - 1 = 2^{n-1} - 1$ additional identical characters beyond that point lying within the same block, and then (additionally) the first character of the next block, thus yielding a unique total string length starting at position j of $2^{n-1} + 1 = j + 1$, as given by the formula above. For the indices in the following $2^{n-1} - 1$ positions of the string (i.e., for $2^{n-1} < j \leq 2^n - 1$), we are dealing with a strictly decreasing integer sequence, terminating in 2. This reflects the fact that, for index j, the uniform symbol prefixes beginning at index point j have all previously been seen within this block, and the smallest unique string consists of the prefix beginning at the current point (index j), proceeding through the end of the block, and including one character beyond the end of that block (which has not yet been previously encountered within the string). For a character at position j (zero-based) within the block, this yields a string length of $(2^n - j) + 1$. Thus, we have $\forall_{j>2^{n-1}}\Lambda_j = (2^n - j) + 1$. To summarize, Λ_j will be an arithmetic sequence, starting at 1, until just beyond the midpoint is reached; and then decreasing until the final value of 2 (e.g., $1, 2, \ldots, \frac{L_b}{2}, \frac{L_b}{2} + 1, \frac{L_b}{2}, \frac{L_b}{2} - 1, \ldots, 2$).

Given this per-block total, and that there are $\frac{x}{vT2^n}$ blocks, we have

$$\sum_{j=0}^{2^n-1} \Lambda_j = \sum_{j=0}^{\frac{2^n}{2}} (j+1) + \sum_{k=0}^{\frac{2^n}{2}-1} (k+1) = \frac{2^{2n}}{4} + 2^n.$$

Having a closed form expression for Λ_j and the equivalence in 4.2, we can now derive an expression for Λ_i.

$$\frac{1}{L} \sum_{i=0}^{L-1} \Lambda_i = \frac{vT}{x} \frac{x}{vT2^n} \left(\frac{2^{2n}}{4} + 2^n \right) = \left(2^{n-2} + 1 \right).$$

Substituting Λ_i into the equation for LZ compression-based entropy rate (4.1), the estimated entropy rate of the string is the following.

$$\begin{aligned} H(W,T) &= \left(2^{n-2} + 1 \right)^{-1} \ln \frac{x}{vT} \\ &= \frac{\ln \frac{x}{vT}}{(2^{n-2}+1)}. \end{aligned}$$

Because the number of symbols is related to the width of the cell and sampling rate, and as we have assumed the minimum width $W_0 = vT$ to ensure at least one sample per cell

$$H(W,T) = \frac{4W_0 \ln \frac{x}{vT}}{(W + 4W_0)},$$

and, therefore,

$$H(W,T) = \frac{4 \ln \frac{x}{vT}}{\frac{W}{vT} + 4}, \tag{4.3}$$

where x and v are independent properties of the path in question, and W and T are parameters that are intrinsic to the methods and apparatus of a particular experiment. The existence of a scaling law containing only four terms, two controlled by the experimenter, and two determined by the path, is one of the key findings of this work.

While choice of units will affect the size of the x, v, T and W_0 terms, we note that the governing terms $\frac{x}{vT}$ and $\frac{W}{vT}$ are distinguished by being of unit dimension; thus *the entropy rate expression is also of unit dimension, and invariant to unit change*. The first of these expressions is the total length of the sampled string; the latter is the number of samples required to cross a bin. This result suggests that for a single line segment, the entropy rate of strings sampled at different resolutions according to bin widths W and temporal inter-sample spacing of T should scale proportional to $O\left(\frac{4 \ln \frac{x}{vT}}{\frac{W}{vT}+4}\right)$.

Somewhat counter-intuitively, the entropy rate for a sequence of non-overlapping line segments of total length x, which are traversed in four-connected manner, is identical to the single line segment derivation above. Consider two cases: a single line segment of length x, and a snaking series of line segments also collectively of length x, which are selected in four-connected manner, but randomly picking a non-overlapping direction at every bin. The single segment linear path induces a string containing $\frac{L}{L_b}$ unique symbols, each repeating L_b times, as described above, and is, therefore, described by (4.3). The snaking path induces a string with exactly the same structure. Each transit of a bin produces L_b symbols. At the end of each bin transit, a new batch of L_b symbols begins, starting with a never before seen character. At the end of the path, in accordance with our assumptions, a unique symbol is emitted. This applies to any mixture of line segment lengths traversed at constant velocity, as long as they are multiples of W, and do not overlap. Any set of paths that generate a repeating structure like the structure for a single line segment will exhibit entropy scaling behavior described by (4.3). Intuitively, the straight line trajectory should have a lower entropy rate than the snaking trajectory because the trajectory can be described by a simple mathematical function. However, the entropy rate of the sequence is evaluated independently of the rule used to generate it. This apparent incongruence between the apparent and actual entropy rates for trajectories is subtle, and outside the scope of this work. However, a further investigation into the role of context into human mobility entropy rate estimation, along the lines of Smith *et al.* [21], appears warranted.

This formulation extends to any number of dimensions as long as the decomposition of that space is a hypercube, and transiting of the hypercube happens hyperface to hyperface along equidistant paths across the hypercube, which is essentially the higher-dimensional generalization of the four-connected path we have assumed. Because the compression — and, therefore, the entropy rate calculation — happens only on the trajectory, which is a single dimensional manifold, as long as the structure of the symbols generated by the trajectory remains the same, the above analysis will hold, and the scaling law will apply. In the case of higher dimensional spaces, W is the single dimensional edge length of the hypercube, and v is the velocity through the hypercubes. Because opposite faces of a hypercube will be W distance apart, by definition, the straight line trajectory through a hyperspace will have

the same symbol structure, and, therefore, the same entropy rate scaling behavior as above. Since there also must exist a path of distance W between adjacent faces of the hyperplane, the non-overlapping path argument above also applies. Therefore, (4.3) holds in general for spaces of arbitrary dimension, decomposed as hypercubes, for non-overlapping paths.

The scaling law exhibits some degree of upper-boundedness against some, but not all, of the assumptions. In particular, paths characterized by repetition will decrease the overall entropy rate by introducing inter-block repetition, that LZ will detect and compress. Violations of the scale assumptions will also decrease entropy, as bin sizes larger than the smallest line segment will cause line segment concatenation with a cell, and, therefore, longer repeating blocks. Similarly, skipping cells due to undersampling will not increase the entropy, as a maximal condition of each symbol in the string being new and unique will already have been reached. However, the addition of noise can disrupt the sequences described here, potentially increasing entropy rate, as expected for additive noise processes. Allowing non-four-connected paths could also increase the entropy in some cases, particularly as cell size increases and clipping becomes more likely, although whether the entropy rate increases or decreases is dependant on the interaction of path and spatial discretization.

Scaling Law Behavior

When proposing scaling laws, it is often useful to examine their limiting behavior. The proposed law is well behaved in the limits for the experimenter controlled parameters. As $T \to 0$, while the length of the string increases, each bin will also be sampled by an ever larger number of repetitions and the entropy rate goes to zero. By contrast, the limit of $H(W, T)$ as $T \to \infty$ is negative infinity. However, this bound does not make sense semantically, because it represents the entropy rate of mobility patterns which are never sampled, which violates our assumption about sampling. As $W \to 0$, the entropy rate tends towards a maximum value $\ln \frac{x}{vT}$, which represents the log of the number of symbols sampled, or the entropy rate of a series of distinct symbols of the given length. As $W \to \infty$, the entropy rate approaches zero, which is sensible, as the entire string would consist of a repetition of the same location symbol.

The proposed law is also well behaved in the path description parameters. As $v \to 0$,

$H(W,T)$ also goes to zero, as we have a path composed of a single repeating symbol. As $v \to \infty$, (putting aside relativistic effects), the entropy rate goes to negative infinity, which, as in the case of T, corresponds to a path that is never sampled, and violates our assumptions about sampling. At a minimum, L must be at least one, or there is no string, and LZ will return the compression of a single symbol, likely a poor approximation of the entropy rate. As the string becomes infinitely long, with an infinite number of distinct blocks, the entropy rate approaches infinity, as would be appropriate.

A natural question is whether the scaling law has any maxima or minima with respect to W or T, as this would imply sampling regimes which might be considered optimal. This behavior can be investigated using the partial derivatives. The partial derivative of $H(W,T)$ with respect to W is

$$\frac{\partial H}{\partial W} = -\frac{\frac{4}{vt} ln(\frac{x}{vT})}{(\frac{W}{vT}+4)^2}. \tag{4.4}$$

The derivative does not have a root with respect to W, so there are no minima or maxima along the W axis for the scaling relationship, implying that no sampling dimension is preferred. Examining the partial derivative of the entropy rate scaling with respect to T yields

$$\frac{\partial H}{\partial T} = \frac{4vW + 16vT - 4vW ln(\frac{x}{vT})}{(4Tv+w)^2}, \tag{4.5}$$

which has a sequence of roots for a given (v, W, x) at

$$T = \frac{W}{4v}\boldsymbol{W}(\frac{4x}{eW}), \tag{4.6}$$

where e is the natural basis and $\mathbf{W}$ is the Lambert W function, which is not solvable analytically, but is readily approximated numerically. This function is defined for $W > 0$ and $v > 0$, which is strictly true in our formulation, as W is a distance, and v is a ratio of distance and time. This implies that for certain values of (x, v, W), there exists a sampling rate corresponding to maximum entropy rate. Sampling beyond this rate will lead to repetition, decreasing the entropy rate. Sampling below this rate will result in removing information, also lowering the entropy rate. This finding is a central outcome of the scaling law, as it

implies that there exists an optimal temporal sampling regime for a given spatial resolution and mobility pattern.

4.2.3 Entropy Rate of Paths with Mixtures of Velocities

While the previous section derived the scaling behavior of the entropy rate of a non-overlapping piecewise linear path, this analysis is unnecessarily limiting for practical application. We seek here to derive an entropy rate for a sequence of non-overlapping line segments traversed with varying velocity. Considering non-overlapping paths as before, (4.3) provides a starting point to examine how entropy rate might sum for non-overlapping paths of straight line segments through space.

We begin by noting that changes in speed undertaken between two samples occuring within the same spatial bin are not observable, being below the spatial sampling rate. The number of symbols emitted when transiting the cell is proportional to the time it takes to cross the cell, divided by the sampling rate. The time taken to cross the cell can be trivially represented as the width of the cell divided by the average speed within the cell, from the definition of average speed ($\bar{v_c} = \frac{W}{T}$). Given that speed changes within a cell are averaged by the emission of symbols, we need only concern the derivation with inter-cell velocity variability.

Given the same linear four-connected path, covering a distance x, consider the case where a fraction α is made at velocity βv, and fraction $(1-\alpha)$ is made at velocity γv, yielding a time-averaged velocity of

$$\bar{v} = \frac{v}{\left(\frac{\alpha}{\beta} + \frac{(1-\alpha)}{\gamma}\right)}.$$

The string length is

$$L' = \frac{\alpha x}{\beta v T} + \frac{(1-\alpha)x}{\gamma v T} = \frac{x}{\bar{v}T}. \tag{4.7}$$

The total entropy rate is then (step-by-step derivation is provided in Section 4.6) the following:

$$\left(\frac{1}{L'}\sum_{i=0}^{L'-1}\Lambda_i\right)^{-1}\ln L'$$
$$=\left(\frac{1}{L'}\left(\sum_{b=1}^{\frac{\alpha x}{\beta v T 2^n}}\left(\frac{2^{2n}}{4}+2^n\right)+\sum_{b=1}^{\frac{(1-\alpha)x}{\gamma v T 2^n}}\left(\frac{2^{2n}}{4}+2^n\right)\right)\right)^{-1}\ln L'$$

and, therefore,

$$H(W,T)=\frac{4\ln\frac{x}{\bar{v}T}}{\frac{W}{\bar{v}T}+4} \tag{4.8}$$

which is the same expression as in equation (4.3), but including time averaged rather than constant velocity. This derivation is generally valid, subject to bounds on the velocity which maintain that at least one symbol per cell must be recorded, and no cells can be skipped by changing velocity.

4.2.4 Impact of Spatial Uncertainty

As most entropy rate calculations of interest will be performed on empirical data, it is important to consider the impact of measurement noise on scaling behavior. If measurement noise dominates, then the scaling behavior described here is of limited utility. However, if the measurement noise has well-behaved statistical properties, it may be possible to derive an expected entropy rate considering these impacts. We seek here to consider the effects of spatial noise on the entropy rate estimates, as we expect timing estimates to be much finer grained than human motion. We assume a GPS-like positioning system, with position error estimates that are normally distributed around the true value μ with standard deviation σ, employing the classic zero mean Gaussian noise model. The probability that a given measurement (a sample from that distribution) lies *further* than distance d from the mean is given by $1-erf\left(\frac{d}{\sigma\sqrt{2}}\right)$.

Now consider taking a measurement at the center point of a generic square bin of physical width W. The probability, p, of a measurement lying outside the distance to the boundary

$\left(\frac{W}{2}\right)$ — and, thus, returning an erroneous spatial bin, and associated symbol — is given by the following.

$$p = 1 - erf\left(\frac{W}{2\sqrt{2}\sigma}\right), \tag{4.9}$$

where draws from this distribution are considered independent

By incorporating the above noise model, and applying a number of further assumptions, the entropy rate can be approximated as follows (step by step derivation is provided in Section 4.6):

$$H(W,T) = \frac{\ln \frac{x}{vT}}{\frac{1}{p} + \frac{1}{pL_b}\left(1 + 2\left(\frac{(1-p)\left((1-p)^{\frac{L_b}{2}} - 1\right)}{p}\right) - (1-p)^{\frac{L_b}{2}}\right)}. \tag{4.10}$$

Recall that $L = \frac{x}{vT}$ and $L_b = \frac{W}{vT}$, where the total path length is x, physical bin width is W, the velocity is v, and inter-sampling period is T. We can further expand (4.10) by substituting $\frac{W}{vT}$ for L_b, and (4.9) for p. If the agent travels distance x with a mixture of velocities, v in (4.10) gets substituted by the time-averaged velocity $\bar{v}$.

Erroneous symbols generated through noise processes may come from a bin traversed earlier in the trajectory, a bin that will be traversed later in the trajectory, or from a bin that will not be encountered by the trajectory. While the occurrence of an erroneous reading in either of the first two categories will yield repetitions (thus, preventing the relevant substrings from being entirely unique), an occurrence of the latter will not. Specifically, we believe that it is considerably more likely that the formula in (4.10) will underestimate the entropy rate in practice, as large enough noise to be effective will disrupt the repetition of symbols, and, therefore, increase entropy rate. However, it is possible to imagine pathological behavior where noise would, for the entire duration it takes to traverse a bin width W at v, perturb the measurement in the direction of the next bin on the trajectory, returning a double length sequence of symbols and thus decreasing the entropy rate. However, for a symmetric error distribution like a Gaussian, we anticipate that this behavior should be rare.

Fig. 4.1 compares the entropy rate measures with (generally top) and without (generally bottom) noise for $5 <= W <= 200$, $0.5 <= T <= 10$, $\bar{v} = 1$, and $x = 1000$. Absent noise, the entropy rate is generally lower over wide ranges of medium and large spatial scales and

sampling periods when compared with the estimate of entropy rate with noise. However, at small physical scales and longer sampling periods, the entropy rate without noise can lead to sequences of entirely unique symbols, whereas there is some repetition in the presence of noise — and, therefore, somewhat lower entropy rate. Assuming a standard deviation of 30 m for GPS, these two entropy rate estimates exhibit a high degree of disparity, particularly for physical scales of around $40 - 80$ m. By contrast, the entropy rate estimates with and without noise approach each other asymptotically as the spatial aggregation scale increases, as expected.

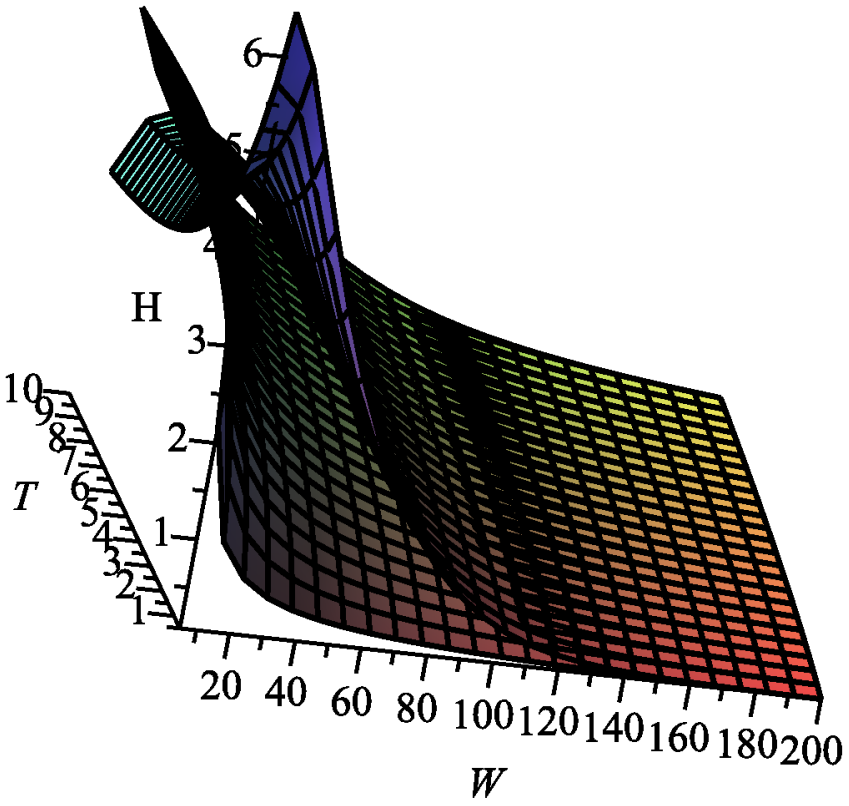


Fig 4.1: Entropy rate measures with (generally top) and without noise (generally bottom)

4.3 Methods

To provide a semi-empirical validation for the model, we compared the results of the theoretical model with the results from two widely employed and stylized simulated models of human

mobility. A single agent traversed a simulated field with a constant speed (v) while following the employed motion models, and agent locations on the grid were recorded according to the spatial and temporal sampling rates. The maximum and minimum sampling periods were set to 512 s and 1 s, respectively. We collected 64 samples for $max(T) = T_0 = 512$ s; therefore, making the number of samples 64×2^m for $T = T_0 2^{-m}$. To collect 64 samples at $T_0 = 512$ s, the agent in the theoretical model had to traverse $64vT_0 = 65536$ m. For other models where the agent moves in a square field, we set the diagonal length of the field to $64vT_0$ to make their comparison with the theoretical model sensible. The minimum value of W for a combination of v and T is vT, and the maximum value of W is $64vT_0$. Each model was applied with and without power law distributed dwelling at nodes, and (for each such variant) with and without additive noise. The two empirical motion models are the following:

- *Random Waypoint Motion Model*: in this model, 100 unique waypoints were drawn uniformly from the field described above. The waypoints described a fully connected graph; that is, the agent could go from a waypoint to any other waypoint. This allows crossing paths, which we assumed absent in the theoretical derivation for simplicity. Transitions from one waypoint to another were drawn uniformly. However, because waypoints were drawn uniformly, the probability of repeated path sequences was low. We investigated transitions with and without dwell time. For transitions with dwell time, dwell time was drawn from a power law distribution with the exponent of -1.8 and maximum dwell time was set to 17 hours, consistent with [5].

- *Power Law-based Motion Model*: in this model, the agent selected an angular direction uniformly from a set $\{5k^\circ : k \in \mathcal{N}^+ \text{ and } 5 \leq 5K \leq 355\}$, and drew the distance for the next step from a power law distribution, which is typically observed in empirical datasets (e.g., [5]). Draws were constrained to ensure that the agent remained in the field. The distance was limited to 0.8 times the characteristic length of the field. Movement directions were resampled until a destination inside the field was generated. In these experiments, -1.55 was chosen as the power law exponent, consistent with reported empirical findings [5]. For the dwell time variant, we employed the same

distribution as for the Random Waypoint model.

We also considered an additive measurement noise model. Each of the above scenarios was run once without any additive noise and once for the noise model. Simple zero mean Gaussian additive measrement noise model was considered, consistent with simple noise models of GPS location measurements. Noise was added to the signal after the agent moved but before simulated measurement took place. A moderate ($\sigma = 10m$) noise level was selected consistent with commodity GPS systems. A theoretical entropy rate was calculated from (4.3), and compared to the empirical measurement calculated according to (4.1).

Several aspects of these simulated motion models depart from the assumptions made when deriving our scaling law. First, each model permits crossing paths, leading to repeated symbols, although are unlikely to produce cyclic paths. Second, we have included variants which include measurement noise and dwelling, neither of which are explicitly accounted for in (4.3). Third, the models can lead to clipping effects explicitly ruled out when deriving 4.3.

Given that the paths were generated in simulation, we have precise control over the sampling rates, bin widths, path length and agent velocity and can, therefore, explicitly calculate the scaling law, and compare them against the Lempel-Ziv derived entropy rates from the trajectory records. Employing bin widths of $W = W_0 2^n = vT2^n$, we can simplify (4.3) into (4.11).

$$H(W,T) = \frac{4ln(L)}{2^n + 4}. \tag{4.11}$$

We use the coefficient of determination (R^2 metric) to understand how well the theoretical curves fit with those from the empirical simulation models, including the model that applies (4.1) to the sequences of the theoretical model. The definition of R^2 is given in (4.12), where $f_1, f_2, \ldots, f_n$ are the predicted values for $y_1, y_2, \ldots, y_n$. R^2 values were calculated in R software environment.

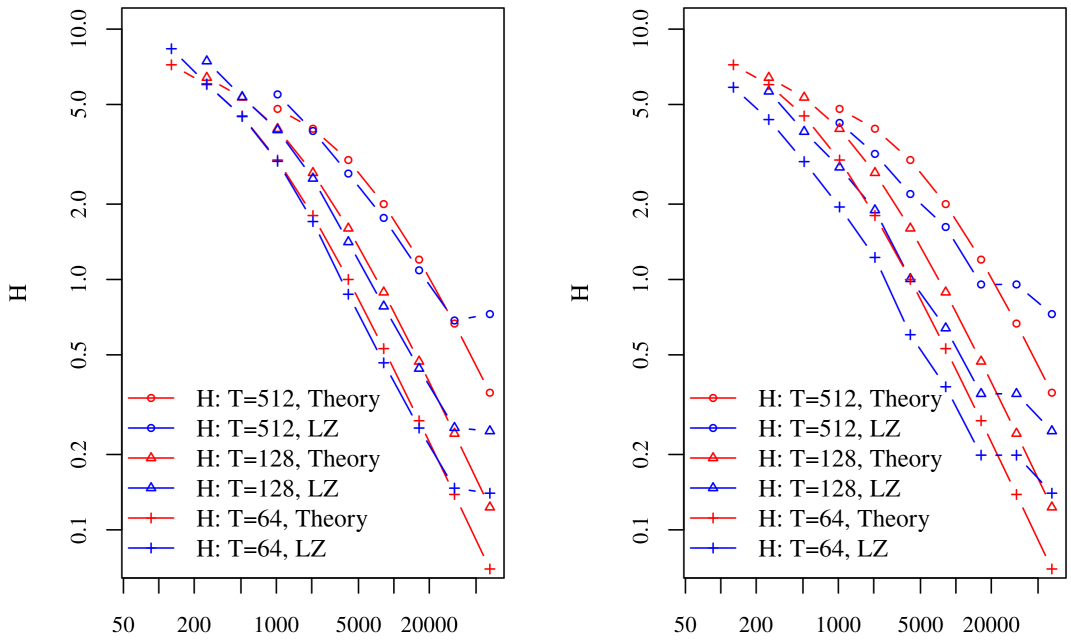
$$R^2 = 1 - \frac{\sum_i^n (y_i - f_i)^2}{\sum_i^n \left(y_i - \frac{1}{n}\sum_i^n y_i\right)^2}. \tag{4.12}$$

We ran the simulations on a Linux-based computing cluster with 96 computational nodes, each having 2 x eight-core Intel E5-2650L (1.8GHz) or Intel E5-2640L (2.0 GHz) Xeon Processors, and 32GB RAM. Jobs were submitted to the cluster through the Torque scheduler. Refer to supporting information *S1_Data* of [158] for the relevant data and code required to generate the data.

4.4 Results

We seek to determine how well the scaling law behaves when compared to paths without non-Gaussian measurement noise, participant non-compliance and other effects that may be present in empirical data. These might obfuscate the underlying behavior, and make comparisons more difficult. Some of the simulated systems here are noise free, but do allow for repeating symbols and cell clipping. Analyzing the behavior of these simulated systems against the theoretical scaling model could provide insight into the impact of breaking these key assumptions on the proposed scaling law's predictions.

Fig. 4.2 presents the comparison between the theoretical model and power law-based models with and without dwelling, and with no added measurement noise in the sequences. In the model without dwelling, the scaling law provides exceptional agreement with the simulation. At very large W, the empirical entropy rate exceed the theoretical, as clipping effects begin to dominate. As the bin width increases, more repetitions occur in the string. Therefore, entropy rate goes down. The theoretical model considers regular patterns of strings. However, because of the stochastic nature of empirical strings, the effect of large bin width may be less dominant in lowering the entropy rate than is the case for the theoretical model. This is why the entropy rate of the empirical models in Fig. 4.2 for large W exceeds that of the theoretical model. As an example, consider two 64-character strings from the alphabet {'0', '1'}, which are expressed, using regular expression, as /0{32}1{32}/ and /1{3}0{31}1{30}/. Here, the second string has a higher entropy rate. The first string has the structure assumed by the theoretical model, while the second indicates a clipped trajectory. The latter may appear as the representation of a trip, at a large bin width, which is derived from power law-based trip segment lengths and dwell times.



(a) Entropy without Dwelling

(b) Entropy with Dwelling

Fig 4.2: Theoretical model generated sequence entropy rate Vs. LZ entropy rate of sequence obtained from power law models.

Fig. 4.3 presents the comparisons between the theoretical model and the noise-free random waypoint-based models with and without dwelling. Similar to the power law based empirical model, entropy rates at large bin widths exceed those of the theoretical model. However, the effect of dwelling is less pronounced than power law-based models, because fewer constraints were placed on the trip length in the random waypoint model. The trip segments, therefore, were longer and fewer trip segments (2 to 5 segments as compared to 186 to 292 for the power law model in the conducted experiments) were required to obtain the desired numbers of location samples. This resulted in fewer dwell occurrences in the random waypoint model than their power law counterparts. The theoretical model shows admirable agreement for the entropy rate scaling behavior for both synthetic mobility models. Deviation from theoretical behavior is apparent for very small and very large values of W.

To show the effects of added measurement noise to the power law and random waypoint

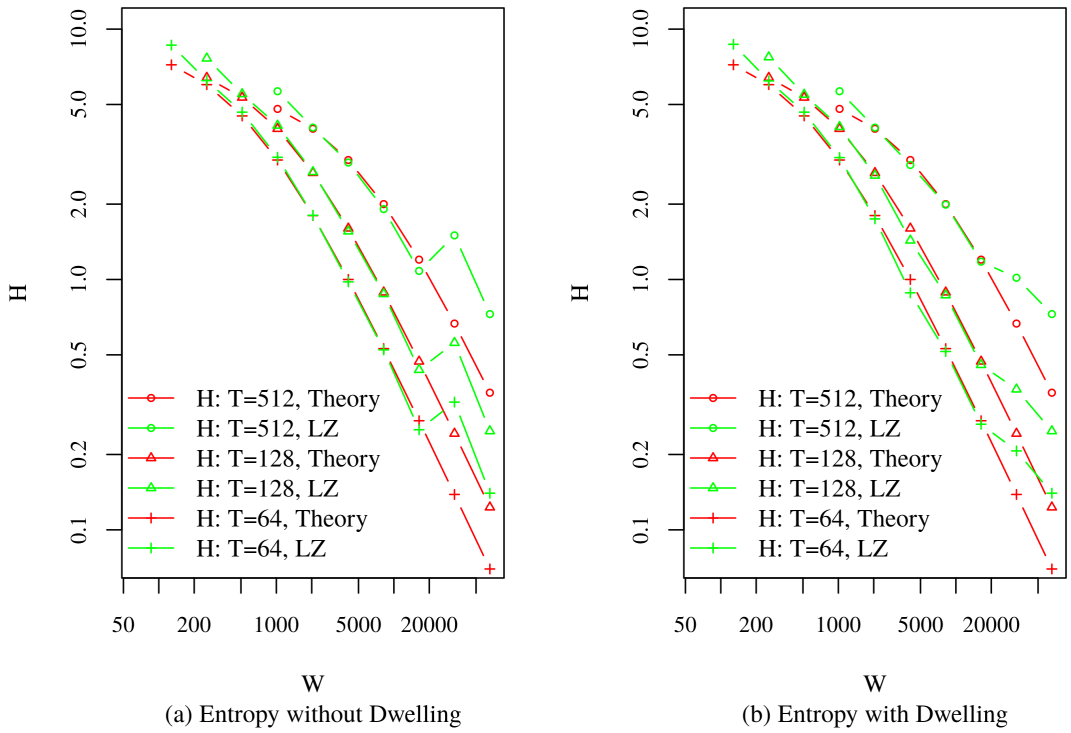
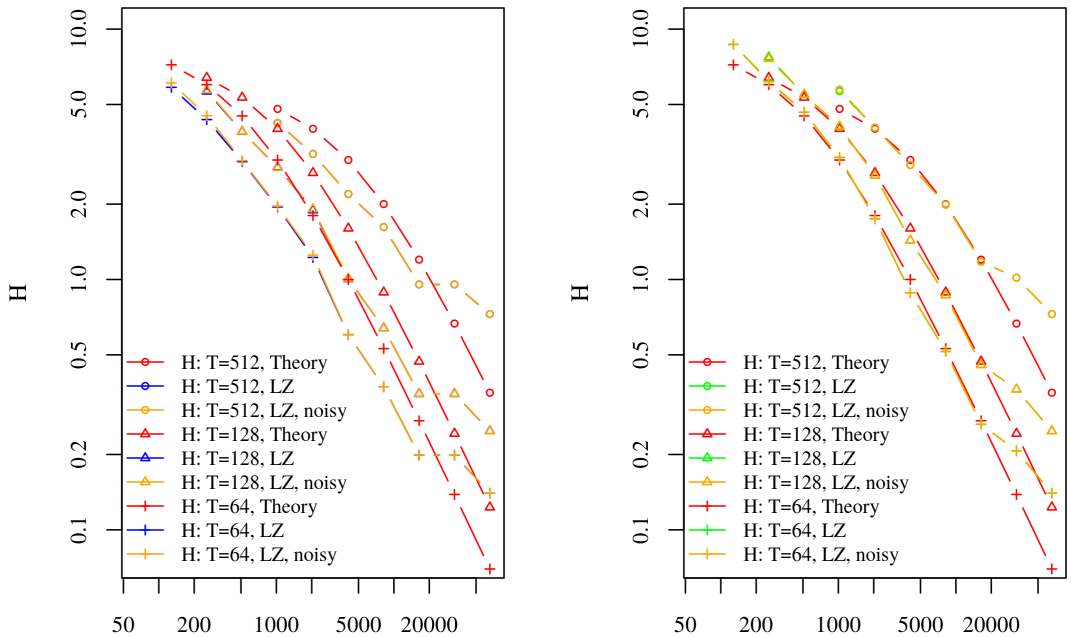


Fig 4.3: Theoretical model generated sequence entropy rate Vs. LZ entropy rate of sequence obtained from random waypoint models.

based models on entropy rate, Fig. 4.4 presents the entropies of the sequences obtained from these models, with dwelling enabled, alongside the entropies of their noisy versions for $\sigma = 10m$, a value typical for consumer GPS systems. Fig. 4.4 shows that the introduced zero mean Gaussian noise does not significantly alter the entropy rate, particularly as grid size increases. The probability that a given measurement falls outside the current grid cell, given the accuracy of GPS systems, is small for the sizes of cells considered. Smaller cells would be more susceptible to noise deviations, and might show greater impact on entropy rate, but that impact would be predominantly sensor noise and not the phenomenon of interest. While compensating for noise using more complex models such as (4.10) may be possible, a simpler solution in some circumstances would be to use bin sizes larger than the expected error, but that still capture the phenomenon of interest.

Fig. 4.5 compares the curves generated by the theoretical and simulation models. For each simulation model, we compare the curves, relating entropy H to W for different values



(a) Power Law Model (b) Random WP Model

Fig 4.4: Theoretical model generated sequence entropy rate Vs. LZ entropy rate of power law and random waypoint models with and without noise, and with dwelling

of T, with the corresponding curves of the theoretical model. Each boxplot in Fig. 4.5 is generated with the R^2 values of fitting the theoretical curves to the curves of the simulation models over all T. All but the power law with dwelling model show exceptional fit quality (in excess of 0.9), and even the poorer fitting models have an R^2 of about 0.8. The shortcomings of the R^2 metric on non-linear models notwithstanding, these results provide us with additional confidence in the fit quality visually evident in the previous figures.

4.4.1 Explanation of Results

The theoretical model provides a surprising degree of agreement with the synthetic mobility models, suggesting that the mechanics of compression have a great deal to do with the scaling behavior reported in the literature. Our derivation indicated that, subject to our assumptions, the scaling model should form an upper bound on the entropy rate, as any deviations from a unique straight line path would reduce repetition in the string, and,

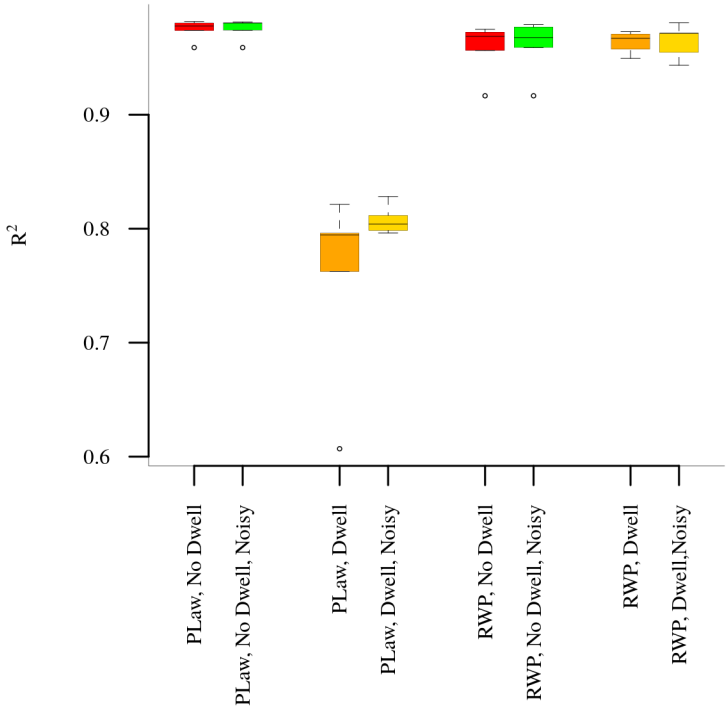


Fig 4.5: Fitness of theoretical curves to simulation models.

therefore, increase the entropy rate. However, when the theory deviates from the prediction, it almost always underestimates the entropy rate calculated from Lemple-Ziv compression. This is primarily due to violations of two of our assumptions, made to make the mathematics tractable.

First, while we assumed a unique termination character during our derivation, we did not supply a unique termination character at the end of strings built from the simulation. This has the counterintuitive result of increasing the estimated entropy rate. Consider a sequence of four symbols. If all symbols are the same, $\sum_{i=0}^{L-1} \Lambda_i = 8$ under our assumption, compared to $\sum_{i=0}^{L-1} \Lambda_i = 3$ according to (4.1). Therefore, theoretical entropy rate drops faster than the LZ-entropy for larger W.

Second, we assumed that the agent traversed the entirety of each block that it encountered; however, this is not necessarily the case in empirical data. For example, a path which traverses cell A, clips cell C and traverses cell B could have a corresponding location string of 'AAAAAAAACCBBBBBB', whereas the theory implicitly assumes that the path must

be 'AAAAAAAABBBBBBBB'. While this assumption was reasonable at small W, at larger scales, real paths are less likely to transit in a four connected manner. This effect also demonstrates that there are representational effects in the compression calculation. With grid and travel path at arbitrary relative orientations, paths which clip the edge of a cell are possible, and increasingly likely with increasing cell size, increasing the entropy rate at larger scales beyond the theoretical prediction.

However, despite these shortcomings, the predicted values showed excellent agreement with the empirical values computed from LZ compression on simulated paths. These results are encouraging for extending our model to incorporate real empirical data, which is confounded by missing data, varying sample sizes and non-Gaussian noise processes. This model should provide a firm theoretical basis for continuing work to address the more difficult situations encountered in real data.

4.5 Discussion

In this chapter, we have described a methodology for estimating the differences in predicted entropy rates over different spatial and temporal scales, with and without Gaussian noise, grounded in the theoretical behavior of the Lempel-Ziv compression algorithm typically used to the calculate mobility entropy rate. We have demonstrated that scaling behavior is to be expected and is inversely proportional to the spatial scale, and proportionate to the logarithm of the sampling rate. From these derivations, we were able to demonstrate that there is a predicted sampling rate of maximal entropy rate, which can be calculated using the Lambert **W** function. This theoretical model was validated against models of simulated movement, and found to provide excellent fits for stylized results, but with declining impact at very large or small spatial scales where our assumptions begin to break down. These results are important for a number of reasons.

First, we establish a strong theoretical foundation for mobility entropy rate scaling behavior observed and reported by a number of other authors [21, 22]. Based on an analysis of the behavior of Lempel-Ziv compression on the kinds of strings created by agents moving through space, we were able to demonstrate that the mobility entropy rate scaling behavior

could be described with only four terms: the length of the path, the average velocity of the agent, the width of the spatial bin, and the period of the sampling rate. Since scaling law encodes both parameters related to agent motion (x, v) and experimental design (W, T), we can conclude that the scaling depends both on agent behavior and the mathematical realization of that path. This finding is important, as it indicates that the scaling behavior encodes the mobile agent's behavior, and is not purely an artifact of mathematics, and, therefore, is itself a potentially useful metric. This finding also opens a clear opportunity to separate the two components of entropy rate scaling, providing the ability to isolate the behavioral fingerprint represented in the data.

Second, the scaling law is general, subject to the assumptions. As the trajectory compressed using Lempel-Ziv itself is a single dimensional manifold, as long as the space decomposition and path definition is analogous to the four-connected path described in the assumptions, the scaling law is valid. Similarly, because LZ compression does not distinguish between symbols, only symbol order, any non-overlapping path that crosses the entirety of a cell along only cardinal directions is also valid. We note that while describing the trajectories of people was our primary motivation, this derivation applies to the trajectory of any agent moving through space, subject to our assumptions.

Third, the structure of the equation indicates that the differences matter. As shown in the results and in previous works [21, 22], changing the scale of measurement can have a significant impact on the resulting entropy rate calculation. Directly comparing mobility entropy rates from experiments with differing spatial and temporal resolutions is not meaningful. Estimates of entropy rate at a common spatio-temporal resolution, either using the upper bound estimate here, or through an empirical estimate, would be required. This outcome is particularly important for spatial scale, as it implies that the results for studies with heterogeneous cell sizes may be confounded by scaling effects, particularly if the frequency of visits to cells of different sizes is significantly different for different participants.

Finally, the scaling law has a maximum value with respect to T, implying that there is a preferred sampling rate for a given spatial and velocity profile. This is an obvious point to use as a common comparator between datasets. Datasets with similar entropy rate maxima will likely have more similar scaling properties than those that do not. This property is

also potentially useful for researchers designing data collection studies, as they could use anticipated average velocity, trip length and spatial bin size to identify a preferred sampling period T.

4.5.1 Limitations and Future Work

The primary limitation in this work is the set of assumptions which made the theoretical analysis tractable. By assuming that the agent was always in motion, and that the path contained no repetitions, and through use of a simple noise model, we have constrained the generalizability of the findings. However, the model matched well against simulated systems, and is relatively straightforward to calculate. The primary goal of any future work should be to extend our results to encapsulate a more broadly representative model of human mobility and noise processes. The second major limitation of our assumptions was that the discretization of space was based on equally dimensioned square grid cells. While this is a reasonable assumption, in practice, researchers have employed cellular tower records to provide the discretization of space (e.g. [5]), leading to a distribution of cell sizes based on the Voronoi diagram of the cell towers' spatial configuration. The irregularity of the cell tower configuration could potentially exacerbate cell clipping effects, and make the entropy rate dependent on the path the agent takes though the cell. A more sophisticated analysis treating both cell shape and path orientation as independent random variables might address these issues; however, that analysis requires a substantial additional body of research. Similarly, time scales from call records are not constant and depend on individual calling patterns. Extending our work so that spatial resolution and sampling rate can also be represented as random variables would be an important step forward. Finally, we validated our scaling law against simulated mobility models. The model provided surprisingly good fits given the strength of the assumptions, and the fact that both simulated systems violated those assumptions. However, the stylized mobility models employed, while popular, have been shown to be imperfect representations of human mobility [13, 118]. It is a priority to validate the scaling law against actual mobility data.

Concluding Remarks

The findings presented here provide a theoretical explanation for the scaling behavior observed in calculations of mobility entropy rate from strings of locations using Lempel-Ziv compression. These results, while based on stylized assumptions, provided a useful approximation of scaling behavior for a wide variety of simulated paths, knowing only the average velocity, even under simulated sensor noise. The theory and simulated results provided close agreement for a wide range of spatial and temporal sampling scales, only breaking down at relatively large (corresponding to long repetitions of single symbols) or very small (corresponding to strings of unique symbols) spatial scales, indicating that our assumptions are plausibly valid. The entropy rate scaling formulation has a maximum at a particular sampling frequency, implying that optimal sampling regimes for given trajectories should exist and are in principle approximatable. This work is an important step in transforming mobility entropy rate from a scientific curiosity into a reliable workhorse of modern mobility and spatial behavior studies. By extending this work to emprical data and less stylized mobility assumptions, a scale-free mobility entropy rate formulation may be derived.

4.6 Supplementary Material: Detailed Scaling Law Derivation

In this supplementary document, we supply the line-by-line derivations to accompany the main results and descriptions of this chapter, subject to the same assumptions. For readability of the detailed derivations, many of the key equations and arguments have been reproduced here. Our goal is to determine the spatiotemporal scaling behavior for Lempel-Ziv compression, according to the following equation.

$$H = \left(\frac{1}{L} \sum_{i=0}^{L-1} \Lambda_i \right)^{-1} \ln L \tag{4.13}$$

The entropy rate of a string S of length L is given by (4.13) as $L \to \infty$, where i is the index of a character in the string (with the first character being at $i = 0$), and Λ_i is the length of the minimum substring beginning at i such that this substring has not previously been observed in the prefix of S terminating at position i. Now consider the sequences of characters of that string S resulting from sampling the agent's location along a 1D trajectory at different levels of spatial and temporal resolutions.

4.6.1 Ranges of Spatial and Temporal Resolution

For simplicity, we consider spatial and temporal sampling rates which scale by powers of two. For each spatial scale, we consider only temporal inter-sampling period regimes in which at least one sample will be measured within the time for the agent to traverse the distance x; that is, in general, $T \leq T_0 = \frac{x}{v}$, which is the upper limit (T_0) of inter-sampling time, as shown in (4.14).

$$T_0 = \frac{x}{v} \tag{4.14}$$

and, for the general case within this range,

$$T = \frac{T_0}{2^m} \qquad \text{where } m \in \mathbb{N}^0. \tag{4.15}$$

The spatial bins are bounded by the following assumptions.

Upper bound: $W = x$**.** At and above this level of spatial scale, all samples from the path are mapped to the same bin.

Lower bound: $W = W_0 = vT$**.** Below this spatial scale, we begin missing transited cells due to undersampling, and all sampled blocks remain unique.

$$W_0 \ = vT = v\left(\frac{T_0}{2^m}\right) \tag{4.16}$$

and, for the general case within this range,

$$W \ = W_0 2^n \qquad \text{where } n \in \mathbb{N}^0 \tag{4.17}$$

4.6.2 Structure of the Sampled Sequence

Both the temporal inter-sampling rate T and the spatial scale W affect the structure of the sampled sequence. At the most fundamental level, the length of the sampled sequence representing the trajectory varies with temporal inter-sampling period T, being given by $L = \frac{x}{vT}$ characters. Moreover, the internal structure of the sequence will differ across both temporal resolutions T and spatial resolutions W. An important insight is that this sequence itself consists of a series of uniformly sized *blocks*, each composed uniformly of a repeated occurrence of a single unique character.

Both the spatial scale and the temporal sampling rate strongly impact this structure. For $W = x$ (i.e., a bin width equal to the total path length), the sampled string consists of a single and homogeneous block of length $L = \frac{x}{vT}$ characters. For the lower mesoscopic bound of the bin size $W = vT$, this sampled sequence of length L consists of $\frac{x}{vT}$ blocks, each of length 1 and consisting of a unique sampled character (reflecting the fact that at the maximal resolution, the successive samples all fall into distinct bins). In general, for a specific temporal scale (associated with inter-sampling time T) and the resolution associated with bin width $W = W_0 2^n$, the sample string of length L will consist of $\frac{x}{W}$ successive blocks, each of length $\frac{W}{vT}$, and each consisting purely of repetitions of one sampled value - the bin into which all of the sampled locations within that block fall; and each such block will be associated with a unique such sampled value. As alluded to above, the minimal spatial scale

yielding a change in entropy rate is $W = vT$; below that level of scale W, entropy rate will remain invariant, as the length of the sequence depends only on temporal scale T, and the sampled values will remain unique and equal in number. For a given level of temporal scale T, we thus can specify $W_0 = vT$, and consider spatial scaling at successive binary powers n of that minimum scale. Thus, for a given n, $W = W_0 2^n$, and we will have N_b blocks as shown in (4.18), each of length L_b characters, as shown in (4.19).

$$N_b = \frac{x}{W} = \frac{x}{W_0 2^n} = \frac{x}{vT2^n} \tag{4.18}$$

$$L_b = \frac{W}{vT} = \frac{W_0 2^n}{vT} = \frac{vT2^n}{vT} = 2^n \tag{4.19}$$

Because of our assumptions of $1D$ trajectories and (to this point) constant speed, the binned values associated with different blocks are distinct and the sampled values within a given block are homogeneous. Therefore, the values of Λ_i all follow a regular pattern, *which depends only on the index within the block, and not on the index within the sampled string as a whole*. Thus, $\Lambda_i = \Lambda_{(i \mod 2^n)}$. We can thus decompose the sum over the entire string ($\sum_{i=0}^{L-1} \Lambda_i$) into nested sums over blocks b and indicies i within each such block, as in (4.20):

$$\begin{aligned} \frac{1}{L}\sum_{i=0}^{L-1} \Lambda_i &= \frac{1}{L}\sum_{i=0}^{L-1} \Lambda_{(i \mod 2^n)} \\ &= \frac{1}{L}\sum_{b=1}^{\frac{x}{vT2^n}} \sum_{j=0}^{2^n-1} \Lambda_j. \end{aligned} \tag{4.20}$$

We now consider the total of the Λ_j values across a block, $\sum_{j=0}^{2^n-1} \Lambda_j$. Because the value of Λ_j depends only on the location of the block (i.e., $\Lambda_i = \Lambda_{(i \mod 2^n)}$), this sum over the Λ_j within a block is identical for different blocks. To derive this sum, we consider two distinct cases – the positions in the first half of the block, and those in the latter half of the block.

The pattern for the Λ_j in the first half of the block is a simple rising sequence. Regardless of the block, the first sample in the block (i.e., $j = i \mod 2^n = 0$) is a unique character not previously seen in the string, and thus $\forall_{j=0}\Lambda_j = 1$. Similarly, for all blocks of length of at least 2, the second sample in the block concatenated with its following symbol (in this or

the next block) has not previously been seen in the string, and thus $\forall_{j=(i \mod 2^n)=1} \Lambda_j = 2$. Using similar reasoning, the lambda values continue to rise within the block up to the index of $j = 2^n/2$ (zero-based). Thus $\forall_{j=(i \mod 2^n) \leq \frac{2^n}{2}} \Lambda_j = j + 1$. That is, for indices up to the halfway point through the string, the substring starting at that point and including j additional subsequent characters (and thus of length $j + 1$) consists purely of repetitions of the same character associated with this block, of successively larger lengths, and has not previously been seen.

We consider now the cases of the Λ_j in the second half of the block. Before discussing the handling of this case, we note that after the final block of the entire string of length L, we assume either a unique terminating character, or the starting character of the initial block, which has never previously been encountered following characters in the final block. We now turn to discuss the characters in the latter half of blocks in general. For characters at indices just beyond the midpoint of their block (i.e., $j = i \mod 2^n = \frac{2^n}{2} = 2^{n-1}$), there is a minimum unique string consisting of the character at that point, $2^{n-1} - 1$ additional identical characters beyond that point lying within the same block, and then (additionally) the first character of the next block, thus yielding a unique total string length starting at position j of $2^{n-1} + 1 = j + 1$, as given by the formula above. For the indices in the following $2^{n-1} - 1$ positions of the string (i.e., for $2^{n-1} < j \leq 2^n - 1$), because the uniform symbol prefixes beginning at index point j have all previously been seen within this block, the smallest unique string consists of the prefix beginning at the current point (index j), proceeding through the end of the block, and including one character beyond the end of that block (which has not yet been previously encountered within the string). For a character at position j (zero-based) within the block, this yields a string length of $(2^n - j) + 1$. Thus, we have $\forall_{j=(i \mod 2^n) > 2^{n-1}} \Lambda_j = (2^n - j) + 1$. Therefore, we can decompose the sum of Λ_j values in a block as follows:

$$
\begin{aligned}
\sum_{j=0}^{L_b-1} \Lambda_j &= \sum_{j=0}^{\frac{L_b}{2}} (j+1) + \sum_{j=\frac{L_b}{2}+1}^{L_b-1} (2^n - j + 1) \\
&= \sum_{j=1}^{\frac{L_b}{2}+1} j + \sum_{2^n-j'=\frac{L_b}{2}+1}^{2^n-j'=L_b-1} (j'+1) \quad \text{, where } j' = 2^n - j \\
&= \sum_{j=1}^{\frac{L_b}{2}+1} j + \sum_{L_b-j'=\frac{L_b}{2}+1}^{L_b-j'=L_b-1} (j'+1) \quad \text{, given that } L_b = 2^n \text{ from (4.19)} \\
&= \sum_{j=1}^{\frac{L_b}{2}+1} j + \sum_{j'=\frac{L_b}{2}-1}^{j'=1} (j'+1) \\
&= \sum_{j=1}^{\frac{L_b}{2}+1} j + \sum_{j'=1}^{\frac{L_b}{2}-1} (j'+1) \\
&= \sum_{j=1}^{\frac{L_b}{2}} j + \frac{L_b}{2} + 1 + \sum_{j'=1}^{\frac{L_b}{2}-1} j' + \frac{L_b}{2} - 1 \\
&= \sum_{j=1}^{\frac{L_b}{2}} j + \frac{L_b}{2} + \sum_{j'=1}^{\frac{L_b}{2}} j' \\
&= 2\sum_{j=1}^{\frac{L_b}{2}} j + \frac{L_b}{2}.
\end{aligned}
\tag{4.21}
$$

From (4.21), we have the sum of the Λ_j across a single block, that is $\sum_{j=0}^{2^n-1} \Lambda_j = \sum_{j=1}^{\frac{L_b}{2}} \Lambda_j$, as given by $2\sum_{j=1}^{\frac{L_b}{2}} j + \frac{L_b}{2}$. Now recognizing that $\sum_{k=1}^{c} k = \frac{c(c+1)}{2}$, (4.21) can be further

reduced as follows:

$$\begin{aligned}
\sum_{j=0}^{2^n-1}\Lambda_j &= 2\sum_{j=1}^{\frac{L_b}{2}} j + \frac{L_b}{2} \\
&= 2\frac{(\frac{L_b}{2}(\frac{L_b}{2}+1)}{2} + \frac{L_b}{2} \\
&= \frac{L_b}{2}(\frac{L_b}{2}+1) + \frac{L_b}{2} \\
&= (\frac{L_b}{2})^2 + \frac{L_b}{2} + \frac{L_b}{2} \\
&= \frac{(L_b)^2}{4} + L_b \\
&= \frac{(2^n)^2}{4} + 2^n \quad \text{, given that } L_b = 2^n \text{ from (4.19)} \\
&= \frac{2^{2n}}{4} + 2^n.
\end{aligned} \tag{4.22}$$

Given (4.22) for the sum of the Λ_j across a single block, we turn our attention now to their sum across all blocks $\frac{1}{L}\sum_{i=0}^{L-1}\Lambda_i$, as is considered in (4.20). By applying $L = \frac{x}{vT}$ and (4.22) into (4.20), we have:

$$\begin{aligned}
\frac{1}{L}\sum_{i=0}^{L-1}\Lambda_i &= \frac{1}{\frac{x}{vT}}\sum_{b=1}^{\frac{x}{vT2^n}}\left(\frac{2^{2n}}{4}+2^n\right) \\
&= \frac{vT}{x}\frac{x}{vT2^n}\left(\frac{2^{2n}}{4}+2^n\right) = \\
&= \frac{1}{2^n}\left(\frac{2^{2n}}{4}+2^n\right) = \\
&= \left(\frac{2^n}{4}+1\right) = \left(2^{n-2}+1\right).
\end{aligned} \tag{4.23}$$

Recalling from (4.13) that $H = \left(\frac{1}{L}\sum_{i=0}^{L-1}\Lambda_i\right)^{-1}\ln L$, and recalling that $L = \frac{x}{vT}$, and (from (4.17)) that $W = W_0 2^n$, the formula for the entropy rate of the string can be simplified to

$$\begin{aligned}
H = \left(\frac{1}{L}\sum_{i=0}^{L-1}\Lambda_i\right)^{-1}\ln L &= \left(2^{n-2}+1\right)^{-1}\ln\frac{x}{vT} = \frac{\ln\frac{x}{vT}}{(2^{n-2}+1)} \\
&= \frac{\ln\frac{x}{vT}}{\left(\frac{\left(\frac{W}{W_0}\right)}{4}+1\right)} = \frac{4W_0\ln\frac{x}{vT}}{(W+4W_0)}.
\end{aligned} \tag{4.24}$$

Recall that the basal (minimum meaningful) spatial scale W_0 varies with the temporal resolution, reflecting the fact that more finely temporally sampled paths can benefit from

additional precision on the spatial scale (and thus a smaller bin size at which the sample begins to return unique values). Specifically, recall from (4.17) that $W_0 = vT$. Thus, for the joint scaling relation, we have

$$\frac{4W_0 \ln \frac{x}{vT}}{(W + 4W_0)} = \frac{4vT \ln \frac{x}{vT}}{(W + 4vT)} = \frac{4 \ln \frac{x}{vT}}{\left(\frac{W}{vT} + 4\right)}. \tag{4.25}$$

While choice of units will affect the size of the x, v, T and W_0 terms, we note that the governing terms $\frac{x}{vT}$ and $\frac{W}{vT}$ are distinguished by being of unit dimension; thus *the entropy rate expression is also of unit dimension, and thus invariant to unit change*. The first of these expressions is the total length of the sampled string; the latter is the number of samples required to cross the bin size. This result suggests that given a continuous, one-dimensional trajectory, the entropy rate of strings sampled at different resolutions according to bin widths W and temporal inter-sample spacing of T should scale as $O\left(\frac{4 \ln \frac{x}{vT}}{\frac{W}{vT}+4}\right)$.

Entropy of Paths with Mixtures of Velocities

We now consider traversing the same distance x, but where a fraction of the distance α is made at velocity βv, and fraction $(1 - \alpha)$ is made at velocity γv. For this case, the total elapsed trip time is $\frac{\alpha x}{\beta v} + \frac{(1-\alpha)x}{\gamma v} = \frac{x}{v}\left(\frac{\alpha}{\beta} + \frac{(1-\alpha)}{\gamma}\right)$. This yields a time-averaged velocity of

$$\bar{v} = \frac{x}{\frac{x}{v}\left(\frac{\alpha}{\beta} + \frac{(1-\alpha)}{\gamma}\right)} = \frac{v}{\left(\frac{\alpha}{\beta} + \frac{(1-\alpha)}{\gamma}\right)}. \tag{4.26}$$

The corresponding string length is

$$\begin{aligned} L' &= \frac{x}{\bar{v}T} \\ &= \frac{x}{vT}\left(\frac{\alpha}{\beta} + \frac{(1-\alpha)}{\gamma}\right) \\ &= \frac{\alpha x}{\beta vT} + \frac{(1-\alpha)x}{\gamma vT}. \end{aligned} \tag{4.27}$$

The total entropy rate is then calculated:

$$\left(\frac{1}{L'}\sum_{i=0}^{L'-1}\Lambda_i\right)^{-1} \ln L' = \left(\frac{1}{L'}\sum_{i=0}^{L'-1}\Lambda_i\right)^{-1} \ln\left(L'\right)$$

$$= \left(\frac{1}{L'}\left(\sum_{b=1}^{\frac{\alpha x}{\beta vT 2^n}}\left(\frac{2^{2n}}{4} + 2^n\right) + \sum_{b=1}^{\frac{(1-\alpha)x}{\gamma vT 2^n}}\left(\frac{2^{2n}}{4} + 2^n\right)\right)\right)^{-1} \ln\left(L'\right)$$

$$
\begin{aligned}
&= \left(\frac{1}{\frac{x}{vT}\left(\frac{\alpha}{\beta}+\frac{(1-\alpha)}{\gamma}\right)} \left(\frac{\alpha x}{\beta v T 2^n}\left(\frac{2^{2n}}{4}+2^n\right) + \frac{(1-\alpha)x}{\gamma v T 2^n}\left(\frac{2^{2n}}{4}+2^n\right)\right)\right)^{-1} \ln\left(L'\right) \\
&= \left(\frac{1}{\left(\frac{\alpha}{\beta}+\frac{(1-\alpha)}{\gamma}\right)} \left(\frac{\alpha}{\beta 2^n}\left(\frac{2^{2n}}{4}+2^n\right) + \frac{(1-\alpha)}{\gamma 2^n}\left(\frac{2^{2n}}{4}+2^n\right)\right)\right)^{-1} \ln\left(L'\right) \\
&= \left(\frac{1}{\left(\frac{\alpha}{\beta}+\frac{(1-\alpha)}{\gamma}\right)} \left(\frac{\alpha}{\beta}\left(\frac{2^n}{4}+1\right) + \frac{(1-\alpha)}{\gamma}\left(\frac{2^n}{4}+1\right)\right)\right)^{-1} \ln\left(L'\right) \\
&= \left(\frac{1}{\left(\frac{\alpha}{\beta}+\frac{(1-\alpha)}{\gamma}\right)} \left(\left(\frac{\alpha}{\beta}+\frac{(1-\alpha)}{\gamma}\right)\left(\frac{2^n}{4}+1\right)\right)\right)^{-1} \ln\left(L'\right) \\
&= \left(\frac{1}{\left(\frac{\alpha}{\beta}+\frac{(1-\alpha)}{\gamma}\right)} \left(\left(\frac{\alpha}{\beta}+\frac{(1-\alpha)}{\gamma}\right)\left(\frac{2^n}{4}+1\right)\right)\right)^{-1} \ln\left(L'\right) \\
&= \left(\frac{2^n}{4}+1\right)\ln\left(L'\right) = \frac{\ln\left(\frac{x}{\bar{v}T}\right)}{\left(\frac{\frac{W}{\bar{v}T}}{4}+1\right)} \\
&= \left(\frac{\frac{W}{\bar{v}T}}{4}+1\right)^{-1} \ln\left(\frac{x}{\bar{v}T}\right) \\
&= \frac{4\ln\frac{x}{\bar{v}T}}{\frac{W}{\bar{v}T}+4}. \qquad (4.28)
\end{aligned}
$$

We emphasize that the above is the same as the formula for the entropy rate $H = \frac{4W_0 \ln \frac{x}{vT}}{(W+4W_0)}$ derived in (4.24) for the case of a fixed velocity, except that the mean velocity $\bar{v}$ is substituted for originally fixed entropy v. While the analysis above considered two segments at different velocities, the derivation readily generalizes *mutatis mutandis* to other mixtures of velocities.

Impact of Spatial Uncertainty

It is well recognized that positioning systems such as GPS are associated with noise. We consider here the effects of such spatial noise on the entropy rate estimates. Employing the classic zero mean Gaussian noise model, we assume that GPS positioning is associated with measurements that are normally distributed around the true value μ with standard deviation σ. The probability of a given GPS measurement (a sample from that distribution) lying further than distance y from the mean is given by $1 - erf\left(\frac{y}{\sigma\sqrt{2}}\right)$. Now consider taking a measurement at the center point of a unidimensional bin of physical width W, which is

measured in the same unit system as y. The probability of a unidimensional measurement lying outside the distance to the boundary, $\frac{W}{2}$, is given by

$$p = 1 - erf\left(\frac{\frac{W}{2}}{\sigma\sqrt{2}}\right) = 1 - erf\left(\frac{W}{2\sqrt{2}\sigma}\right). \tag{4.29}$$

Now consider a sequence of measurements, as considered earlier. In the presence of noise, we can relate the Λ_j values within a block to a *truncated geometric distribution*. A random variable Y following a truncated geometric distribution with probability p of success and up to k tries, where the k^{th} draw is a success if all previous ones fail, has an expected value of

$$\begin{aligned} E[Y] &= \sum_{i=1}^{k-1} i\left((1-p)^{i-1}p\right) + k\left(1 - \sum_{i=1}^{k-1}\left((1-p)^{i-1}p\right)\right) \\ &= \frac{1-(1-p)(1-p)^{k-1}}{p} \\ &= \frac{1-(1-p)^k}{p}. \end{aligned} \tag{4.30}$$

For simplicity and as an approximation, we consider the draw associated with each element of the sum in (4.30) as independent, and as occurring from the center of the bin.

To compute Λ_j for each index j of samples in a block, we assume that a sample ends the unique sequence starting at j if the sample is erroneously reported to lie outside of the current bin. That is, we consider that if an incorrect value is sampled (i.e., if the positioning system erroneously reports a location outside of the current bin), that it will represent a repetition that terminates any unique sequence. If we consider a draw at a given sample position j, we treat the number of tries to obtain an erroneous value as following a truncated geometric distribution, where the number of tries is bound by the the length of the unique sequence that would start at position j when noise is absent. This maximum value is dictated by the position and the probability of achieving a value from outside of the bin is given by the value p. Therefore, Λ_j can be approximated by the expected value of this truncated geometric distribution.

For the case of multiple draws from this distribution associated with determining Λ_j at position j, we consider the discrepancy in the measurements independent. Adapting the formula in (4.21) for the probabilistic case, we can decompose the sum of the Λ_j for the

block as

$$2\sum_{j=1}^{\frac{L_b}{2}}(\text{times to repeat up to } j \text{ tries}) + \text{times to repeat up to } \frac{L_b}{2} \text{ tries},$$

where the 'times to repeat up to m' times are considered to reach the value m if and only if no erroneous reading has occurred, and is otherwise immediately truncated.

For our simplified case, we, therefore, approximate a total of the Λ_j across the current block of

$$\begin{aligned}
&2\sum_{j=1}^{\frac{L_b}{2}}\frac{1-(1-p)^j}{p}+\frac{1-(1-p)^{\frac{L_b}{2}}}{p}\\
&=\frac{1}{p}\left(2\left(\sum_{j=1}^{\frac{L_b}{2}}1-(1-p)^j\right)+1-(1-p)^{\frac{L_b}{2}}\right)\\
&=\frac{1}{p}\left(2\left(\frac{L_b}{2}-\sum_{j=1}^{\frac{L_b}{2}}(1-p)^j\right)+1-(1-p)^{\frac{L_b}{2}}\right)\\
&=\frac{1}{p}\left(2\left(\frac{L_b}{2}-\left(\frac{(1-p)\left((1-p)^{\frac{L_b}{2}}-1\right)}{(1-p)-1}\right)\right)+1-(1-p)^{\frac{L_b}{2}}\right)\\
&=\frac{1}{p}\left(2\left(\frac{L_b}{2}+\left(\frac{(1-p)\left((1-p)^{\frac{L_b}{2}}-1\right)}{p}\right)\right)+1-(1-p)^{\frac{L_b}{2}}\right)\\
&=\frac{1}{p}\left(\left(L_b+2\left(\frac{(1-p)\left((1-p)^{\frac{L_b}{2}}-1\right)}{p}\right)\right)+1-(1-p)^{\frac{L_b}{2}}\right)\\
&=\frac{1}{p}\left(L_b+1+2\left(\frac{(1-p)\left((1-p)^{\frac{L_b}{2}}-1\right)}{p}\right)-(1-p)^{\frac{L_b}{2}}\right).
\end{aligned} \tag{4.31}$$

Now, summing up across the $N_b = \frac{L}{L_b}$ blocks, we have a denominator to (4.13) of

$$
\begin{aligned}
\frac{1}{L}\sum_{i=0}^{L-1}\Lambda_i &= \frac{1}{L}\sum_{b=1}^{\frac{L}{L_b}}\left(\frac{1}{p}\left(L_b+1+2\left(\frac{(1-p)\left((1-p)^{\frac{L_b}{2}}-1\right)}{p}\right)-(1-p)^{\frac{L_b}{2}}\right)\right)\\
&= \frac{1}{L}\frac{L}{L_b}\left(\frac{1}{p}\left(L_b+1+2\left(\frac{(1-p)\left((1-p)^{\frac{L_b}{2}}-1\right)}{p}\right)-(1-p)^{\frac{L_b}{2}}\right)\right)\\
&= \frac{1}{pL_b}\left(L_b+1+2\left(\frac{(1-p)\left((1-p)^{\frac{L_b}{2}}-1\right)}{p}\right)-(1-p)^{\frac{L_b}{2}}\right)\\
&= \frac{1}{p}+\frac{1}{pL_b}\left(1+2\left(\frac{(1-p)\left((1-p)^{\frac{L_b}{2}}-1\right)}{p}\right)-(1-p)^{\frac{L_b}{2}}\right).
\end{aligned}
\tag{4.32}
$$

By substituting (4.32) into (4.13), we can express the entropy rate, in the presence of white noise, as

$$
H = \frac{\ln\frac{x}{vT}}{\frac{1}{p}+\frac{1}{pL_b}\left(1+2\left(\frac{(1-p)\left((1-p)^{\frac{L_b}{2}}-1\right)}{p}\right)-(1-p)^{\frac{L_b}{2}}\right)}. \tag{4.33}
$$

Recall that $L=\frac{x}{vT}$ and $L_b=\frac{W}{vT}$, where the total path length is x, physical bin width is W, the velocity is v, and inter-sampling period is T. We can further expand (4.33) by substituting $\frac{W}{vT}$ for L_b, and (4.29) for p. If the agent travels distance x with a mixture of velocities, v in (4.33) is substituted by the time-averaged velocity $\bar{v}$.

4.7 Acknowledgments

We would like to acknowledge the Natural Sciences and Engineering Research Council of Canada for providing funding, and Dr. Mark A. Smith of Sandia National Laboratories for initial discussions regarding entropy rate scaling effects.

4.8 Addendum

The manuscript in this chapter has been reformatted, and some paragraphs/sentences of the published version have been modified/added/deleted, based on edits proposed by the

examining committee, for inclusion in the dissertation. No substantial changes to the results/findings were made.

Chapter 5

Manuscript 3

Title: Multiscale Entropy Rate Analysis of Complex Mobile Agents

Citation: Paul T, Stanley, K, Osgood, N. Multiscale Entropy Rate Analysis of Complex Mobile Agents. Submitted to Science. January, 2017.

Abstract Predicting the motion of an object is a central scientific question. For deterministic or stochastic processes, models exist which characterize motion with a high degree of statistical reliability. For complex systems, or those where objects have a degree of agency, characterizing motion is far more challenging. The information entropy rate of motion through a discrete space can place a limit on the predictability of even the most complex or history-dependent actor, but the variability in measured encountered locations is inexorably tied to the spatial and temporal resolutions of those measurements. This relation depends on the path of the actor, and can be used to derive a general scaling law for mobility entropy rate, which depends on the spatial and temporal resolution and the marginal path properties within each cell along the path. Correcting for spatial and temporal effects through regression yields the marginal path properties and a measure of mobility entropy rate robust to changes in dimension, allowing comparison of mobility entropy rates between data sets. Employing this measure on empirical datasets yields novel findings, from the similarity of taxicabs to driftwood, to the predictable lives of undergraduates, to the browsing habits of Canadian moose.

Relationship to this Thesis The theoretical model developed in Manuscript 2 primarily focuses on the effects of spatio-temporal scale on the mobility entropy rate, as well as pa-

rameters that depend on human behaviors (average speed) or the study (sequence length). Although the model was validated against stylized mobility models, it was not compared with real mobility traces. This work extends the initial model to a general scaling model of mobility entropy rate. The final model was validated with six empirical datasets, which are not constrained to humans; and consider diverse mobility traces of animals, birds, taxicabs, and ocean drifters. The validated model, which encodes behavioral and quantization parameters, is the main research goal of this dissertation. The analysis of quantization effects on entropy rate reveals how the observed changes are in agreement with the agent behaviors, allowing qualitative comparison of different populations. This paper leverages the empirical behaviors analyzed in Manuscript 1 to extend the theoretical model in Manuscript 2, providing a complete model of entropy scaling. The model will enable researchers to compare aggregate behaviors of two populations from the respective results of mobility studies, carried out at different spatio-temporal resolutions. Detailed derivation of the entropy rate scaling model is provided as a supplementary material (Section 5.5) at the end of this chapter.

5.1 Introduction

How predictable is the motion of an object though space? For many deterministic or stochastic systems, this question can be answered with a degree of certainty using well-established models of physical systems. However, for complex physical systems, or systems where the actors have a degree of agency, this question remains difficult to answer. In their seminal work, Song *et al.* demonstrated how the measured entropy rate of a person moving through discretized space could be used to estimate the predictability, in the limit, of human actors [7]. While their analysis focused on the daily habits of individuals, their analysis technique is generally applicable to any object moving through a discretized space. However, their core conclusion – that human mobility is inherently predictable – was only established for the dataset that they considered: predominantly urban dwellers who own and use cellular phones on a regular basis. By employing cellular call records - routinely employed by subsequent works [5, 81, 83] - they not only biased their analysis to a particular demographic, but to a particular spatial and temporal resolution. More recent empirical research has established that the estimated predictability of human mobility is contingent on the scale [22] and structure [21, 151] of the data, and underlying mobility model assumptions Manuscript 2. While Song *et al.* made a foundational contribution to quantifying mobility predictability in complex systems, results obtained through their technique are only applicable to the population and spatio-temporal resolution represented in the data.

This paper builds upon the results of Manuscript 2 to derive a general solution to the scaling of mobility entropy rate estimates in a discretized space. The model shows excellent agreement with empirical data for agents as diverse as a university students, taxicabs, moose, or buoys. Employing the scaling law allows researchers to analyze the predictability of the mobility traces at spatial scales consistent with the underlying mobility, to renormalize mobility data to common spatio-temporal resolutions for meta-analysis, to predict the impact of increasing or decreasing the spatio-temporal resolution of their study, and – through the analysis of the scaling law parameters – come to actionable conclusions about the relative mobility behaviors of individuals or populations.

5.2 Materials and Methods

Song et al.'s central methodological insight [7]– that an agent's passage through a discretized space creates a sequence of symbols corresponding to the locations traversed (a visit string) – constituted a foundational contribution. Like all strings, the visit string has an intrinsic information content which can be readily approximated using Lempel-Ziv compression. As string length tends towards infinity, the compressibility of the string tends towards the information entropy rate of the string, which describes the limit of the string's predictability [28]. However, different spatial and temporal resolutions will create different symbolic representations of the same trajectory (for example, doubling spatial bin size will re-render the string AABBCCDD as AAAACCCC), and, therefore, lead to different intrinsic entropy calculations for each spatio-temporal representation, limiting the generalizability of the results. The entropy rate can be estimated as

$$H = \left(\frac{1}{L} \sum_{i=0}^{L-1} \Lambda_i \right)^{-1} \ln L \tag{5.1}$$

where Λ_i is the length of the shortest substring starting at position i that has not previously been encountered, and L is the length of the string overall, as used in Song *et al.* [7] and subsequent works [5, 81, 83]. In Manuscript 2, we presented a theoretical framework inspired by the spatio-temporal effects observed in other studies [21, 22] using Song et al.'s approach. They specifically considered regular spatial binning and temporal sampling of constrained paths, but the contribution was limited by key assumptions, which allowed Osgood *et al.* to split the summation in (5.1) into n distinct substrings, characterized by repetitions of identical symbols corresponding to the traversal of a single cell.

When considering paths which traverse implicitly (due to accuracy limitations) or explicitly (due to experimental structure) discretized spaces and times, the discretization itself can be leveraged to relax our previous assumptions [158]. Actors with agency can traverse a cell along a non-linear path, and potentially stop along the way, leading to

$$t_i = \frac{d}{v_i^*} + t_{d_i} \tag{5.2}$$

where t_i is the time to traverse the i^{th} cell, d is the width of the cell, v_i^* is the apparent velocity across the cell and t_{d_i} is the total (stationary) dwell time within the cell. The

apparent velocity across the cell is the average velocity, while moving, across the length of the cell, and incorporates the actual average velocity, and the path length through the cell.

When an agent traverses a cell, the length of the sequence of symbols correspondingly emitted is simply the time to traverse the cell divided the by the sampling rate:

$$L_i = \frac{t_i}{T} = \frac{1}{T}\left(\frac{d}{v_i^*} + t_{d_i}\right) \tag{5.3}$$

where L_i is the number of symbols generated per cell. From Osgood et al., the summation of Λ_i values for all positions in the string is

$$\sum_{i=0}^{L-1} \Lambda_i = \sum_{i=1}^{n} \left[\frac{L_i^2}{4} + L_i\right] \tag{5.4}$$

Where L_i is the length of the ith block of identical symbols (emitted in the course of transiting the ith cell) and n is the total number of blocks in a string. Denoting the sampling period as T, substituting $\frac{t_i}{T}$ for L_i, and solving for H, we obtain (see supplementary material (Section 5.5))

$$H(d,T) = \frac{\log L}{\frac{d^2}{4LT^2}\sum_{i=1}^{n}\frac{1}{v_i^{*2}} + \frac{1}{4LT^2}\sum_{i=1}^{n} t_{d_i}^2 + \frac{2d}{4LT^2}\sum_{i=1}^{n}\frac{t_{d_i}}{v_i^*} + \frac{4dT}{4LT^2}\sum_{i=1}^{n}\frac{1}{v_i^*} + \frac{4T}{4LT^2}\sum_{i=1}^{n} t_{d_i}} \tag{5.5}$$

where L is the length of the string, and d is the cell size. The numerator of this scaling law is the entropy of a string of unique symbols of length L, and the denominator is the amount by which that value is scaled. The sums aggregate terms involving v_i^* and t_{d_i}, which are intrinsic mobility parameters of the agent, which impact the L_i according to (5.3). Note that each term in the denominator corresponds to a number of samples or symbols in the string implicitly summed over a block to get L_i, and summed over all blocks, and represent the marginal impact of that property across the path. When divided by L, these values represent an average impact of that particular term over the entire string. If we consider v_i^* and t_{d_i} to be well-behaved random variables with means that are broadly invariant under resampling, then the overall average should converge, independent of d and T [159]. If continuous distributions for v_i^* and t_{d_i} are sampled at increasing coarse resolutions, the mean should be stable, as long as the discrete approximation provides a stable representation of the continuous distribution. This variance would be an artifact of the downsampling process and not indicative of the

underlying distribution, implying that (5.5) will only be valid over a range of scales, where the measurement scale adequately represents the mobility pattern. Similarly, if there is inherent scaling behavior within a dataset, where specific phases of motion are evident at particular scales, this approximation would only be valid for a single phase, and the more complex functional dependence of v_i^* and t_{d_i} on d and T can be empirically approximated (see supplementary material (Section 5.5)).

To validate our model, we computed the entropy scaling behavior of six distinct datasets comprising human, animal, and complex physical systems. For details on data selection, pretreatment and cleaning, see the Materials and Methods section in the supplementary material (Section 5.5).

Saskatchewan Human Ethology Dataset (SHED) 7 and 8: GPS/WiFi based location records from smartphone-based data collection over a four-week period of 63 and 75 university-affiliated participants, primarily undergraduates, in the summer and fall of 2016 in a mid-sized Canadian city. Participants who returned at least 15 records at 8-hour intervals were retained [141]. We expect these datasets to be largely similar to other human mobility datasets [36].

Roman Taxis: Over 21 million GPS records of locations at $\approx$ 7 s intervals, publically available in the Crawdad repository [160]. Over $350,000$ records from the top 59 taxi drivers, who returned at least 15 records at 8-hour sampling intervals, based on number of reported records were retained.

Moose: in study area of south-central Saskatchewan, telemetered with GPS tracking collars [161]. Moose tend to live solitary nomadic lives, browsing and sleeping at their pleasure over a home range.

Antarctic petrel: movements characterized by $55,176$ GPS traces of petrel behavior [162]. Petrels graze the surface of the water for small fish and crustaceans, only occasionally returning to their nesting sites during breeding season.

Buoys: in the Juan de Fuca Strait. Nine drifters were released off the coast of Vancouver, Canada to study the impact of surface and tidal currents in distributing pollution

generated by the city and port [163]. As buoys have no agency, they follow paths dictated by tides and currents.

Data were cleaned according to the procedures in the supplementary material (Section 5.5) to eliminate erroneous datapoints, such as individuals with sporadic GPS records. Following Chapter 4, the space covered by the dataset was gridded into cells 4 km across, then downsampled using a quadtree decomposition to a minimum grid cell size of either 62.5 or 15.625 meters (see supplemental material). Trajectories through the discretized space were rendered as visit strings. These strings were downsampled at regular intervals, with a maximum period (inter-sampling interval) of eight hours and a minimum of between 1 minute and 1 hour, depending on the structure of the data (see supplementary material (Section 5.5)). Trajectory duration was conserved to compare the same paths through time, leading to decreasing L with T, as longer sampling periods yielded shorter strings for the same trajectories.

Each generated string for each entity in each dataset had its entropy approximated using (5.1), implemented in custom C++ code. Entropies for each dataset at each spatial and temporal resolution were averaged over agents to generate the entropy rate central tendency for each dataset and each spatio-temporal resolution. Assuming that the means are stable under resampling, the summations in (5.5) can be treated as constants (denoted C_1 to C_5 in the supplementary material (Section 5.5)). The summation terms from (5.5) were fit to each entropy rate over (d,T) for each dataset using Eureqa, which employs evolutionary algorithms, from Nutonian Inc. [164, 165], using absolute error as the (global) optimization metric, approximating the marginal values in each of the sums as a constant. Mean squared error is reported as a goodness of fit metric.

5.3 Results

Changing spatial and temporal resolution changes the distribution of repeated characters within substrings sampled from a single cell. At the smallest spatial and temporal scales, fine motion is captured, but stationary periods will be strongly represented for regularly immobile agents such as undergraduate students. As the inter-sample interval increases, substrings

become increasingly short, until visit strings such as those associated with brief commutes are dropped entirely. Similarly, as spatial bin size increases, longer repeated substrings are expected, and, therefore, lower entropy rates. For a full characterization of the dependency of visit string length on spatial and temporal sampling regimes for the examined datasets, see the supplementary material (Section 5.5).

Fitting was able to achieve an excellent match between model and data for most datasets. Mean squared error was less than 10% of the total span of entropies calculated empirically. While they offer limited reliability in nonlinear fitting, R^2 values were greater than 0.9. Table 5.1 summarizes the fit quality and resulting coefficients from Qian et al. [22]. Surfaces denoting the model over the range of d and T values considered, and the calculated entropies, are shown in Fig. 5.1. It is clear from these results that the model in (5.5) provides an accurate description of how mobility entropy rates vary across a wide range of spatial and temporal scales.

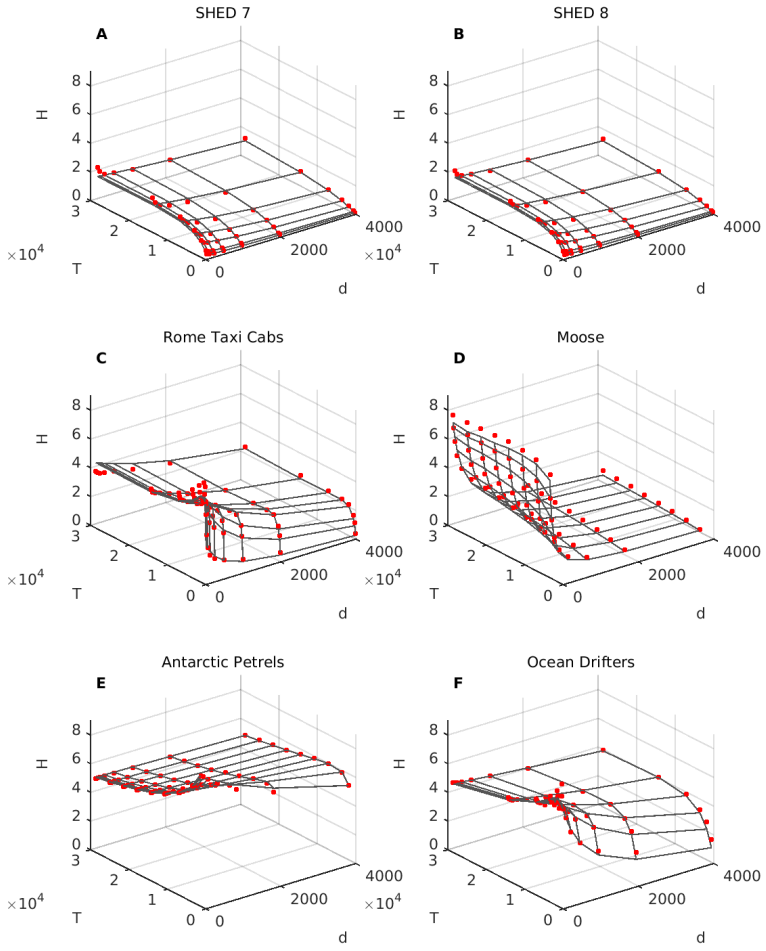


Fig 5.1: Entropy surface and empirical points for **A**) university students during Summer term, **B**) university students during Fall term, **C**) taxicabs in Rome, **D**) moose in south-central Saskatchewan, **E**) Antarctic petrels, and **F**) buoys in the Juan de Fuca Straight. d is in meters, and T in seconds, H is in bits. Petrels exhibited the greatest entropy. Notable similarities exist between students, regardless of season, and between taxis and driftwood. Moose have unique profiles reflective of their nomadic nature. Departures from theory are evident in the student datasets for large d and T, implying a change in scaling behavior.

Table 5.1: Constants after fitting equation (5.5) using nonlinear regression, with R^2 and Mean Squared Error. Fits are largely independent of squared velocity but show variability in squared dwell time and linear terms.

Dataset	$\sum_{i=1}^{n} \frac{1}{v_i^{*2}}$	$\sum_{i=1}^{n} {t_{d_i}}^2$	$\sum_{i=1}^{n} \frac{t_{d_i}}{v_i^*}$	$\sum_{i=1}^{n} \frac{1}{v_i^*}$	$\sum_{i=1}^{n} t_{d_i}$	R^2	MSE
SHED7	2.24E-02	6.42E+10	2.59E+07	1.65E+01	2.89E+06	0.93	0.027
SHED8	2.02E-05	5.59E+10	2.13E+07	4.54E+01	2.82E+06	0.95	0.016
Taxi	1.63E-05	2.75E+08	4.77E+05	9.40E+01	4.34E+05	0.93	0.1498
Moose	1.14E-01	4.78E+10	3.07E+09	2.47E+04	2.07E+07	0.98	0.093
Petrel	4.92E-09	4.75E-03	7.43E+05	4.53E-05	1.12E+06	0.996	0.005
Buoy	1.25E-01	8.22E+08	2.44E+06	1.05E+01	9.66E+05	0.98	0.045

Deviations from the model exist, particularly for small d across all datasets, likely the impact of unmodeled Gaussian noise in the measured locations [158]. Deviation between model and empirical entropy rates are also evident at large d and T, particular for the human mobility datasets. The scaling law underestimates the entropy rate at large cell sizes and sampling periods, indicating that the sampling of the path parameters has become sufficiently coarse as to make the means shift.

Undergraduate mobility is characterized by relatively low entropies at small T, as they are likely to spend prolonged intervals at the university or home, exhibiting string containing long runs of repeated location symbols, making the visit strings highly compressible. Taxi drivers are often on the move, and generally have higher entropies than students at all resolutions. This finding confirms Song et al.'s hypothesis that mobility entropy rate can be used to compare behavior between populations [7]. While it is common to find an undergraduate in the same spatial location after a half-hour, the same cannot be said for taxis, and this holds true across a range of spatial scales.

Because at small temporal scales, both students and taxis have a reasonable probability of being in the same location, increasing the inter-sample interval increases the entropy rate by decreasing the number of observed repetitions over the string. Petrels have few repeating substrings (see supplementary material (Section 5.5)), and, therefore, increasing the sampling

rate does not change the compressibility of a string, already well represented as a sequence of unique non-repeating symbols, and instead the entropy rate associated with the same paths decreases with rising T according to the numerator $\log L$. Petrels are the most entropic of all the datasets examined, although other seabirds, such as albatross, would be expected to have even greater mobility entropy rates across a wider variety of spatial scales.

Moose mobility entropy rate is distinct from both humans and birds in that it is almost invariant with analyzed sampling rates. Instead, moose mobility entropy rate falls sharply with d, likely due to grazing behavior, where short wanders happen nearly continuously. Once cell size is large enough to incorporate these meandering paths, entropy rate reaches a stable value, at around 500 m. Plateaus in the scaling behavior can indicate spatial or temporal scales at which spatial behaviors become indistinguishable. Entropy rate scaling analysis can provide insight into what characteristic spatial scales are important for populations under study.

The similarity between surfaces in Fig. 5.1 can be encapsulated in the values of the marginal path properties. According to (5.5), the values of each marginal property can be described as shown in Table 5.1. The values for $\frac{1}{v_i^{*2}}$ are substantially lower than the others, even given that the maximum value of d^2 considered is over 1.6 million square meters. All mobility traces have a negligible dependence on $\frac{1}{v_i^{*2}}$. Both SHED datasets have similar values for all remaining terms, indicating a degree of similarity. Likewise, the taxi and buoy coefficients are always within an order of magnitude of each other. The petrel dataset is distinct for having a negligible dependence on ${t_{d_i}}^2$ and $\frac{1}{v_i^*}$, but comparable dependence on $\frac{t_{d_i}}{v_i^*}$ and t_{d_i} as other datasets, reinforcing our hypothesis that the entropy rate scaling is due to the sampling rate of the mobility.

5.4 Discussion

The information content of a set of trajectories is a concise description of the disorder of the motion, and is related to the limit of predictability for that trajectory set [7]. Like many trajectory-based measures, this is contingent on the spatial and temporal resolution of the measurement, which is jointly a function of the marginal path properties across discretized

space and the spatial and temporal scales of measurement, according to a well-defined and empirically validated scaling law. In the scaling law, we used regression to estimate the mean mobility parameters. In the case of more complex variability with d and T, a semi-empirical approach can be adopted (see supplementary material (Section 5.5)) at the cost of theoretical rigor.

The scaling behavior itself summarizes and exposes characteristics of trajectories, particularly across datasets gathered at differing resolutions. From a simple examination of the scaling surfaces, we obtain the following novel findings:

Taxis are not like students: Students are less sensitive to changes in spatial and temporal scale than taxis, and have a lower overall entropy, consistent with spending time in class or at home, suggesting that entropy rate and its scaling properties are an appropriate tool for investigating the relative mobility behavior of human populations.

Taxis are like driftwood: Taxis and buoys both exhibit sensitivity to spatial and temporal resolution at the same scales, likely driven by least-cost paths through a flowing medium.

Moose movement traces change at scales below 0.25 **km**^{2}**:** The sharp increase in mobility entropy rate for moose across a range of sampling periods indicates a difference in observable behavior above and below that spatial scale.

Mobility entropy scaling has limits: Petrel paths are highly entropic, implying that the observations are at or above the spatial temporal resolution of their characteristic mobility scale.

The scaling law presented here accurately reproduces the mobility entropy rate for a wide variety of agents moving under their own agency or under the influence of complex deterministic systems. However, we have only considered the aggregate mobility entropy rates across all paths and have not considered individual mobile entropy rate, stratified within-subject mobility entropy rate, or the probability of predicting the next location under constraints as presented in Smith *et al.* [21], all of which are fertile areas for future research. While the scaling law provides exceptional fidelity to empirical data over a wide variety of spatial

scales, it is implicitly tied to the data through the regression-derived coefficients. However, for the scales and systems measured presented here, excellent agreement was obtained. The femtosecond behavior of moose, or light-year binned trajectories of buoys are unlikely to be of scientific interest. However, the behavior of humans over kilometers, measured on the order of hours, is of interest, and showed increasing disagreement in entropy rate values, with the theory in the student datasets. When disagreements with the theory do arise, this indicates a potential phase change in observable mobility behavior, and the scales at which this occurs have intrinsic scientific interest.

The central theoretical contribution of this work is the ability to separate path properties from measurement scale properties in entropy rate calculations. Comparison between students, taxis and buoys has demonstrated that classes of mobility entropy rate are likely to exist, manifested through the social, psychological and physical constraints of the system. When employing this methodology to describe datasets, a vocabulary of mobility classes could emerge, providing further insights.

Song *et al.* established that mobility entropy rates could characterize the predictability of human mobility traces [20]. This seminal work allowed within-subject comparison of overall path quality, but was limited by the scale dependence of the metric employed. Osgood *et al.* observed that much of the scaling behavior in mobility entropy rate could be accounted for by examining the structure of visit strings for stylized trajectories [158]. By extending Osgood et al.'s work, we were able to obtain a general scaling law that has been validated for empirical mobility datasets from students to moose and taxis to driftwood, and to describe novel findings about the relative properties of these datasets. While the work here has successfully been applied to complex phenomenon centered on populations with a degree of agency and on complex physical paths, it should be equally valid for systems currently well described using stochastic models, such as financial transactions, fluidic phenomena or particle behavior.

5.5 Supplementary Material

This section provides details on the theory and derivation of the model of entropy rate scaling presented in the main paper. A step-by-step derivation of the scaling law is provided. The presence of maxima/minima of the scaling law is discussed, and the behavior of the scaling law is described in the limits of the spatio-temporal resolution. This section also provides details on data collection, study population, data pre-processing, and spatio-temporal quantization of the data. For the datasets considered in our study, this section also describes effects of spatio-temporal quantization on aggregate run-length distribution of visit strings and the related impact on dictionary size. The fitting protocols applied for developed model are also described.

The theoretical model and the dataset details in this section complement the presentation and logic in the main manuscript to provide a clear insight into the internals of the model and how this applies to the comparison of mobility studies performed on different agents and/or at different spatio-temporal resolutions.

5.6 Theory

For this model, the LZ-based entropy rate, given in (5.6), is the method used to estimate the entropy rate of the mobility string, as employed by other researchers [5, 8, 21]. Our implementation of (5.6) is available at `https://github.com/tuhinpaul/lz_entropy_rate`. In (5.6), L is the length of the sequence and Λ_i is the length of the smallest sub-string that begins at the zero-based index i and was not encountered in positions 0 to $(i-1)$.

$$H = \left(\frac{1}{L} \sum_{i=0}^{L-1} \Lambda_i \right)^{-1} \log L \tag{5.6}$$

The theoretical model of the scaling of entropy rate with spatio-temporal resolution, proposed by Osgood *et al.* estimates entropy rate as

$$H(d,T) = \frac{4 \log \frac{x}{\bar{u}T}}{\left(\frac{d}{\bar{u}T} + 4\right)} \tag{5.7}$$

where d is the spatial scale, T is the sampling interval, x is the total travel distance, and $\bar{u}$ is the average velocity [158].

In this work, we extend the theoretical model, and apply it to empirical datasets from a wide variety of sources. The extended model considers the velocity and dwell time of the agents in formulating the entropy rate. Dwell times follow a power-law distribution, and constitute a major part of the real life mobility traces [5, 151].

Let the spatio-temporal resolution be represented as a tuple (T, d), where T is the sampling interval and d is the side length of a square cell in the spatial grid. We assume square cells for simplicity in spatial quantization, and define d as the length of an edge, or the characteristic length of a cell.

The time spent while in motion in the i^{th} cell is expressed as t_{m_i}. Without considering the dwell time, the apparent average velocity in the i^{th} cell on the path of an agent is $\frac{d}{t_{m_i}}$, as shown below

$$v_i^* = \frac{d}{t_{m_i}}. \tag{5.8}$$

The agent may traverse a distance of $k_i d$ in the i^{th} cell where $k_i \in \mathbb{R}^+$. The actual velocity might vary considerably but the average velocity as observed by the experimenter will appear as v_i^*.

The total time spent in the i^{th} cell is the sum of the time spent in motion and the dwell time. The time spent in motion inside the i^{th} cell is $\frac{d}{v_i^*}$ (from (5.8)). The total time spent in the i^{th} cell is, therefore, expressed as

$$\begin{aligned} t_i &= t_{m_i} + t_{d_i} \\ &= \frac{d}{v_i^*} + t_{d_i} \end{aligned} \tag{5.9}$$

Total dwell time in the i^{th} cell is the sum of dwells in the i^{th} cell:

$$t_{d_i} = \sum_k t_{d_i}^k \tag{5.10}$$

The observable average velocity considering dwell time, $\tilde{v}_i$, while passing the i^{th} cell can, therefore, be expressed as

$$\tilde{v}_i = \frac{d}{t_i} = \frac{d}{\frac{d}{v_i^*} + t_{d_i}}. \tag{5.11}$$

Let the agent travel through n cells on its entire path, treating repetitions of a cell separately. The number of blocks of repeatiting strings along the path would be represented by n. The total time spent in motion on the entire path is expressed as t_m, which is the sum of each t_{m_i}, as

$$t_m = \sum_{i=1}^{n} t_{m_i}. \tag{5.12}$$

Considering nd as the total traversed length with an apparent average velocity of v^*, and considering motion only,

$$v^* = \frac{nd}{t_m}. \tag{5.13}$$

The total dwell time along the entire path is the summation of dwell times in each cell:

$$t_d = \sum_{i=1}^{n} t_{d_i} \tag{5.14}$$

The observable average velocity for the entire path, considering dwell times as well, is represented as v^*.

$$\begin{aligned} \tilde{v} &= \frac{nd}{t_1 + t_2 + \ldots + t_n} \\ &= \frac{nd}{(t_{m_1} + t_{d_1}) + (t_{m_2} + t_{d_2}) + \ldots + (t_{m_n} + t_{d_n})} \\ &= \frac{nd}{\sum_{i=1}^{n} t_{m_i} + \sum_{i=1}^{n} t_{d_i}} \end{aligned} \tag{5.15}$$

From (5.12), we find that $\sum_{i=1}^{n} t_{m_i} = t_m$. From (5.13), we find that $t_m = \frac{nd}{v^*}$. Substituting them into (5.15), we find $\tilde{v}$ as

$$\begin{aligned} \tilde{v} &= \frac{nd}{\sum_{i=1}^{n} t_{m_i} + \sum_{i=1}^{n} t_{d_i}} \\ &= \frac{nd}{t_m + t_d} \\ &= \frac{nd}{\frac{nd}{v^*} + t_d} = \frac{d}{\frac{d}{v^*} + \frac{t_d}{n}}. \end{aligned} \tag{5.16}$$

There are n blocks along the entire path. Let L_i represent the length of the i^{th} block, and L be the length of the entire string. For simplicity, we assume that L_i is an even integer, and is approximated by

$$L_i = \frac{t_i}{T} = \frac{1}{T}\left(\frac{d}{v_i^*} + t_{d_i}\right). \tag{5.17}$$

Similarly, L can be expressed as

$$L = \frac{t}{T} = \frac{1}{T}\left(t_m + t_d\right). \tag{5.18}$$

Assuming a unique terminating symbol at the end of the string representing the path, and using the same decomposition as the theoretical model [158], we can express $\sum_{i=0}^{L-1} \Lambda_i$ in (5.6) as

$$\begin{aligned}
\sum_{i=0}^{L-1} \Lambda_i &= \sum_{i=1}^{n} \sum_{j=1}^{L_i} \Lambda_j \\
&= \sum_{i=1}^{n} \left[2 \sum_{j=1}^{\frac{L_i}{2}} j + \frac{L_i}{2} \right] \\
&= \sum_{i=1}^{n} \left[\frac{L_i^2}{4} + L_i \right].
\end{aligned} \tag{5.19}$$

Substituting (5.17) into (5.19), we can express $\sum_{i=0}^{L-1} \Lambda_i$ as

$$\begin{aligned}
\sum_{i=0}^{L-1} \Lambda_i &= \sum_{i=1}^{n} \left[\frac{\left(\frac{1}{T}\left(\frac{d}{v_i^*} + t_{d_i}\right)\right)^2}{4} + \frac{1}{T}\left(\frac{d}{v_i^*} + t_{d_i}\right) \right] \\
&= \frac{1}{4T^2} \sum_{i=1}^{n} \left[\left(\frac{d}{v_i^*} + t_{d_i}\right)^2 + 4T\left(\frac{d}{v_i^*} + t_{d_i}\right) \right] \\
&= \frac{1}{4T^2} \sum_{i=1}^{n} \left[d^2 \frac{1}{v_i^{*2}} + t_{d_i}^2 + 2d\frac{t_{d_i}}{v_i^*} + 4dT\frac{1}{v_i^*} + 4Tt_{d_i} \right] \\
&= \frac{1}{4T^2} \left(d^2 \sum_{i=1}^{n} \frac{1}{v_i^{*2}} + \sum_{i=1}^{n} t_{d_i}^2 + 2d \sum_{i=1}^{n} \frac{t_{d_i}}{v_i^*} + 4dT \sum_{i=1}^{n} \frac{1}{v_i^*} + 4T \sum_{i=1}^{n} t_{d_i} \right).
\end{aligned} \tag{5.20}$$

The entropy rate $H(d,T)$ from (5.6) can, therefore, be expressed as

$$
\begin{aligned}
H(d,T) &= \left(\frac{1}{L}\sum_{j}\Lambda_j\right)^{-1}\log L \\
&= \frac{4T^2\log L}{\frac{d^2}{L}\sum_{i=1}^{n}\frac{1}{v_i^{*2}} + \frac{1}{L}\sum_{i=1}^{n}t_{d_i}^2 + \frac{2d}{L}\sum_{i=1}^{n}\frac{t_{d_i}}{v_i^*} + \frac{4dT}{L}\sum_{i=1}^{n}\frac{1}{v_i^*} + \frac{4T}{L}\sum_{i=1}^{n}t_{d_i}}.
\end{aligned}
\tag{5.21}
$$

Let the summations in (5.21) be quantified as shown in (5.22) - (5.26).

$$C_1 = \sum_{i=1}^{n}\frac{1}{v_i^{*2}}, \tag{5.22}$$

$$C_2 = \sum_{i=1}^{n}t_{d_i}^2, \tag{5.23}$$

$$C_3 = \sum_{i=1}^{n}\frac{t_{d_i}}{v_i^*}, \tag{5.24}$$

$$C_4 = \sum_{i=1}^{n}\frac{1}{v_i^*}, \tag{5.25}$$

$$C_5 = \sum_{i=1}^{n}t_{d_i}. \tag{5.26}$$

Substituting (5.22) - (5.26) into (5.21), we can express $H(d,T)$ as

$$H(d,T) = \left(d^2\frac{C_1}{4T^2L} + \frac{C_2}{4T^2L} + 2d\frac{C_3}{4T^2L} + d\frac{C_4}{TL} + \frac{C_5}{TL}\right)^{-1}\log L. \tag{5.27}$$

Although t_i values can be approximated from GPS data, t_{m_i}, t_{d_i}, or v_i^* values can not be reliably extracted without additional speed data. We assume that the sums of the terms involving these quantities approximate the true sums within the over distances and sampling rates of interest, or that the values of d and T describe a single regime of the model. At large sampling intervals and kilometer-level spatial quantization, the deviations of the sums become significant, and the model is expected to deviate at those coarse spatio-temporal resolutions.

5.6.1 Variable Coefficient Analysis

In the main body of this chapter, we argued that C_1 — C_5 could be regarded as independent of d and T if the means were stable, which was demonstrated to hold true for all datasets for all but the largest values of d and T. Here we extend the analysis to capture how C_1 — C_5 might vary with d and T.

We can express $\frac{C_5}{TL}$ in (5.27) as follows.

$$\begin{aligned} \frac{C_5}{TL} &= \frac{\sum_{i=1}^{n} t_{d_i}}{T\frac{1}{T}\left(t_m + t_d\right)} && \text{From (5.26) and (5.18)} \\ &= \frac{t_d}{t_m + t_d} && \text{From (5.14)} \end{aligned} \tag{5.28}$$

Therefore, the term $\frac{C_5}{TL}$ corresponds to the fraction of dwell time in the total travel time. $d\frac{C_4}{TL}$ in (5.27) can be expressed as

$$\begin{aligned} d\frac{C_4}{TL} &= \frac{\sum_{i=1}^{n} \frac{d}{v_i^*}}{T\frac{1}{T}\left(t_m + t_d\right)} && \text{From (5.25) and (5.18),} \\ &= \frac{t_m}{t_m + t_d} && \text{From (5.8) and (5.12).} \end{aligned} \tag{5.29}$$

The quantity $d\frac{C_4}{TL}$ corresponds to the fraction of non-dwell time along the entire path. $d^2\frac{C_1}{4T^2L} + \frac{C_2}{4T^2L} + 2d\frac{C_3}{4T^2L}$ in (5.27) can be expressed as

$$\begin{aligned} d^2\frac{C_1}{4T^2L} + \frac{C_2}{4T^2L} + 2d\frac{C_3}{4T^2L} &= d^2\frac{\sum_{i=1}^{n}\frac{1}{v_i^{*2}}}{4T^2L} + \frac{\sum_{i=1}^{n} t_{d_i}^2}{4T^2L} + 2d\frac{\sum_{i=1}^{n}\frac{t_{d_i}}{v_i^*}}{4T^2L} \\ &= \frac{\sum_{i=1}^{n}\frac{d^2}{v_i^{*2}} + \sum_{i=1}^{n} t_{d_i}^2 + \sum_{i=1}^{n} 2\frac{dt_{d_i}}{v_i^*}}{4T^2L} \\ &= \frac{\sum_{i=1}^{n}\left(\frac{d}{v_i^*} + t_{d_i}\right)^2}{4T^2\frac{1}{T}\left(t_m + t_d\right)} && \text{From (5.18)} \\ &= \frac{\sum_{i=1}^{n} {t_i}^2}{4T\left(t_m + t_d\right)} && \text{From (5.9)} \\ &= \frac{t}{4T}\sum_{i=1}^{n}\frac{{t_i}^2}{t^2} && \text{From (5.9).} \end{aligned} \tag{5.30}$$

In (5.30), $\frac{t_i}{t}$ is the fraction of the time spent in the i^{th} cell. Let this ratio be expressed as f_i. Then, we can rewrite (5.30) as

$$\begin{aligned} d^2\frac{C_1}{4T^2L} + \frac{C_2}{4T^2L} + 2d\frac{C_3}{4T^2L} &= \frac{t}{4T}\sum_{i=1}^{n} f_i^2 \\ &= \frac{L}{4}\sum_{i=1}^{n} f_i^2 \qquad \text{From (5.18).} \end{aligned} \tag{5.31}$$

By substituting (5.31) in to (5.27), $H(d,T)$ can be rewritten as

$$H(d,T) = \left(\frac{L}{4}\sum_{i=1}^{n} f_i^2 + d\frac{C_4}{TL} + \frac{C_5}{TL}\right)^{-1} \log L. \tag{5.32}$$

If we know the distribution of f_i, we can approximate $\sum_{i=1}^{n} f_i^2$ when n changes due to change in (T,d). By expressing the sum as $f(d,T)$, we can then rewrite $H(d,T)$ from (5.32) as

$$H(d,T) = \left(\frac{L}{4}f(d,T) + d\frac{C_4}{TL} + \frac{C_5}{TL}\right)^{-1} \log L. \tag{5.33}$$

We can rely on empirical methods to estimate $f(d,T)$. Using Eureqa [165], we empirically found a solution for $f(d,T)$, which works in general form across datasets:

$$f(d,T) = C_6 + C_7 * d + C_8 * T. \tag{5.34}$$

By substituting (5.34) into (5.33), we can, therefore, express $H(d,T)$ as

$$H(d,T) = \left(\frac{L}{4}\left(C_6 + C_7 * d + C_8 * T\right) + d\frac{C_4}{TL} + \frac{C_5}{TL}\right)^{-1} \log L. \tag{5.35}$$

5.6.2 Scaling Law Behavior

Knowledge of maxima/minima of the entropy rate for a particular d or T may be useful in the design and evaluation of a mobility study to assess extreme values of the entropy rate. Similarly, the behavior of the model at the limits of d and T may ensure that if the model conforms to the desirable behaviors governed by the structure of the location string at those limits.

Maxima/Minima: Note that, L is a function of T. Substituting (5.18) into (5.27),

$$H(d,T) = \frac{4Tt\log t - 4Tt\log T}{d^2C_1 + C_2 + 2dC_3 + 4dTC_4 + 4TC_5}. \tag{5.36}$$

To check if $H(d,T)$ has a maxima/minima at any d or T, we need to differentiate (5.27) or (5.36) with respect to d and T. We find the relation

$$T = -\frac{2\left(C_1 d + C_3\right)}{C_4} \tag{5.37}$$

by differentiating (5.27) with respect to d. However, because $C_1, \ldots, C_5$ are positive numbers, no practical d or T can be found from (5.37), as all roots are negative in d and T.

For convenience, we use (5.36) to differentiate $H(d,T)$ with respect to T, and find the following relation dictating the presence of maxima/minima:

$$\begin{aligned}
&\frac{T}{\log\frac{t}{T} - 1} = \frac{C_1 d + 2C_3 d + 2C_2}{C_4 d + C_5} \\
\implies &\frac{T}{\log t - \log T - 1} = \frac{C_1 d + 2C_3 d + 2C_2}{C_4 d + C_5} \\
\implies &T = (\log t - 1 - \log T)\,\frac{C_1 d + 2C_3 d + 2C_2}{C_4 d + C_5} \\
\implies &T = (\log t - 1)\,\frac{C_1 d + 2C_3 d + 2C_2}{C_4 d + C_5} - \log T \frac{C_1 d + 2C_3 d + 2C_2}{C_4 d + C_5} \\
\implies &T + \log T \frac{C_1 d + 2C_3 d + 2C_2}{C_4 d + C_5} = (\log t - 1)\,\frac{C_1 d + 2C_3 d + 2C_2}{C_4 d + C_5}.
\end{aligned} \tag{5.38}$$

Given C_1 — C_5, for a given d, (5.38) can be solved using numerical analysis to find the T pertaining to a maxima/minima of $H(d,T)$.

Behavior of $H(d,T)$ at Limits: We examine if the model has desirable behavior at the limits of T and d. Behavior at the limits would explain whether the model conforms to theoretical constraints. For a constant T, dictionary size grows as d approaches 0. In the limiting case, all repetitions will be the result of dwelling, where C_2 and C_5, from (5.23) and (5.26) respectively, pertain to the dwelling of the agent. The repetition from dwelling scales the maximum entropy, $\log L$, of L symbols as follows:

$$\begin{aligned}
\lim_{d\to 0} H(d,T) &= \lim_{d\to 0} \frac{4T^2 L\log L}{d^2C_1 + C_2 + 2dC_3 + 4dTC_4 + 4TC_5} \qquad \text{From (5.27)} \\
&= \frac{4T^2 L\log L}{C_2 + 4TC_5}
\end{aligned} \tag{5.39}$$

When the cell size is very large ($d \to \infty$), all samples fall into the same cell, resulting in zero entropy rate, which is independent of the temporal resolution. From (5.27), this is mathematically presented as follows:

$$\begin{aligned} \lim_{d\to\infty} H(d,T) &= \lim_{d\to\infty} \frac{4T^2 L \log L}{d^2 C_1 + C_2 + 2dC_3 + 4dTC_4 + 4TC_5} \\ &= 0. \end{aligned} \tag{5.40}$$

For convenience, we use (5.36) to find the entropy rates at the limits of T. When d is a constant and $T \to 0$, entropy rate goes to zero as we end up with longer and longer strings of repeating locations with high compressibility:

$$\begin{aligned} \lim_{T\to 0} H(d,T) &= \lim_{T\to 0} \frac{4Tt\log t - 4Tt\log T}{d^2 C_1 + C_2 + 2dC_3 + 4dTC_4 + 4TC_5} \\ &= 0. \end{aligned} \tag{5.41}$$

When d is a constant and $T \to \infty$, then entropy rate is undefined because

$$\begin{aligned} \lim_{T\to\infty} H(d,T) &= \lim_{T\to\infty} \frac{4Tt\log t - 4Tt\log T}{d^2 C_1 + C_2 + 2dC_3 + 4dTC_4 + 4TC_5} \\ &= \lim_{T\to\infty} \frac{\frac{4Tt\log t}{T\log T} - \frac{4Tt\log T}{T\log T}}{\frac{d^2 C_1 + C_2 + 2dC_3 + 4dTC_4 + 4TC_5}{T\log T}} \\ &= \lim_{T\to\infty} \frac{\frac{4t\log t}{\log T} - 4t}{\frac{d^2 C_1 + C_2 + 2dC_3 + 4dTC_4 + 4TC_5}{T\log T}} \\ &= \text{undefined}. \end{aligned} \tag{5.42}$$

This is sensible because when $T \to \infty$, we can not have enough samples to evaluate $H(d,T)$ at that T for varying d.

We can also consider how the model behaves at extreme values of speed and dwell times. If $\forall_i\ t_{d_i} \to 0$, then entropy rate should depend on apparent average velocities at each cell and dwell times do not effect the entropy rate. The model conforms to this case of motion without dwelling:

$$\begin{aligned} \lim_{\forall_i\ t_{d_i}\to 0} H(d,T) &= \lim_{\forall_i\ t_{d_i}\to 0} \frac{4T^2 \log L}{\frac{d^2}{L}\sum_{i=1}^{n}\frac{1}{v_i^{*2}} + \frac{1}{L}\sum_{i=1}^{n} t_{d_i}^2 + \frac{2d}{L}\sum_{i=1}^{n}\frac{t_{d_i}}{v_i^*} + \frac{4dT}{L}\sum_{i=1}^{n}\frac{1}{v_i^*} + \frac{4T}{L}\sum_{i=1}^{n} t_{d_i}} \\ &= \lim_{\forall_i\ t_{d_i}\to 0} \frac{4T^2 \log L}{\frac{d^2}{L}\sum_{i=1}^{n}\frac{1}{v_i^{*2}} + \frac{4dT}{L}\sum_{i=1}^{n}\frac{1}{v_i^*}}. \end{aligned} \tag{5.43}$$

However, if the dwell time at the j^{th} cell approaches ∞, all samples after reaching that cell will be the same. The apparent average speed within the cell would approach 0. Because the cell would get stuck in the j^{th} cell, the time in that cell and the length of the substring emanating from that cell will be determined by the dwell time in that cell considering total observation time. The dwell time in the j^{th} block, in this case, is t_{d_j} (5.44), and the length of the j^{th} block is L_j (5.45).

$$t_{d_j} = t - \sum_{i=1}^{j-1} t_i \tag{5.44}$$

$$L_j = \frac{t_{d_j}}{T} = \frac{t - \sum_{i=1}^{j-1} t_i}{T} \tag{5.45}$$

Considering (5.19) for the j^{th} block, the entropy rate can be expressed as

$$\begin{aligned} \lim_{\exists_j \ t_{d_j}\to\infty} H(d,T) &= \lim_{\exists_j \ t_{d_j}\to\infty} \frac{4T^2 \log L}{\frac{d^2}{L}\sum_{i=1}^{n}\frac{1}{v_i^{*2}} + \frac{1}{L}\sum_{i=1}^{n} t_{d_i}^2 + \frac{2d}{L}\sum_{i=1}^{n}\frac{t_{d_i}}{v_i^*} + \frac{4dT}{L}\sum_{i=1}^{n}\frac{1}{v_i^*} + \frac{4T}{L}\sum_{i=1}^{n} t_{d_i}} \\ &= \frac{4T^2 \log L}{\frac{d^2}{L}\sum_{i=1}^{j-1}\frac{1}{v_i^{*2}} + \frac{1}{L}\sum_{i=1}^{j} t_{d_i}^2 + \frac{2d}{L}\sum_{i=1}^{j-1}\frac{t_{d_i}}{v_i^*} + \frac{4dT}{L}\sum_{i=1}^{j-1}\frac{1}{v_i^*} + \frac{4T}{L}\sum_{i=1}^{j} t_{d_i}}. \end{aligned} \tag{5.46}$$

If the agent is observed for infinite time, the entropy rate according to (5.46) would approach 0.

If the apparent average velocities in the cells approaches ∞, then v_i^* values would not effect the entropy rate in practice and the entropy rate would depend on dwell times:

$$\begin{aligned} \lim_{\forall_i \ v_i^*\to\infty} H(d,T) &= \lim_{\forall_i \ v_i^*\to\infty} \frac{4T^2 \log L}{\frac{d^2}{L}\sum_{i=1}^{n}\frac{1}{v_i^{*2}} + \frac{1}{L}\sum_{i=1}^{n} t_{d_i}^2 + \frac{2d}{L}\sum_{i=1}^{n}\frac{t_{d_i}}{v_i^*} + \frac{4dT}{L}\sum_{i=1}^{n}\frac{1}{v_i^*} + \frac{4T}{L}\sum_{i=1}^{n} t_{d_i}} \\ &= \frac{4T^2 \log L}{\frac{1}{L}\sum_{i=1}^{n} t_{d_i}^2 + \frac{4T}{L}\sum_{i=1}^{n} t_{d_i}}. \end{aligned} \tag{5.47}$$

5.7 Data Collection and Features

We used six empirical datasets encompassing mobility of humans [141], taxi cabs [160], animals [161], [162], and ocean drifters [163] to evaluate the performance of the theoretical model. Human mobility patters were taken from the Saskatchewan Human Ethology

Datasets (specifically, SHED7 and SHED8) [141], which are linked to the ongoing development of iEpi [135]. The SHED datasets contain detailed mobility, activity, and contact traces from university students and staff. We considered the mobility traces from the taxi cab mobility study conducted by Bracciale *et al.* in Rome, Italy [166], [160] as a contrasting human mobility dataset. Understanding the taxi trace patterns can play an important role as a contrast to the patterns of undergraduates. Taxi cab traces incorporate mobility traces of random people, and are expected to encounter popular routes and important urban locations. We also consider the mobility patterns of wild animals, less constrained by urban enviornments. Entropy rate could help understand animal behavior and predictability at a particular time window and spatial quantization. We assessed the model with the GPS traces that were collected from collars mounted on moose [161] and Antarctic petrels [162]. Because the scaling law does not require movements with agency, we also use the mobility tracks of ocean surface drifters [163] to validate that the model applies in general to complex physical pheonmenon as well. The drifter data also includes times spent on the land.

The durations of the studies behind the datasets vary largely, as shown in Table 5.2. The mentioned duration of SHED7 and SHED8, taxi cab, and ocean drifter studies in Table 5.2 are based on all the available date values in the datasets. The traces in the taxi cab study were sampled at much smaller intervals than in other datasets. Therefore, we limited the location traces to the first fifteen days of the study to make entropy rate calculation feasible within a reasonable amount of time for small T values. For the moose dataset, records between Jan 2012 and Feb 2015 are considered, and this is reflected in Table 5.2. The records in the petrel dataset span from Dec 2011 to Jan 2014.

The agents/participants of each dataset were passed through a filtering process to ensure that they had a minimum number of records at large sampling intervals. The details of the filtering process are described in Data Mediation (Section 5.7.2). The location traces in some datasets are accompanied with accuracy values, as indicated in Table 5.2. Their application is described in Data Mediation (Section 5.7.2) as well.

The base interval in Table 5.2 refers to the approximate interval of data collection, as we observed in the data or was available from the corresponding study.

5.7.1 Dispersion Maps

Fig. 5.2 - Fig. 5.7 show the dispersion of the agents/participants of the datasets, over three days, as heat maps of visited locations. All locations visited by all participants on the selected days are considered. For a given spatial quantization, all locations within a spatial bin were grouped together. The quantization process is described in Data Mediation (Section 5.7.2). All the locations in a group were represented by the location (latitude, longitude) = $(lat_{group}, lon_{group})$ as follows:

$$lat_{group} = \frac{\min(\text{latitude in group}) + \max(\text{latitude in group})}{2} \tag{5.48}$$

$$lon_{group} = \frac{\min(\text{longitude in group}) + \max(\text{longitude in group})}{2} \tag{5.49}$$

The center and zoom level were set manually to make the presentation legible because otherwise the maps needed to cover larger areas, zooming out the locations of interest. The manual selection of the map center and zoom level dropped out some visited locations - a trade-off made to make the maps comprehensible. Based on the frequency of visits, the visited locations are colored using a gradient from red (most visited) to green (least visited). The scale bars on the bottom-left corners of the maps indicate different distances because of the variation in the speed and span of movement of the corresponding agents. In the SHED (Fig. 5.2 and Fig. 5.3) and taxi (Fig. 5.4) studies, all participants take part in the study at or around the same time. Locations of agents in other datasets may not overlap. We find that SHED participants (Fig. 5.2 and Fig. 5.3) visit similar locations from day to day and

Table 5.2: Dataset details

dataset	SHED7	SHED8	Taxi	Moose	Petrel	Drifter
Duration (days)	35	29	30	1143	777	130
#(Agents)	63	75	316	36	124	9
#(Accepted agents)	56	70	59	36	124	9
Accuracy available?	Y	Y	N	N	N	N
Base Interval	5 min	5 min	15 s	1 hr	30 min	10 min

show different behaviors in the weekend. Similar trends are found in the taxi dataset (Fig. 5.4) but their dispersion varies significantly based on spatio-temporal resolution. Moose (Fig. 5.5) visit locations within a larger home range with a few hotspots. Ocean drifters (Fig. 5.7) change their place significantly from day to day. Similar behaviors are observed in petrels (Fig. 5.6). Changing spatio-temporal resolutions redistribute hot spots in the maps because of movement of agents to new places. The change in hotspots is less pronounced in humans and moose than in other datasets, indicating their relatively slower movement. Change of hotspots in the taxi dataset is well pronounced. The change of locations of petrel and ocean dirfters are slightly obscured by the relatively large area of span by these agents.

5.7.2 Data Mediation

We consider a bounding box as a simple constraint to remove erroneous location samples in the datasets, rejecting records outside of it. The bounding box is either computed from the minimum and maximum latitude and longitude values found in the data or arbitrarily from the expected span of the participants' movements in the study (e.g., a city).

In each dataset, the location traces are collected roughly at specific intervals. For each agent, we divide the entire duration of the available data into time windows, called duty cycles. The length of a window, T_w, is the same as the base/fundamental data collection interval of the dataset. We assign one sample to each time window. If location accuracy data are available, the most accurate closest to the start of the time window is chosen. If accuracy information is unavailable, we consider the sample closest to the beginning of the window. If location accuracy data are available, they are used to break the ties between samples having the same timestamps. If a tie still exists or if accuracy is not available, the tie is resolved randomly. A time window with no samples is not assigned any sample from other windows. All unassigned samples are dropped. The resultant sequences of GPS traces, are mapped to spatial grids at different spatial quantization, and further down-sampled using different sampling periods. Entropy rates calculated from these strings are used to ascertain the effectiveness of the model at different spatio-temporal quantization. We generate location sequences at different spatio-temporal resolutions as follows:

Down-sampling: For a down-sampling interval T, we choose every $\left(\frac{T}{T_w}\right)^{\text{th}}$ trace, if available, from the string constructed using time window T_w, as mentioned above. Down-sampling reduces the length of the string used to calculate entropy rate. For each dataset, we use values for T such that $T \bmod T_w = 0$. The chosen values, therefore, depend on the base interval of data collection. The list of down-sampling intervals are provided in Table 5.3.

Spatial Quantization: Unlike down-sampling, spatial quantization does not change the string length. For a down-sampling interval T and spatial resolution d, each record in the down-sampled string is mapped to a cell in a spatial grid. The mapping depends on a record's distance from the top-left corner of the study bounding box. If (x_0, y_0) and (x, y) are the (longitude, latitude) tuples of the top-left corner of the bounding box and the record respectively, the row and column values (row, col) of the mapped cell are computed with the following formulae.

$$\text{row} = \min\left(1, \left\lceil \frac{hd((x,y),(x,y_0))}{d} \right\rceil\right) \tag{5.50}$$

$$\text{col} = \min\left(1, \left\lceil \frac{hd((x,y),(x_0,y))}{d} \right\rceil\right) \tag{5.51}$$

where the function $hd(g_1, g_2)$ indicates the Harvesine distance [167] between geographic coordinates g_1 and g_2. We used the same set of spatial resolutions for all datasets: $\frac{4000}{2^n}$ meters, where $1 \leq n \leq 9$. This range encompasses the accuracy of commodity GPS hardware, as employed in [141], to the transmission range of cell towers in urban areas, as described in [7].

For SHED 7, SHED 8, and taxi datasets, we accepted participants having at least fifteen location records (arbitrarily decided) for the largest sampling interval used for the corresponding dataset. For other datasets, we only discarded erroneous location records as described above.

Table 5.3: Down-sampling intervals of different datasets

dataset	Down-sampling Intervals
SHED7	5 min, 10 min, 30 min, 1 hr, 2 hr, 4 hr, 8 hr
SHED8	5 min, 10 min, 30 min, 1 hr, 2 hr, 4 hr, 8 hr
Taxi	1 min, 5 min, 10 min, 30 min, 1 hr, 2 hr, 4 hr, 8 hr
Moose	1 hr, 2 hr, 3 hr, 4 hr, 5 hr, 6 hr, 7 hr, 8 hr
Petrel	30 min, 1 hr, 2 hr, 3 hr, 4 hr, 5 hr, 6 hr, 7 hr, 8 hr
Drifter	10 min, 30 min, 1 hr, 2 hr, 4 hr, 8 hr

5.7.3 Individual Entropy Rate Distribution

We use (5.6) to compute the entropy rate of the location sequence of each participant at each (T, d) pair. This gives us (p, T, d, H) tuples where p is the participant ID and H is the computed entropy rate.

In Fig. 5.8, boxplots of the same color pertain to the same sampling interval. For any given sampling interval, the spatial quantization grows as we move to the right, from 15.625 m to 4 km using the scaling factor of 2. For the SHED datasets, we removed the results at spatial quantization of 15.625 m and 31.25 m for lack of precision. We can see from the distribution of individual entropy rates in Fig. 5.8 that for a given (T, d) pair, the range of individual entropy rates across is small. Therefore, for a given (T, d) pair applied on a dataset, we consider the average of H across all participants as the aggregate entropy rate for simplicity.

When the spatial quantization becomes coarser, decreasing entropy rate is the expected behavior because longer repeating sequences of the same symbol make the string more compressible, as observed in Fig. 5.8. The plots of SHED 7 and SHED 8 tell us that humans are more unpredictable at a finer spatial granularity, if observed at large intervals. However, as the spatial scale becomes coarser, the difference in predictability due to differences in sampling interval decrease. Changing sampling interval in the moose dataset does not significantly change the entropy rate distribution, which can be ascribed to their tendency to move randomly while grazing. However, low entropy at coarser spatial scales indicate that

they change position steadily. Compared to moose, petrels move larger distances. Therefore, their entropy rates do not fall as sharply as those of moose when observed at large intervals. Ocean drifters demonstrate similarity to petrels at large sampling intervals; but at smaller sampling intervals, their entropy rates decline faster as d increases. This indicates that their positions are probably changing slower than those of petrels. It's interesting that as sampling period increases, taxi cabs become more predictable. This might be the result of controlled routes taken by the cab drivers, popular destinations of the passengers, and short spans of movement (mostly in central Rome) as shown in Fig. 5.4. Depending on data availability and base sampling interval, different down-sampling intervals were used for the datasets, as presented in Table 5.3. The comparison of individual entropy rate distributions of the datasets is presented in Fig. 5.9. Humans are more predictable, which is not surprising. Birds are more unpredictable than land animals. Placement of taxi cabs between the extremes is sensible as they are always on the move, but constrained to human activity patterns.

The location string from an agent's mobility not only depends on the sampling interval and spatial quantization, but also the agent behaviors such as speed and propensity to visit some place repeatedly. The entropy rate represents the compressibility in the string, which depends on the string structure, which in turn depends on the spatio-temporal resolution and agent behaviors. Two important aspects of the string structure are run length distribution and dictionary size. A run is a sequence of the same symbol and the dictionary is the set of unique symbols in the string. Comparing run length distributions and dictionary dynamics under different sampling regimes may provide important insight into the relation of entropy rate with agent behaviors.

5.7.4 Aggregate Run Length Distribution

In the sequence of location records, each run of the same location is considered a run. Fig. 5.10 shows the run length distributions of the datasets, computed across all participants/users. Run lengths of different participants are not combined. Run lengths exhibit a Pareto distribution. The curves for the SHED datasets demonstrate significant tails.

Distribution graphs by T: Fig. 5.11 - Fig. 5.16 show the dependence of aggregate run-length distribution on d for different fixed T values. We observe the following:

- The curves representing smaller d have steeper slopes. Small d is not supportive of repetitions, and therefore, the exponent of the power law becomes large, emulating shorter run lengths.
  - The change in slope with varying d is more prominent at smaller T than larger ones. At larger T, consecutive samples are more likely to be different than at smaller T; therefore, making d smaller does not have as much effect as in smaller T.
- In some datasets we see prominent tails at small T and large d.
  - In the taxi dataset, the tail is visible only at T = 1min. Therefore, the tail indicates a critical time constant related to waiting in traffic in Rome.
  - The ocean drifter dataset does not show such tails at large T, and such tails are absent in the petrel dataset. The agents in these datasets move to new spatial cells, which are less likely to be revisited (especially for small d), faster than the agents in other datasets.
  - Moose curves have comparatively influential tails at larger d even when T is large. This indicates that moose do not move long distances within 8 hours.
- SHED datasets have prominent tails for all T, mostly at large d, indicating a relatively small geographic span of movement.
- The moose curves indicate that the run-length value at which a pair of of curves, having consecutive d values, intersect increase as d values increase in the pair. Such overlaps are visible in the taxi and ocean drifter datasets as well, apparent at smaller T. This indicates that higher d is favorable for large run-lengths.

Distribution graphs by d: Fig. 5.17 - Fig. 5.22 show the dependence of aggregate run-length distribution on T. We observe the following:

- At small d values, the slopes of the log-log curves, as T varies, do not change significantly.
  - The lines at higher T move downward because the number of observed samples go down as T increases.
- At a small d, run-lengths tend to become smaller. This is why, for any T, frequencies of smaller run-lengths are higher at smaller d values.
- At larger d values, T has more effect on the run-length distributions.
  - At small T and large d values, the consecutive samples are more likely to be the same and therefore, longer run-lengths can be found. In a Pareto distribution, this would translate to a smaller exponent and, therefore, the curves have larger angles with the X-axis.
  - Increasing T when d is large makes the curves steeper. This means that for the power law distribution, the exponent is higher and shorter run-lengths are the norm. At large T, even if d is large, consecutive samples are less likely to be the same, especially if the agent is moving further. This behavior is prominent in the Petrel and ocean drifter datasets. Moose and taxi datasets also show similar behaviors.
- The tails at small T, which are more prominent at large d, indicates multi-modal distribution because larger run-lengths appear at small T values. Longer tails in the moose dataset indicate their steady movement. The tails of students' run-length distributions indicate larger run-lengths even at small d and T because humans spend a significant amount of time indoors and sitting in the same place.

Relating Run-length to Entropy Rate: Contrasting the slopes of the logarithmic curves in Fig. 5.10 with the order of datasets in Fig 5.9 based on the entropy rates, we find that steeper slopes in run length distribution corresponds to higher entropy rates. By comparing the per-dataset run length distribution graphs of Fig. 5.11 - Fig. 5.16, we see

that datasets demonstrating high entropy rates in Fig. 5.9 have mostly shorter run lengths overall.

As we can see in Fig. 5.8, the entropy rate is normally the lowest at the smallest T and the largest d, due to repetitions in the observed sequences. Large portions of run-length distribution curves at these d and T values are mostly fluctuating tails (Fig. 5.11 - Fig. 5.16) except petrels because they change their places quickly (Fig. 5.15).

The highest entropy rate is associated with the smallest d, as found in Fig. 5.8. Larger T and smaller d together are favourable for raising the entropy rate. However, the length of the available sequence length after down-sampling at a larger T affect the entropy rate. Therefore, for the lowest d, entropy rate at a larger T may be lower.

5.7.5 Growth of Dictionary Size

This section shows how the dictionary size grows with the length of the location string under spatio-temporal quantization for each dataset, and how that effects the entropy rate.

If any agent in a dataset visits a spatially quantized location, it is considered an unique dictionary element, irrespective of how many participants or agents visit that location. The growth of dictionary size as the aggregate sequence length of all agents in the dataset grows is presented in Fig. 5.23 - Fig. 5.28, where each figure represents one dataset.

For all datasets, when T is small, the slopes are smaller for larger values of d. Small T favors repetition but if d is large, many of these repeated symbols fall in the same cell. As a result, dictionary size grows relatively slowly. As shown in Fig. 5.27, petrels rapidly move to new places and their dictionary size growth curves (pertaining to different d values) exhibit almost the same slope at large T whereas there's significant difference between slopes representing different d values for small T. When T is large, petrels are expected to move to farther locations at the next sampling time than in the case of a smaller T. If they could move large distances in between two samples even at small T, differences in slope among curves would have diminished for small T. Thus, the slope difference between curves representating different d for small T reflects the speed limit of participants/agents. This is, however, different for taxis (Fig. 5.25) because although they can move quickly, they move within a constrained area along well-defined paths.

For a given T, as d increases, the slope of the tangent to the curve at the the root decreases. This indicates that dictionary size grows slower at large d, which is the expected behavior as explained above. However, for the petrel and ocean drifter datasets, increasing d does not have as much effect on the growth of the dictionary size as in other datasets. Even at large values of T, slopes of petrel and ocean drifter datasets remain significantly higher than other datasets at large d values. This reflects highly nomadic behavior of petrels and ocean drifters. For these datasets, the curves representing different d values come closer as T increases.

Although the slopes of SHED datasets look similar to taxi and moose datasets because of the scaling of X and Y axes in the graphs, the growth of dictionary size in taxi and moose datasets is larger than the SHED datasets (Fig. 5.29 — Fig. 5.34) where the ratios of dictionary sizes to the total sequence lengths are plotted. For all datasets, increasing T increases the ratio of new locations visited. For the values of the ratios for different (T, d) tuples for different datasets, please refer to Fig. 5.29 — Fig. 5.34.

Petrels show large ratios of unique locations in their location history, and the ratio does not decrease significantly, apart from an initial decrease for large d. For small T and large d the ratio is above 0.4, whereas for small d, the ratio is close to 1. Ocean drifters also exhibit large ratios, but the ratios fall significantly as d increases or the sequence length grows, implying that the agents in the the ocean drifter are less nomadic than petrels. As T increases, the ratios become higher for large d, indicating that the drifter do not have strong repetitive patterns. For moose, the ratio between dictionary size and string length is relatively constant as the sequence lengths grow, but the ratios are lower at large d values. This is in agreement with their slower nomadic lifestyle. For taxi cabs, the ratios drop as sequence lengths grow, especially for small T. This is because taxi cabs spend their time on defined paths within a small area (e.g., a city). At large T, the ratios do not drop at the same rate when sequence lengths grow because the samples are more likely to be different. Moreover, we have smaller sequence lengths than at smaller T. For SHED datasets, the ratio of the dictionary size to the sequence lengths is comparatively much lower than that of other datasets (below 0.1 for $T = 5\text{min}$). This indicates that routes taken by humans and places where they spend most time are well-defined, and compared to other datasets, they do not

visit many distinct places. When T is small, the ratio for any d is small because of too many samples. As T becomes larger, the differences of the ratios for different d becomes prominent. The ratio of new locations remin largely steady over the sequence lengths. However, as T grows, we have fewer locations in the sequence to evaluate this behavior.

The datasets that demonstrate large ratios of new locations in Fig. 5.29 - Fig. 5.34 also demonstrate high entropy rates in Fig. 5.9. Figure 5.35 shows the distribution of the aggregate dictionary growth ratios across datasets for three (T, d) tuples similar to Fig. 5.9. Comparing Fig. 5.35 and Fig. 5.9 shows that the ordering of datasets according to the ratio is in agreement with the order based on entropy rate. This indicates that the number of new locations significantly contribute to the value of the entropy rate.

5.7.6 Summary

The heat maps of visit locations in Fig. 5.2 - Fig. 5.7. show that different populations exhibit different degrees of uncertainty in their day-to-day mobility patterns. Differences in the mobility entropy rates of different population can be explained by the differences in mobility feaures like runlength distribution and growth of dictionary sizes. This proves that these mobility features are correlated with the mobility entropy rate. The datasets and (d, T) pairs showing large entropy rates have shorter run lengths and large ratio of dictionary size to the sequence length.

5.8 Fitting Protocols

We used the Eureqa software [164, 165] to derive the constant terms in our model, shown in (5.27), from the location sequences after spatio-temporal quantization. Eureqa [165] is an artificial intelligence-based data non-linear regression tool, which estimated the parameters in (5.27) via global-optimization based nonlinear regression. The input to Eureqa for data regression is a set of $(dset, T, d, L, lzH)$ tuples where lzH is the aggregate entropy rate for the data set $dset$ at spatio-temporal quantization (T, d) and L is the corresponding average sequence length. We used R^2-based goodness of fit and mean squared error as the error metrics to evaluate fit performance.

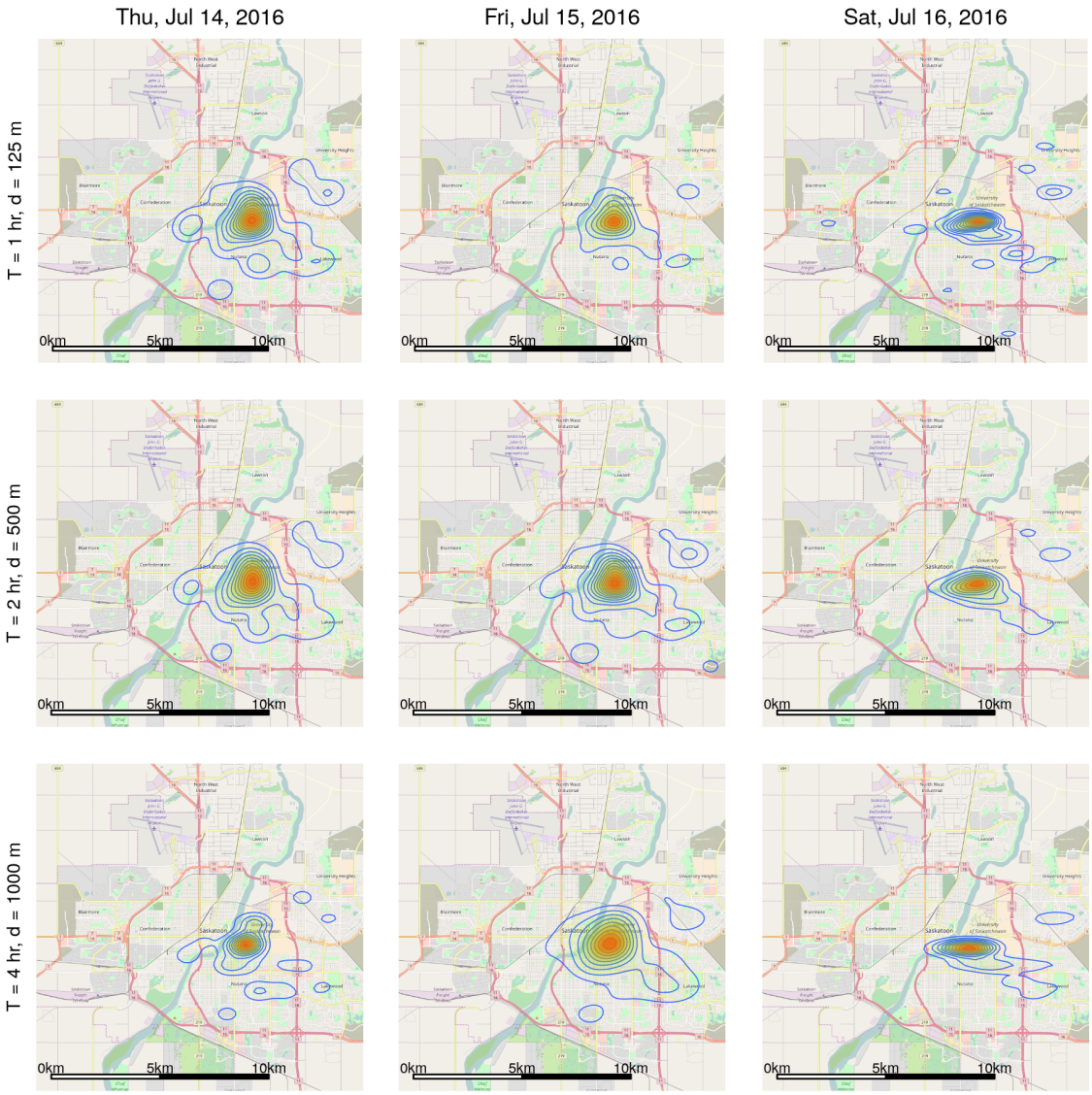


Fig 5.2: Heatmap of the dispersion of participants (undergraduate students) of SHED 7 over three consecutive days in the summer of 2016. Each column represents a day and each row represents a (T, d) pair. The participants visited fairly similar locations, which were centered around the University of Saskatchewan. Dispersion in the weekend is different than the weekdays. Weekdays exhibit visually similar dispersion.

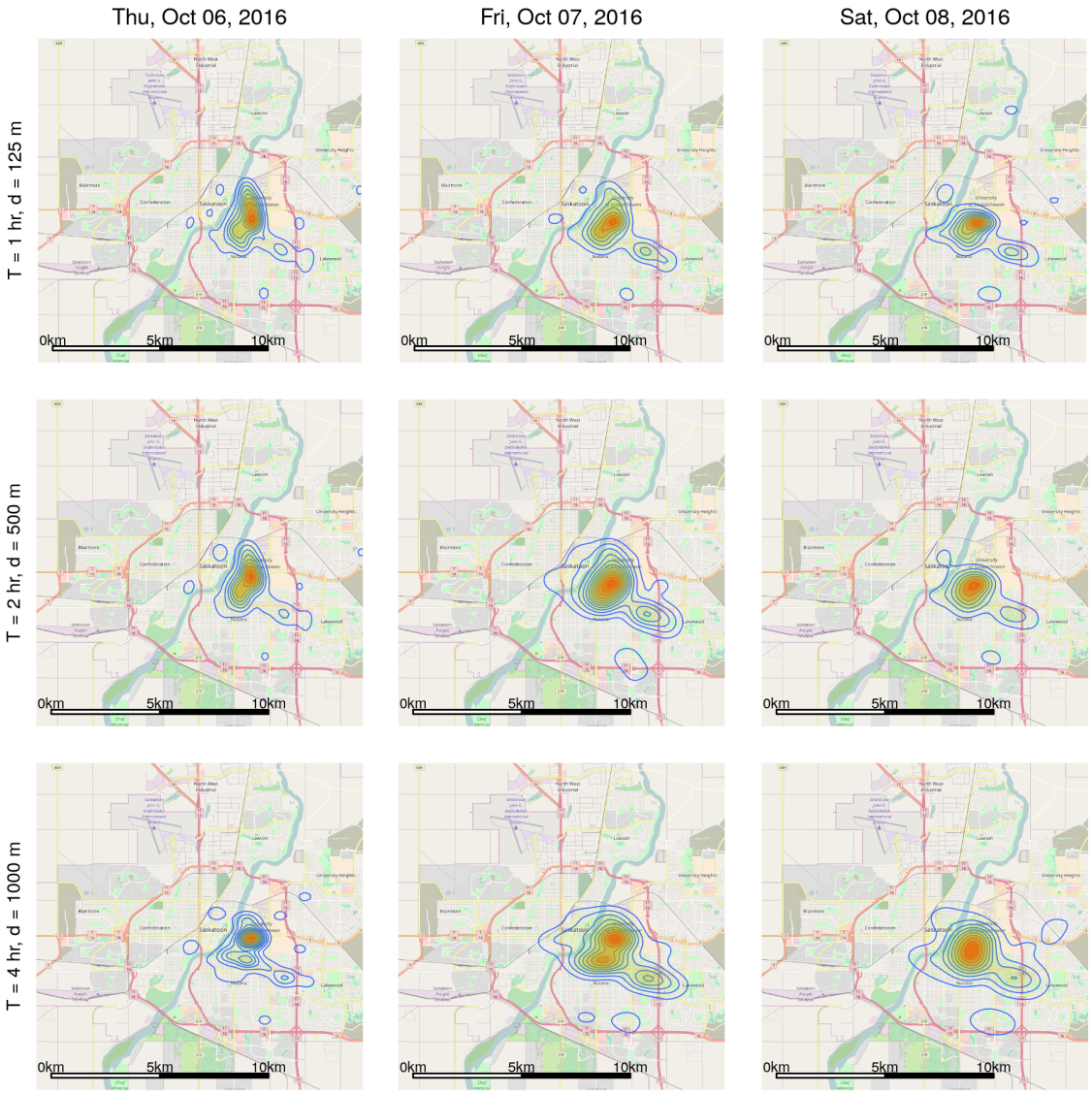


Fig 5.3: Heatmap of the dispersion of participants (undergraduate students) of SHED 8 over three consecutive days in the fall of 2016. Each column represents a day and each row represents a (T, d) pair. The participants visited fairly similar locations, which were centered around the University of Saskatchewan. Dispersion in the weekend is slightly different than the weekdays. Weekdays exhibit visually similar dispersion. The data exhibit visually less dispersion changes than the summer data in Fig. 5.2.

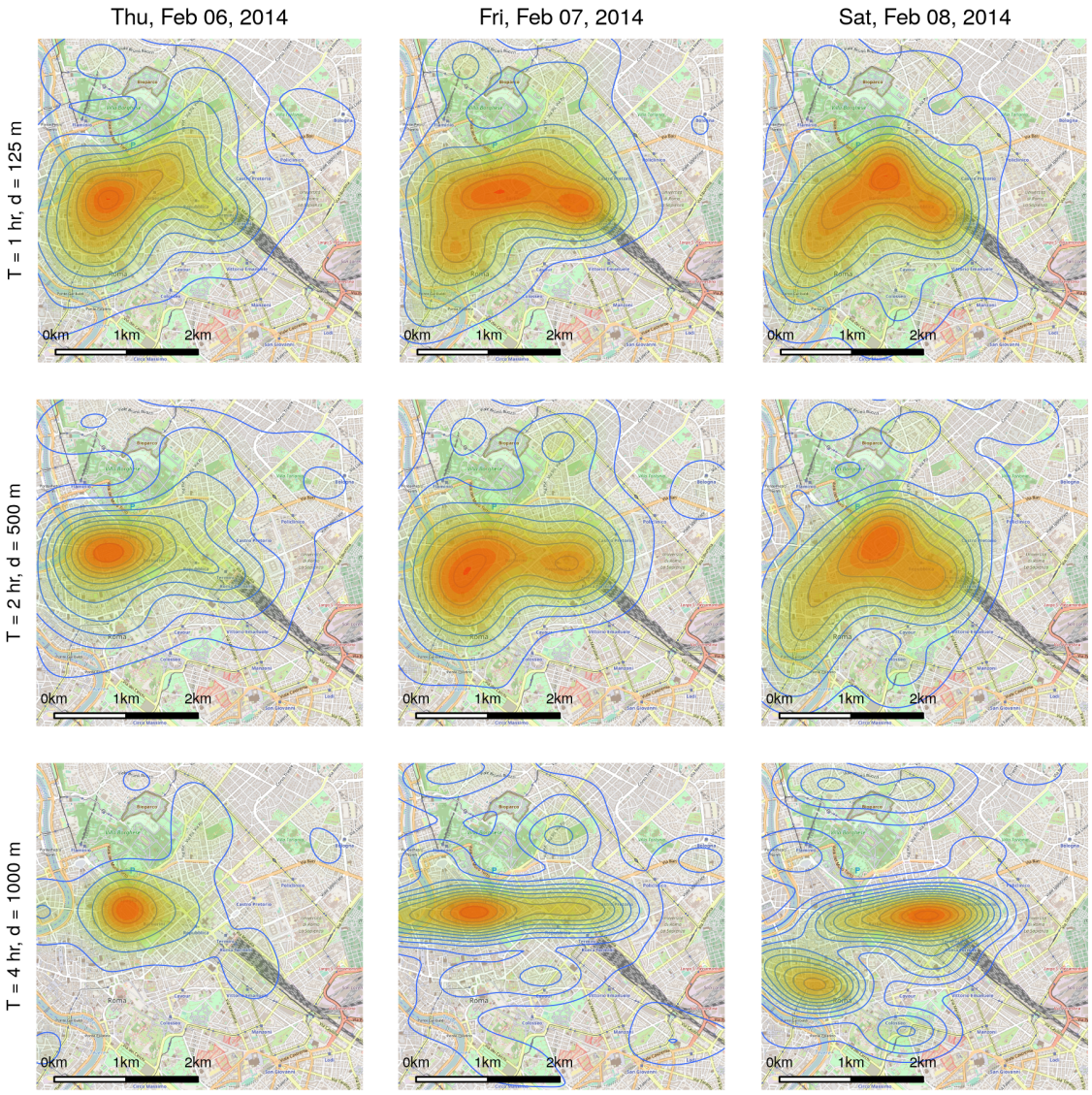


Fig 5.4: Heatmap of the dispersion of taxi cabs tracked in Rome over three consecutive days. The map area is much smaller than the maps shown for undergraduate students in Fig. 5.2 and Fig. 5.3. Taxicabs demonstrate aggregate human movement behaviors. In the weekday, the locations are more concentrated to a hotspot but two hotspots are visible in the weekend.

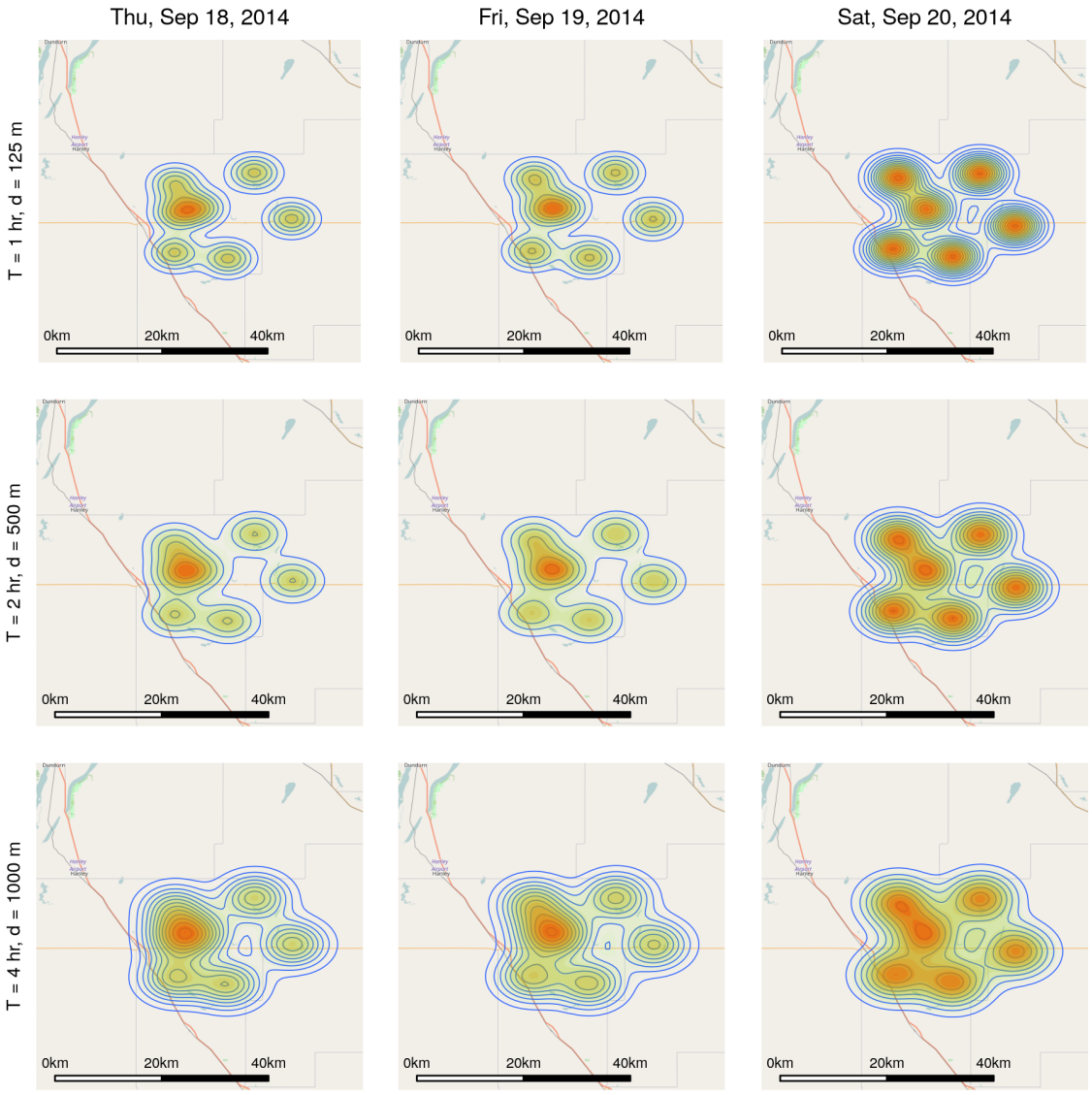


Fig 5.5: Heatmap of the dispersion of the tracked moose over three consecutive days. The hotspots appear visually stable, which indicate steady grazing behavior.

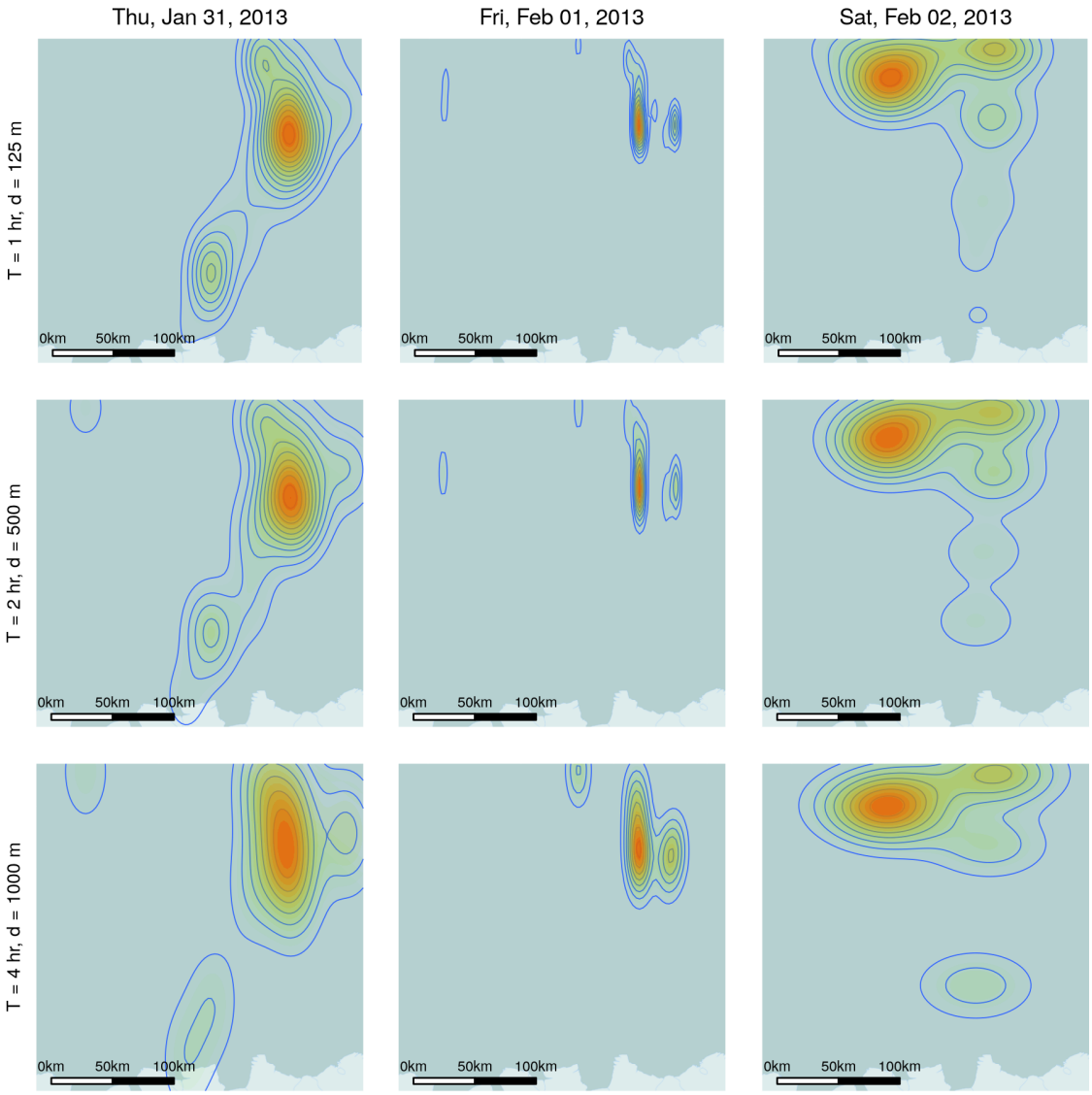


Fig 5.6: Heatmap of the dispersion of Antarctic Petrels over three consecutive days. Their locations change largely from day to day because petrels fly over wide areas and do not stick to specific locations for long.

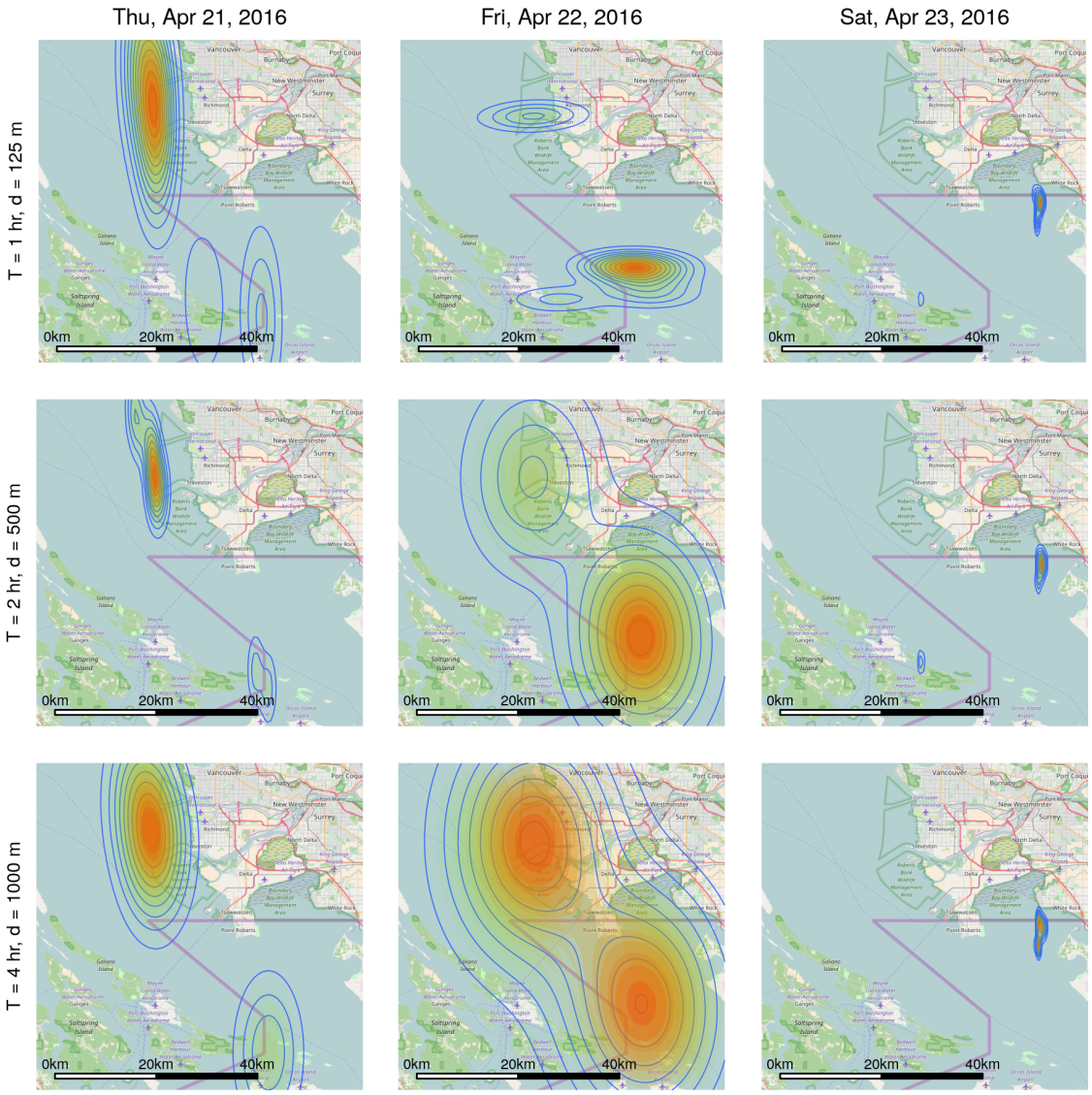


Fig 5.7: Heatmap of the dispersion of ocean drifters over three consecutive days. Similar to petrels, ocean drifters move to different places due to sea currents as time passes, and show no ties to specific locations.

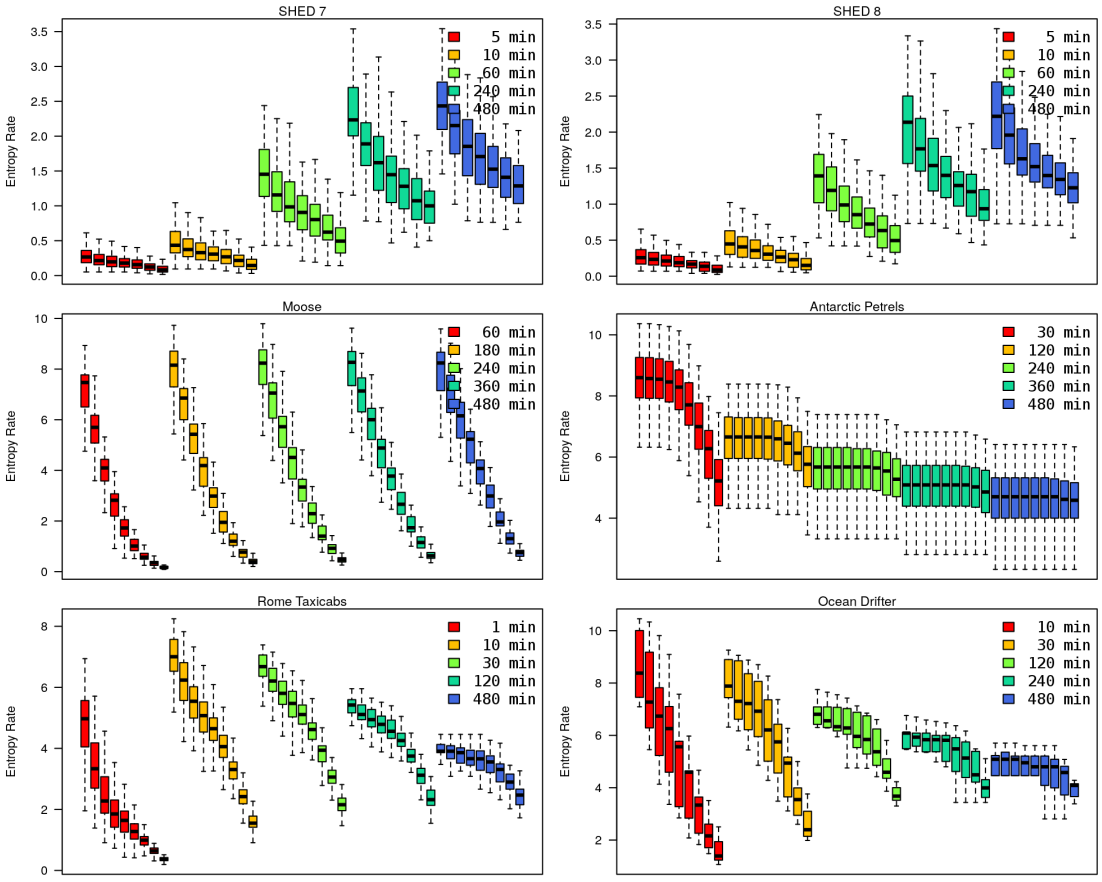


Fig 5.8: Distribution of individual H across (T, d) tuples. The color of a boxplot represent the value of T, as the legends indicate. Values of d increase to the right within a color band. Entropy rate always decreases as d increases; however, entropy rate decreases in different manners. The fall in entropy rate in faster moving agents (petrels, taxicabs, and ocean drifters) are basically different than that for comparatively slower agents (undergraduate students and moose).

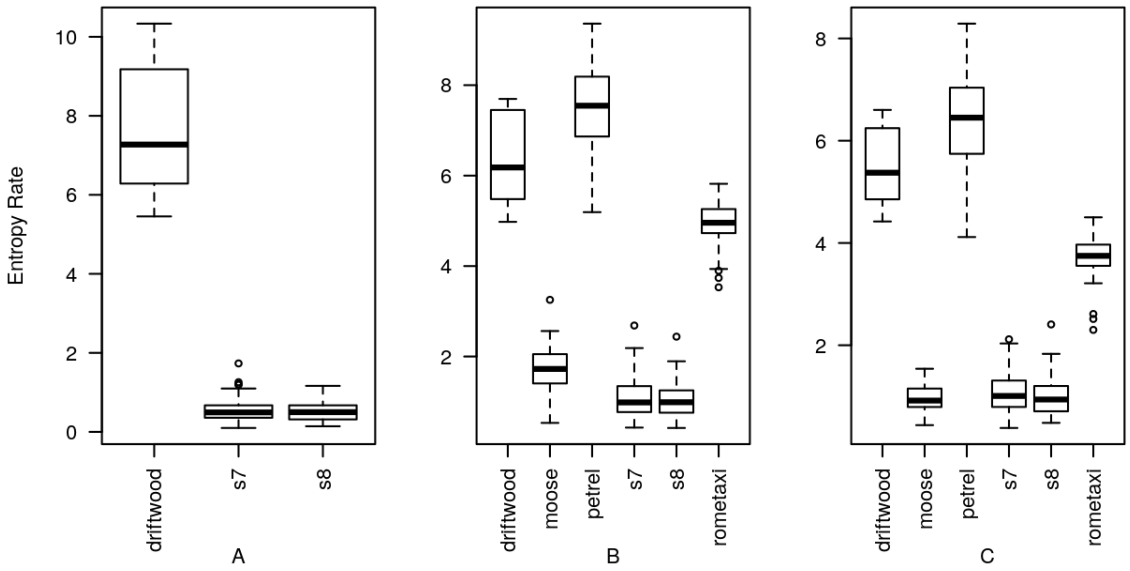


Fig 5.9: Comparison of individual entropy rate distributions of the datasets: **(A)** $(T, d) = (10min, 62.5m)$ **(B)** $(T, d) = (60min, 250m)$ **(C)** $(T, d) = (4hrs, 1km)$. Undergraduate students always demonstrate lower entropy rate than other agents, whereas fast moving petrels have the highest. The order among datasets is retained across spatio-temporal resolutions of location measurement.

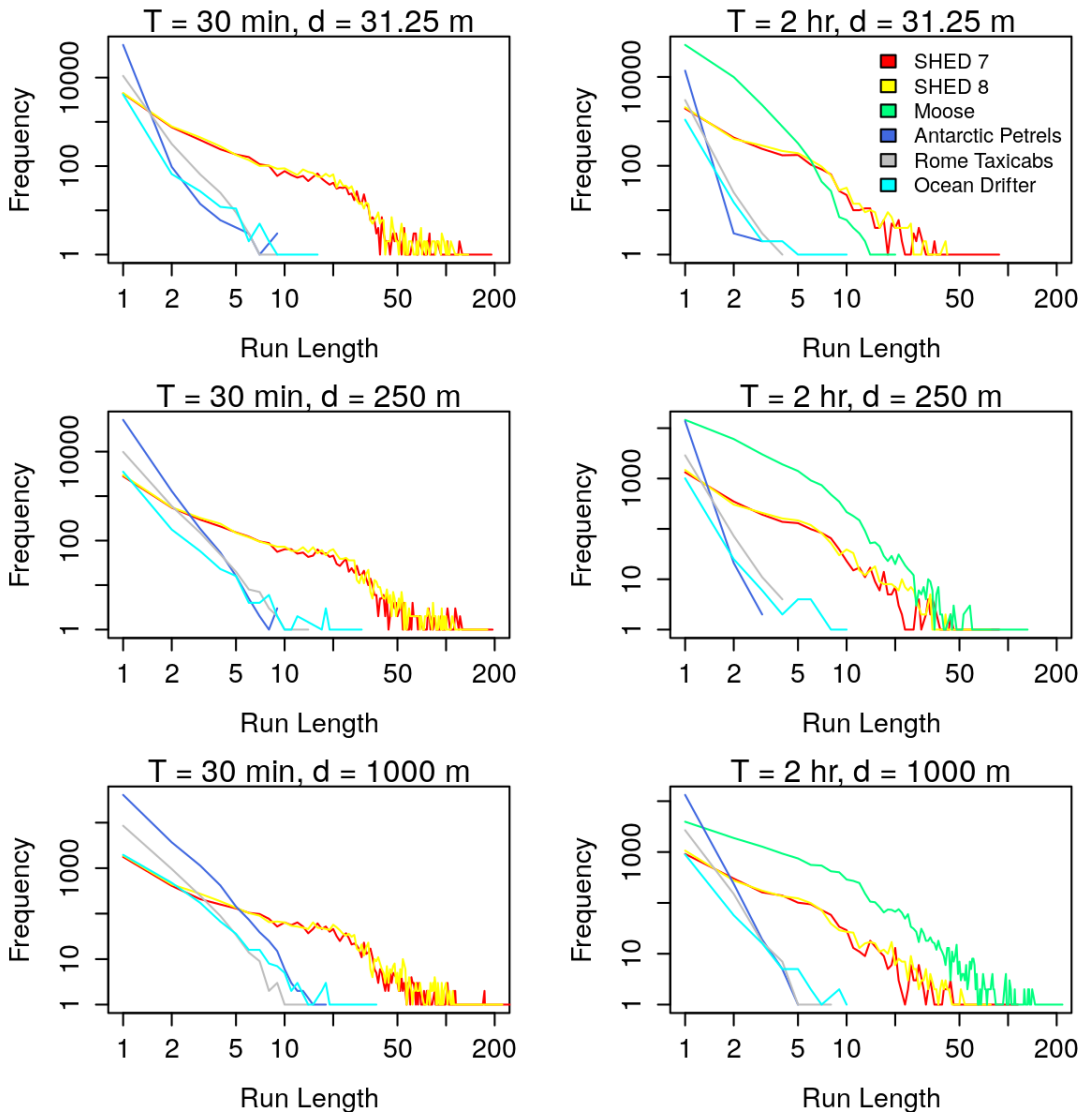


Fig 5.10: Run length distributions of the datasets, aggregated across all participants.

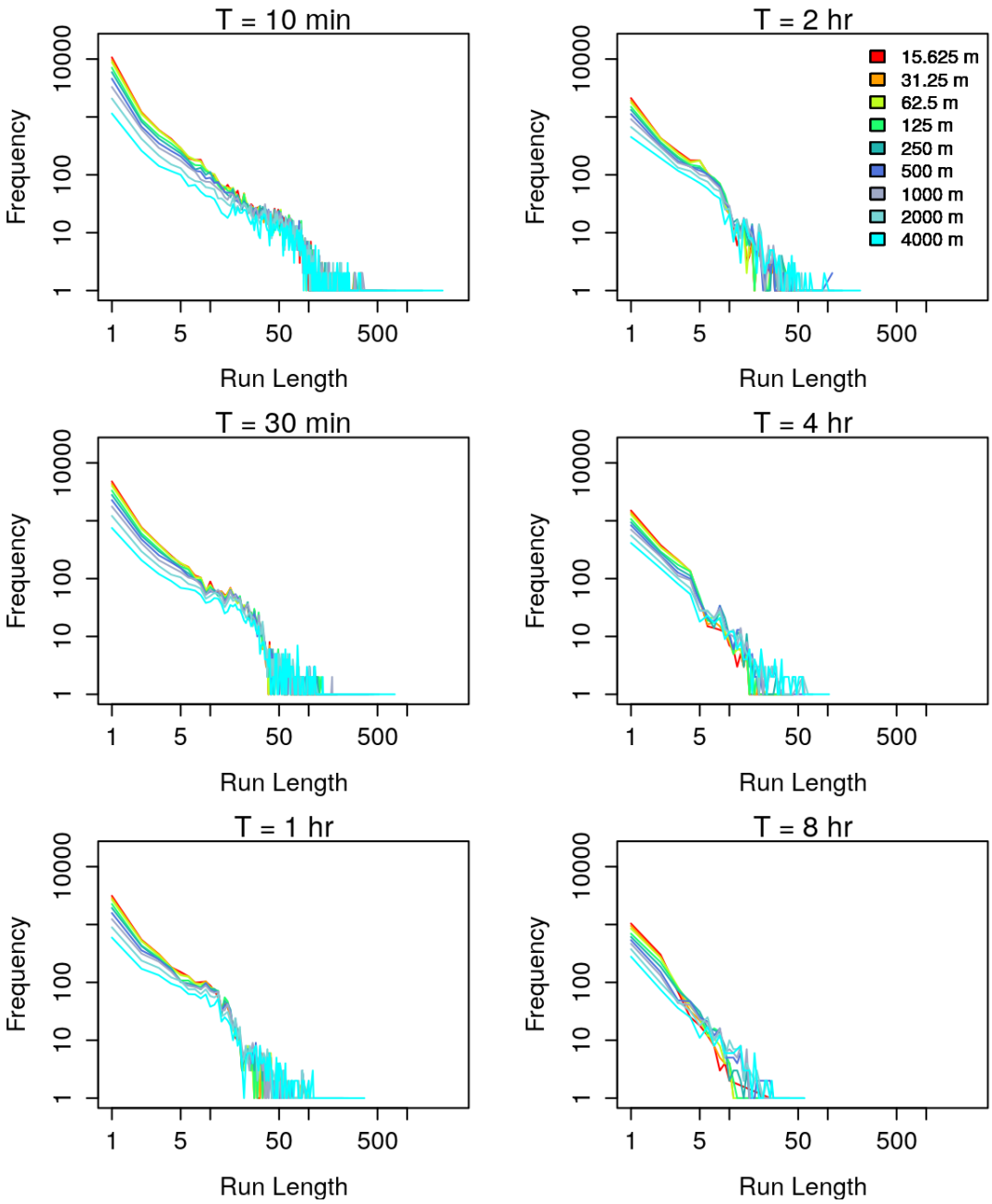


Fig 5.11: Aggregate run length distributions of the SHED 7 Dataset by T

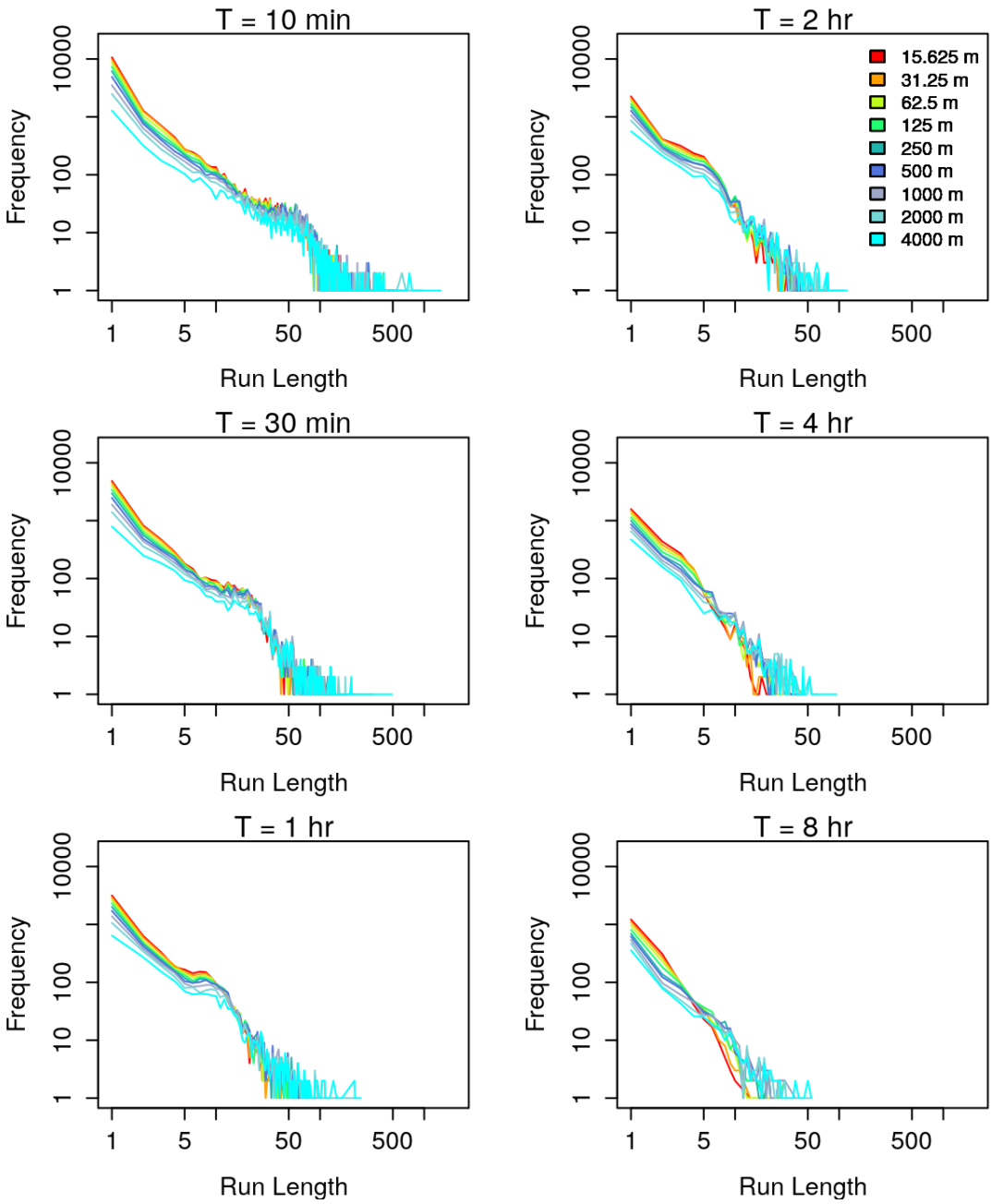


Fig 5.12: Aggregate run length distributions of the SHED 8 Dataset by T

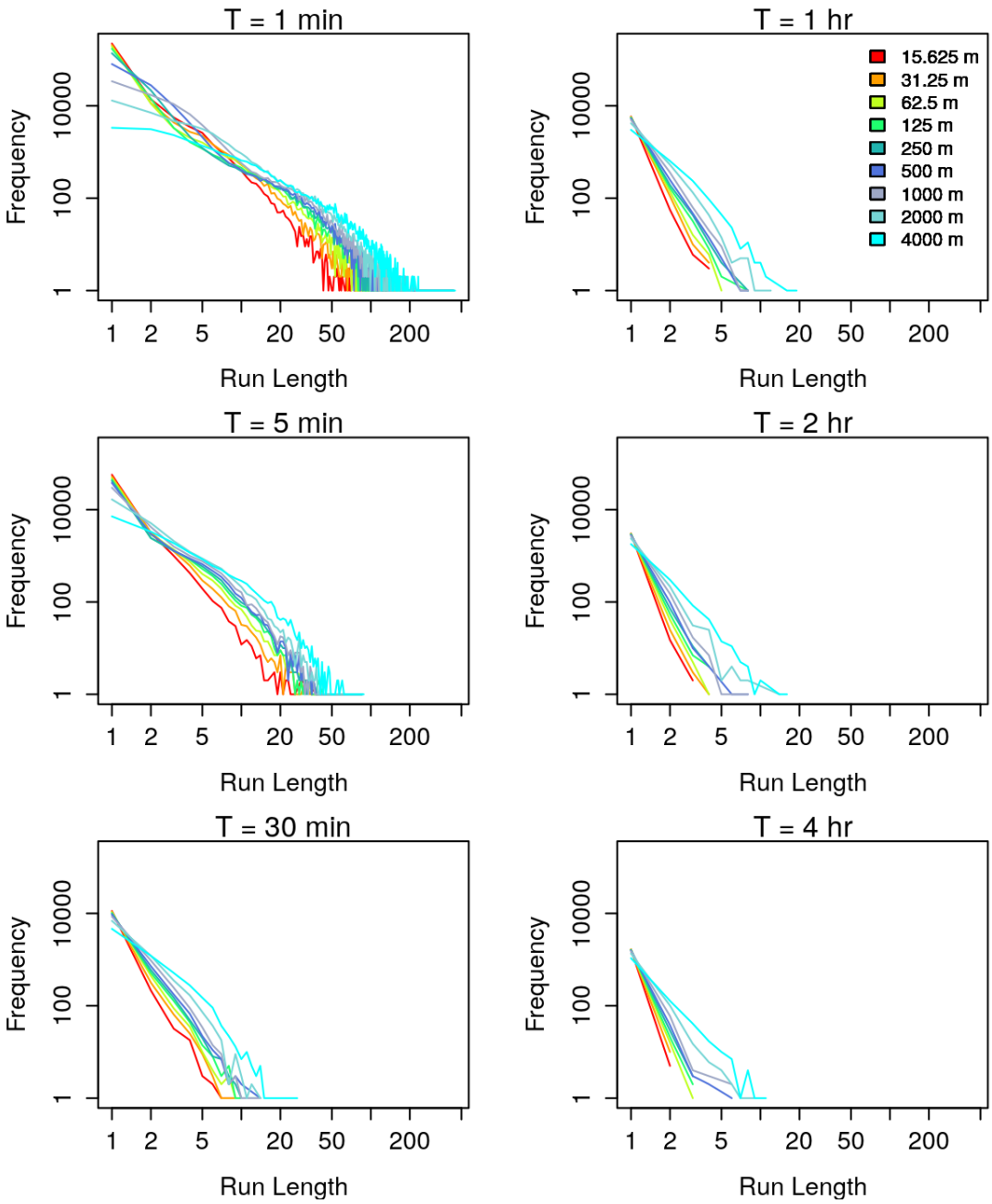


Fig 5.13: Aggregate run length distributions of the Taxi Dataset by T

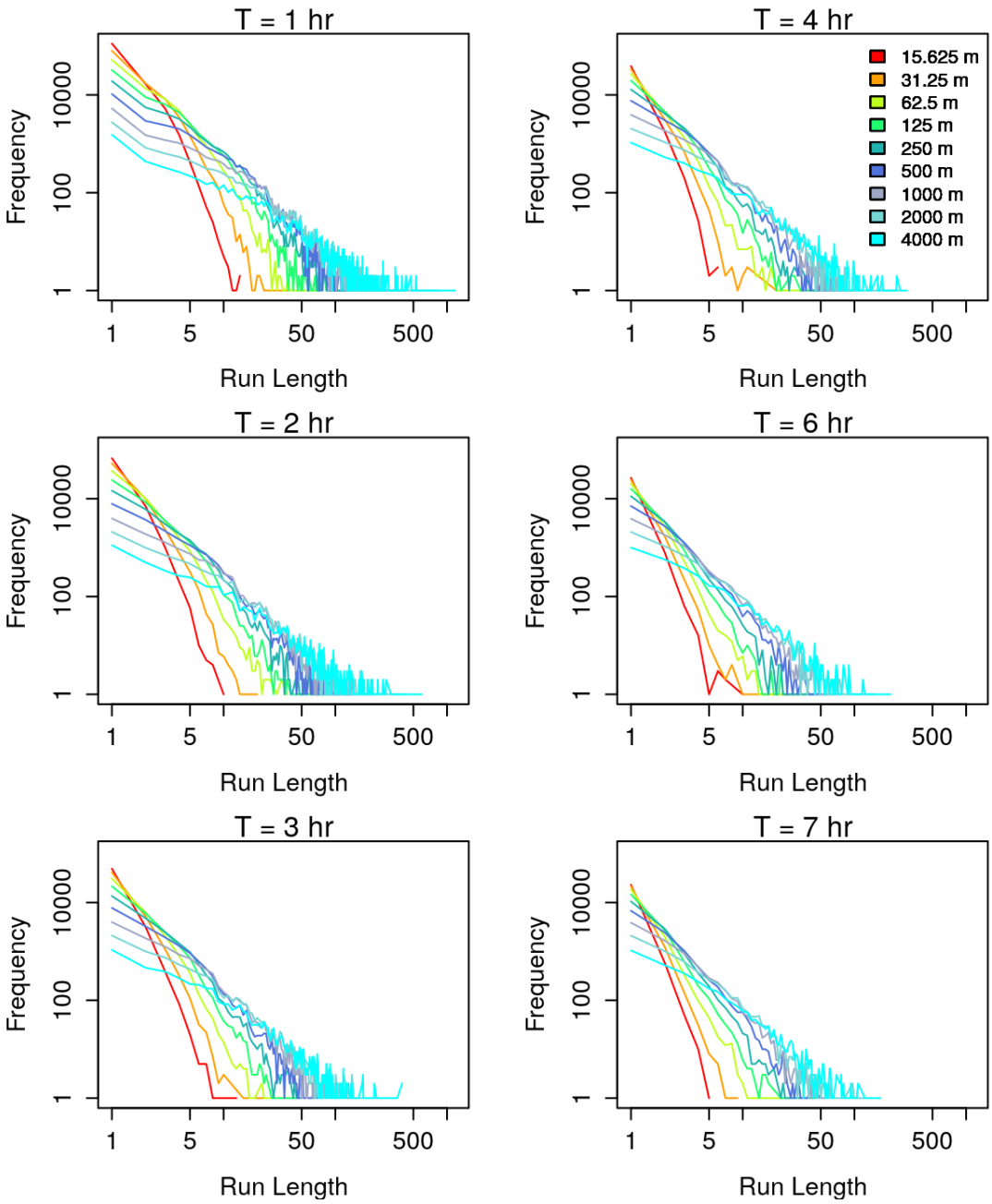


Fig 5.14: Aggregate run length distributions of the Moose Dataset by T

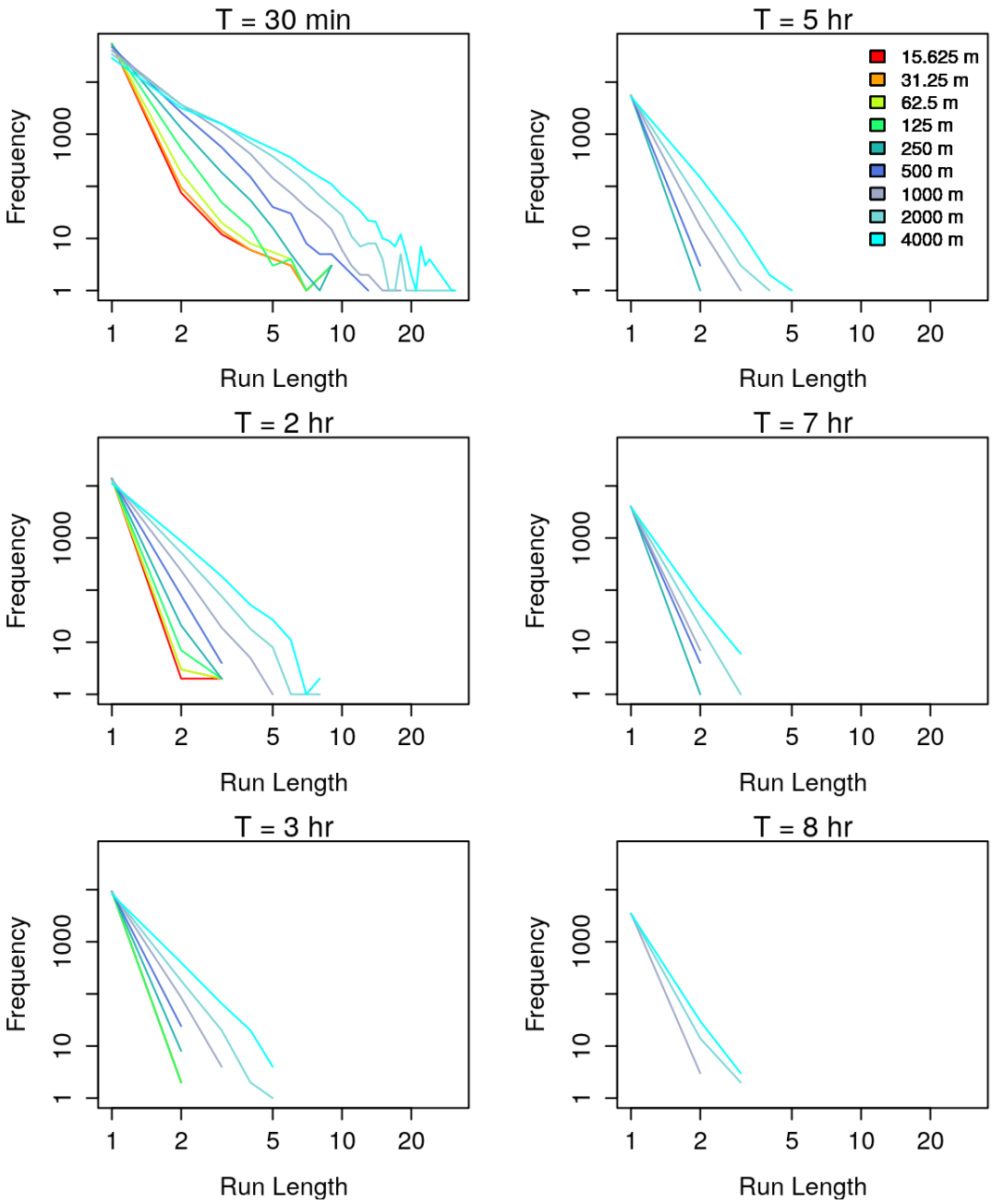


Fig 5.15: Aggregate run length distributions of the Petrel Dataset by T

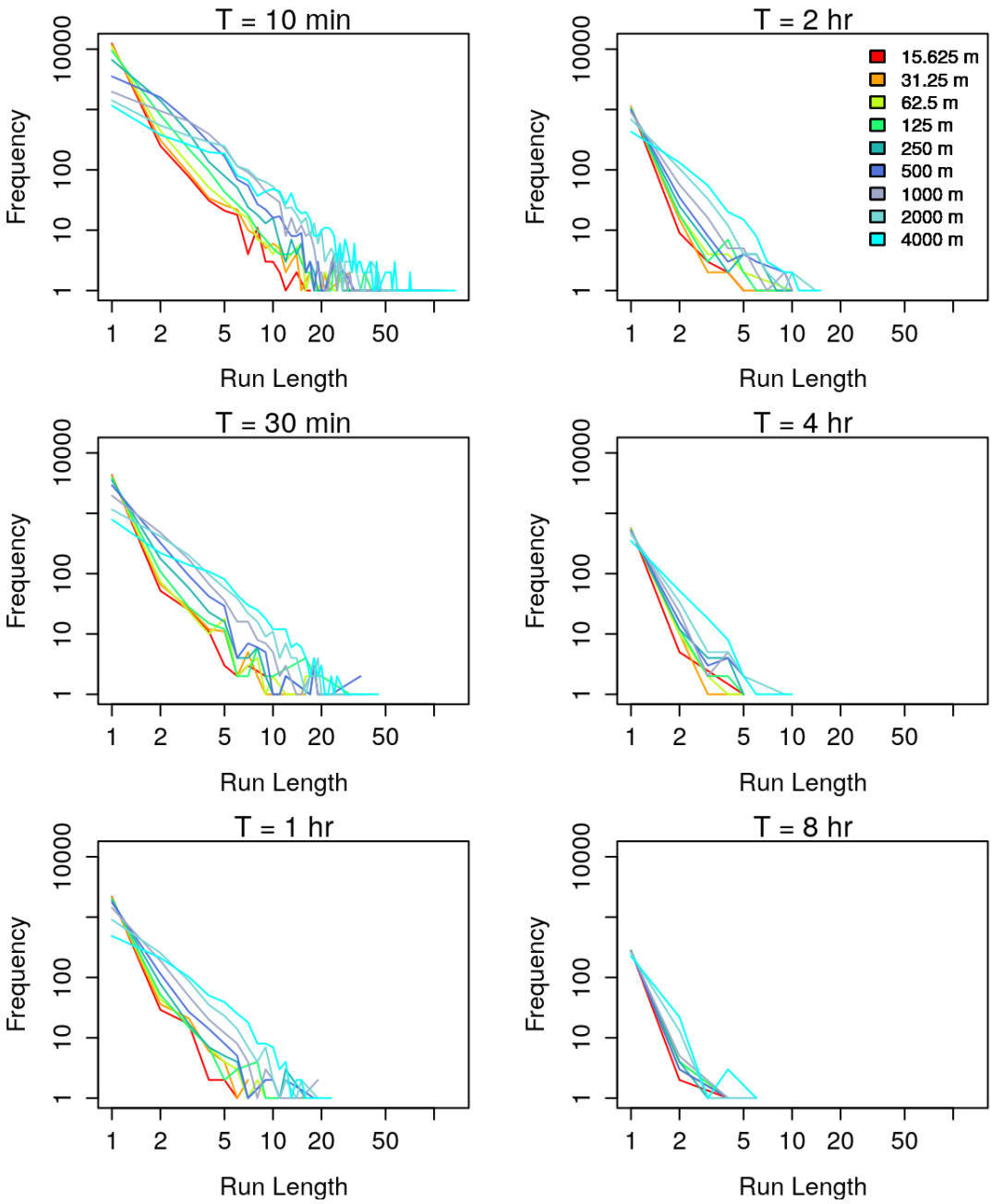


Fig 5.16: Aggregate run length distributions of the Ocean Drifter Dataset by T

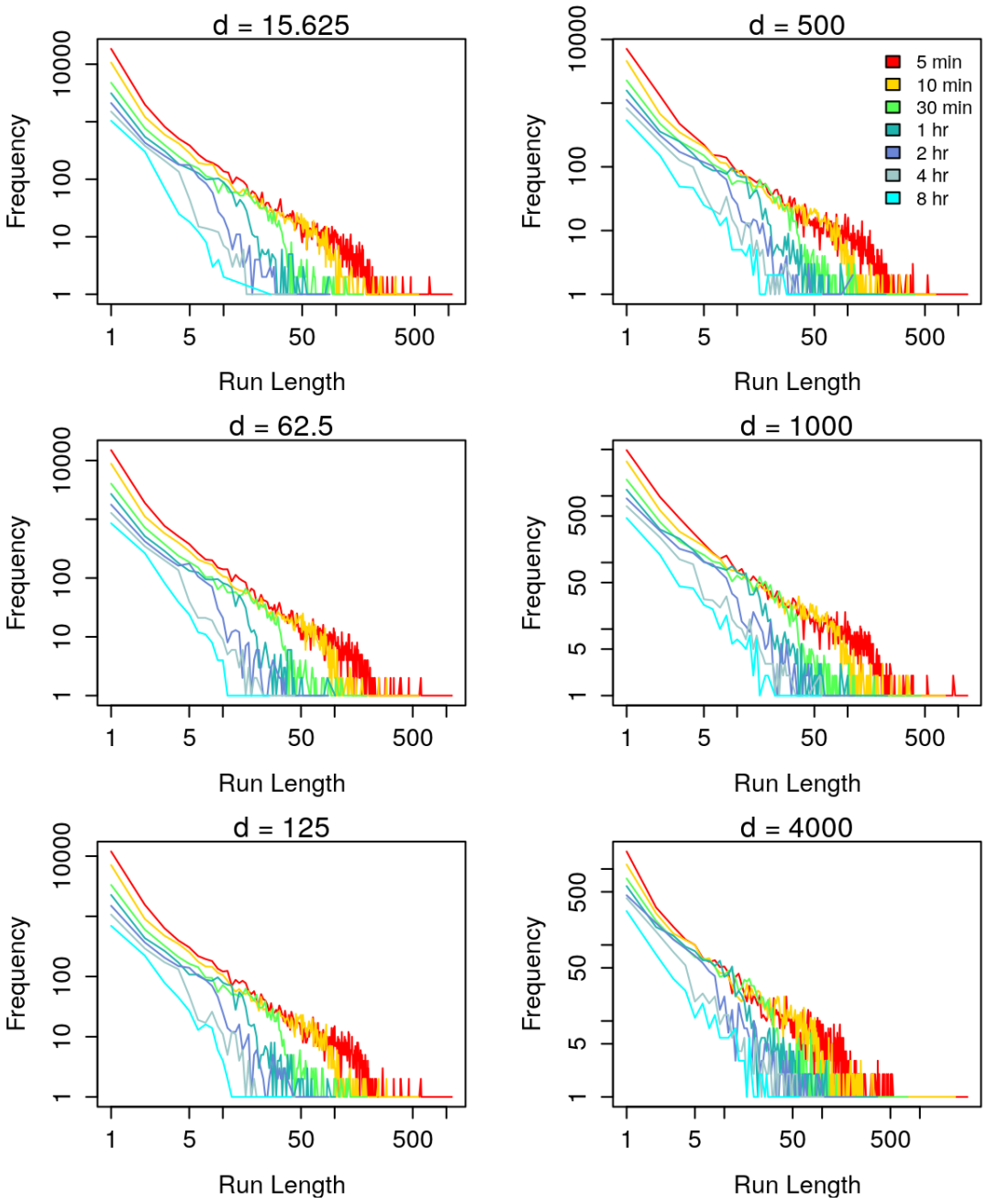


Fig 5.17: Aggregate run length distributions of the SHED 7 Dataset by d

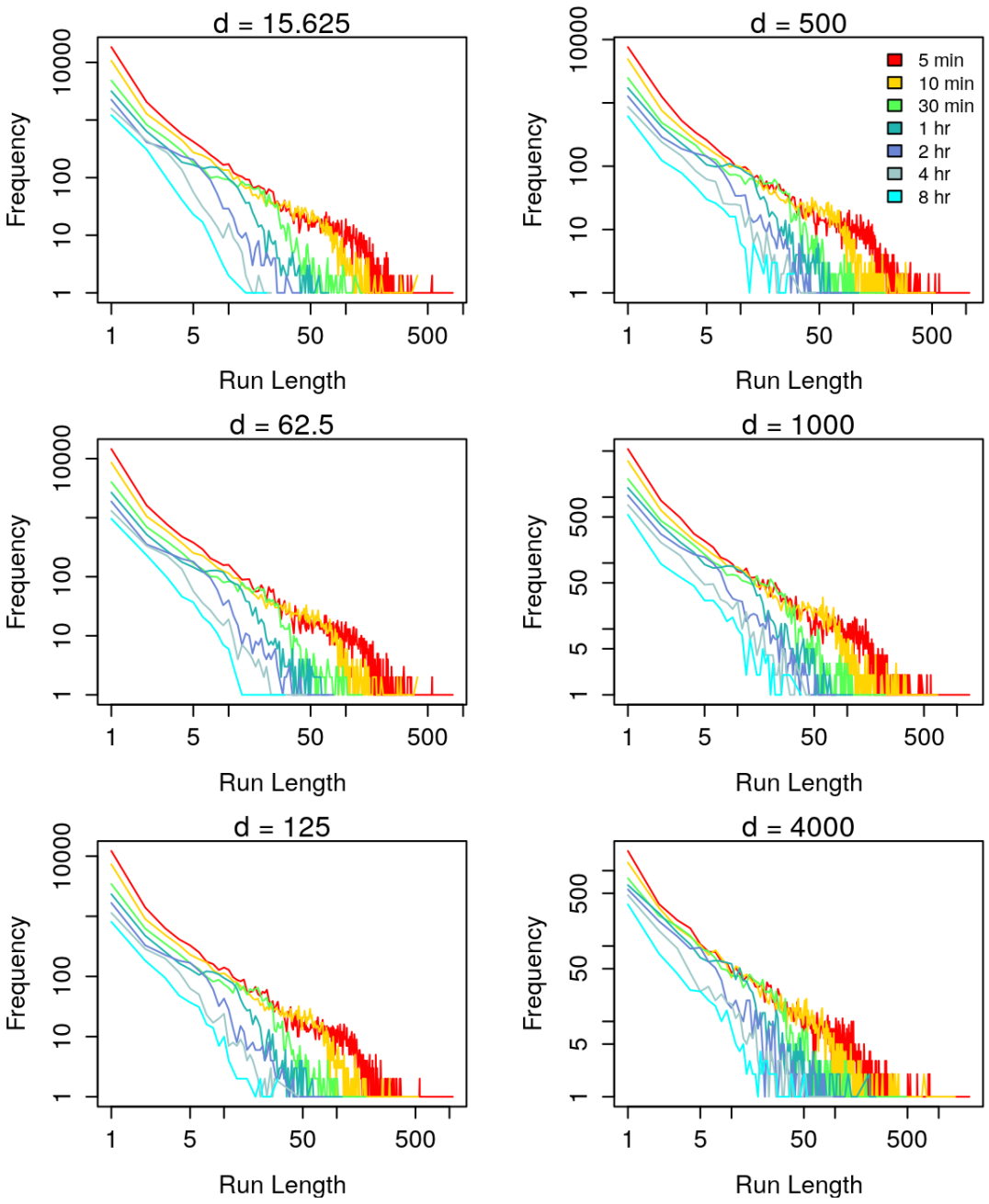


Fig 5.18: Aggregate run length distributions of the SHED 8 Dataset by d

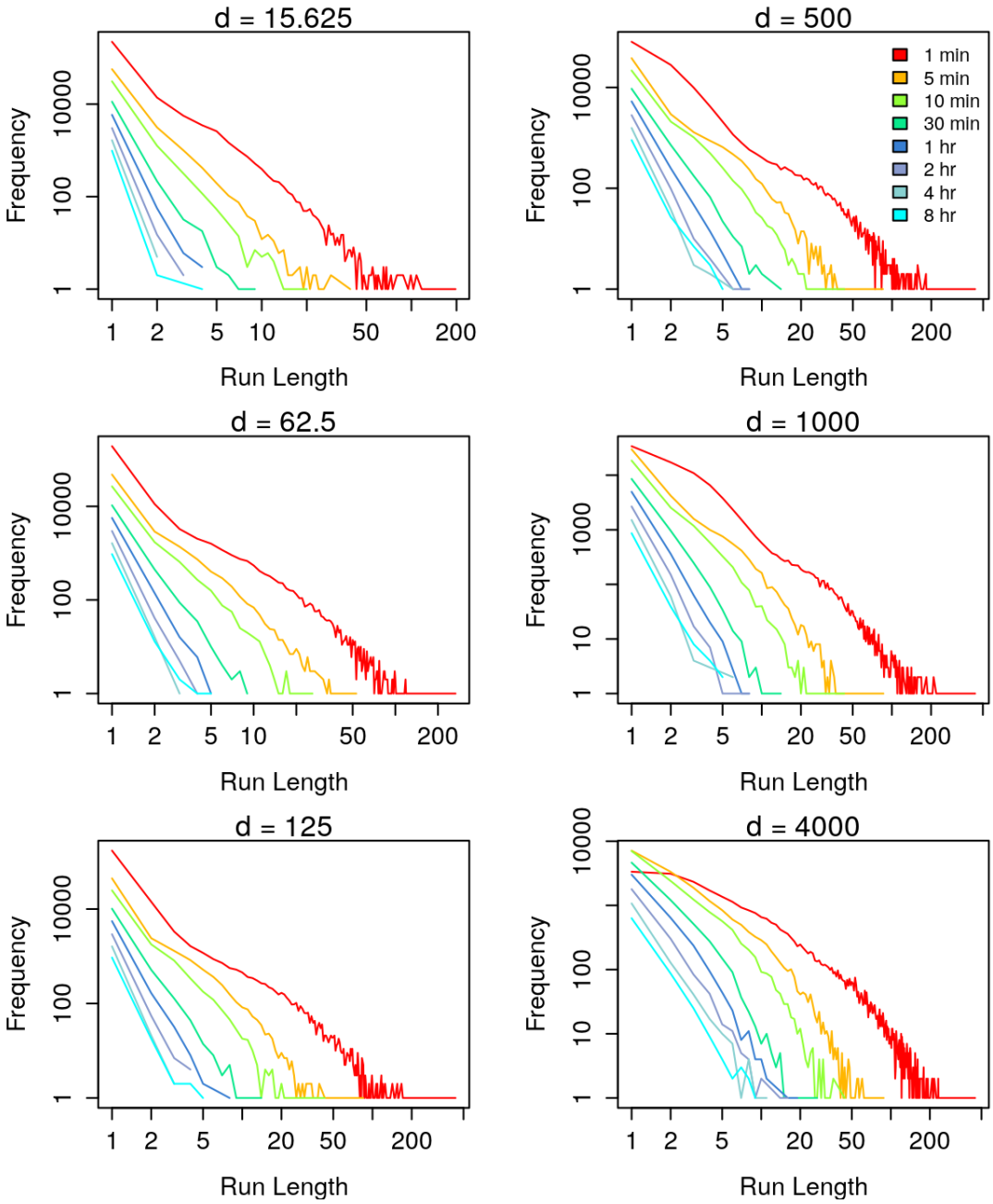


Fig 5.19: RunAggregate run length distributions of the Taxi Dataset by d

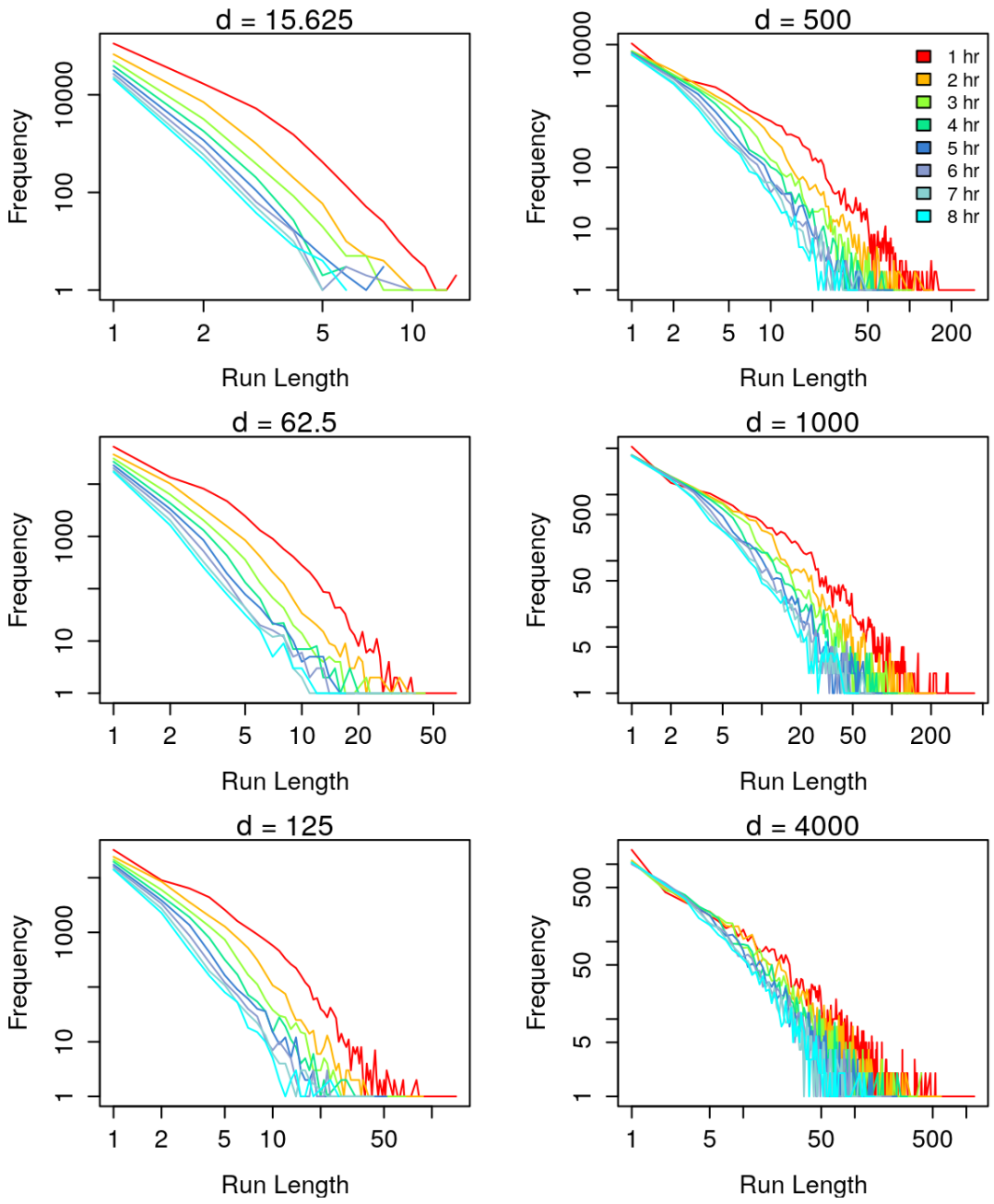


Fig 5.20: Aggregate run length distributions of the Moose Dataset by d

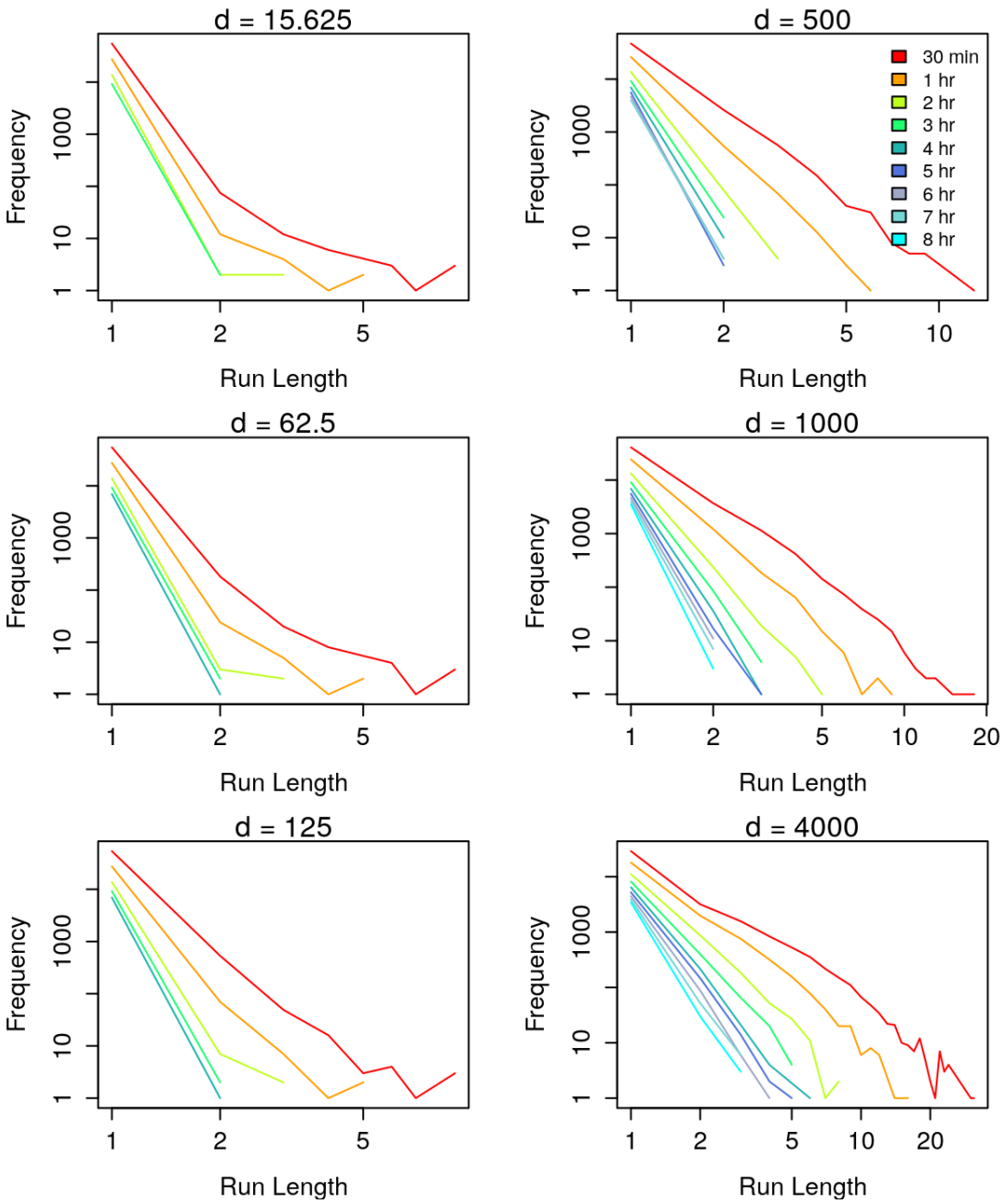


Fig 5.21: Aggregate run length distributions of the Petrel Dataset by d

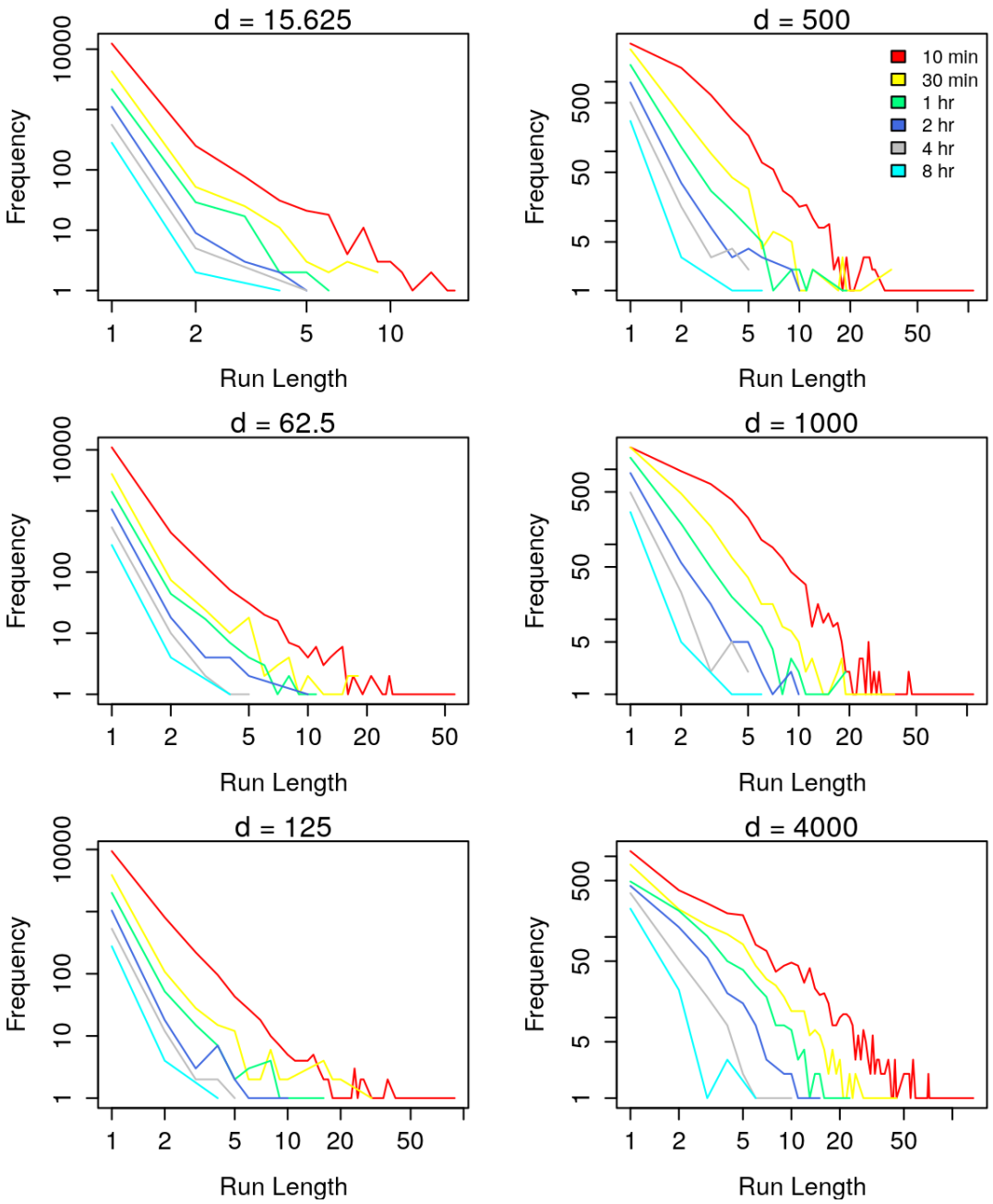


Fig 5.22: Aggregate run length distributions of the Ocean Drifter Dataset by d

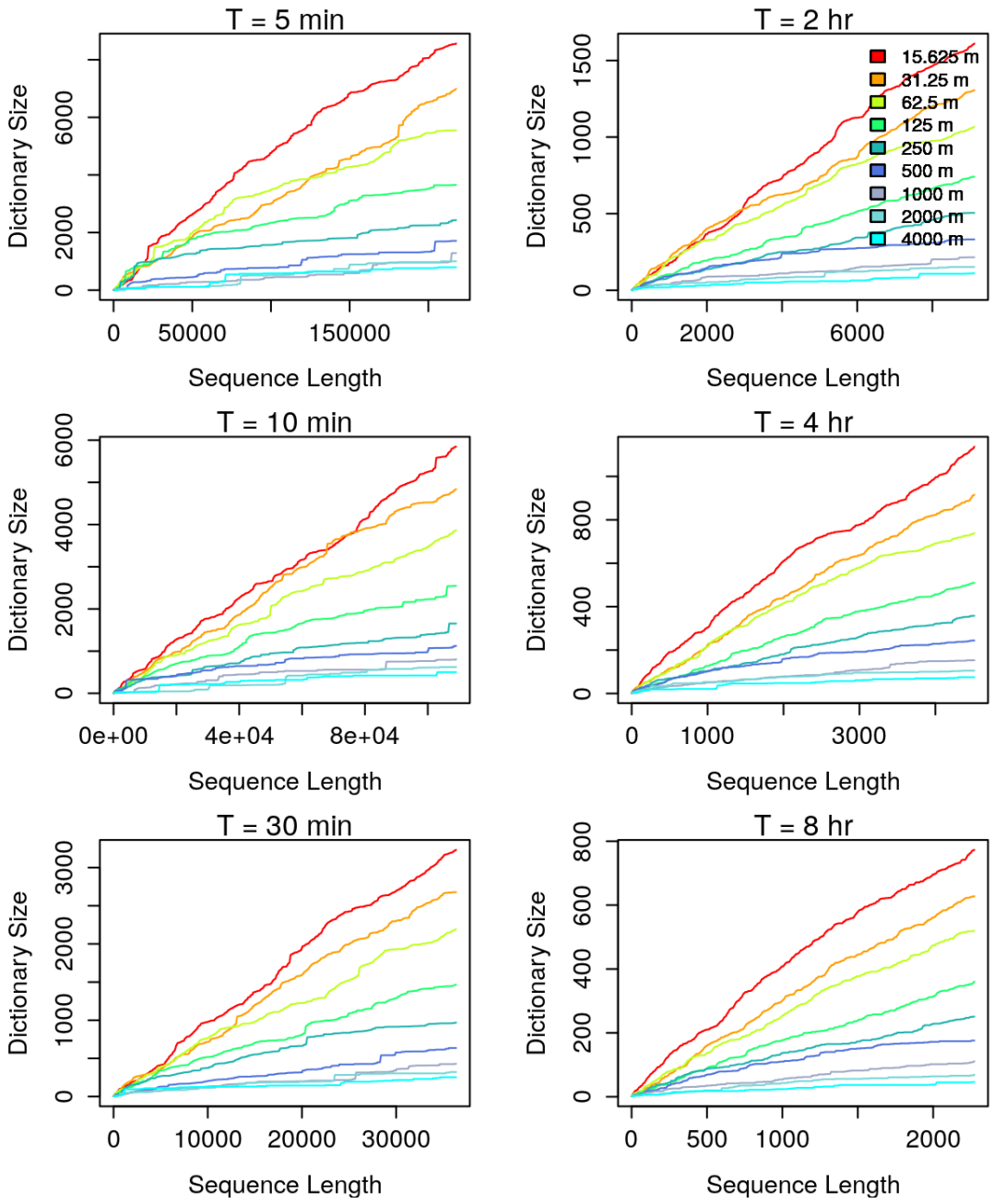


Fig 5.23: Growth of dictionary size in SHED 7

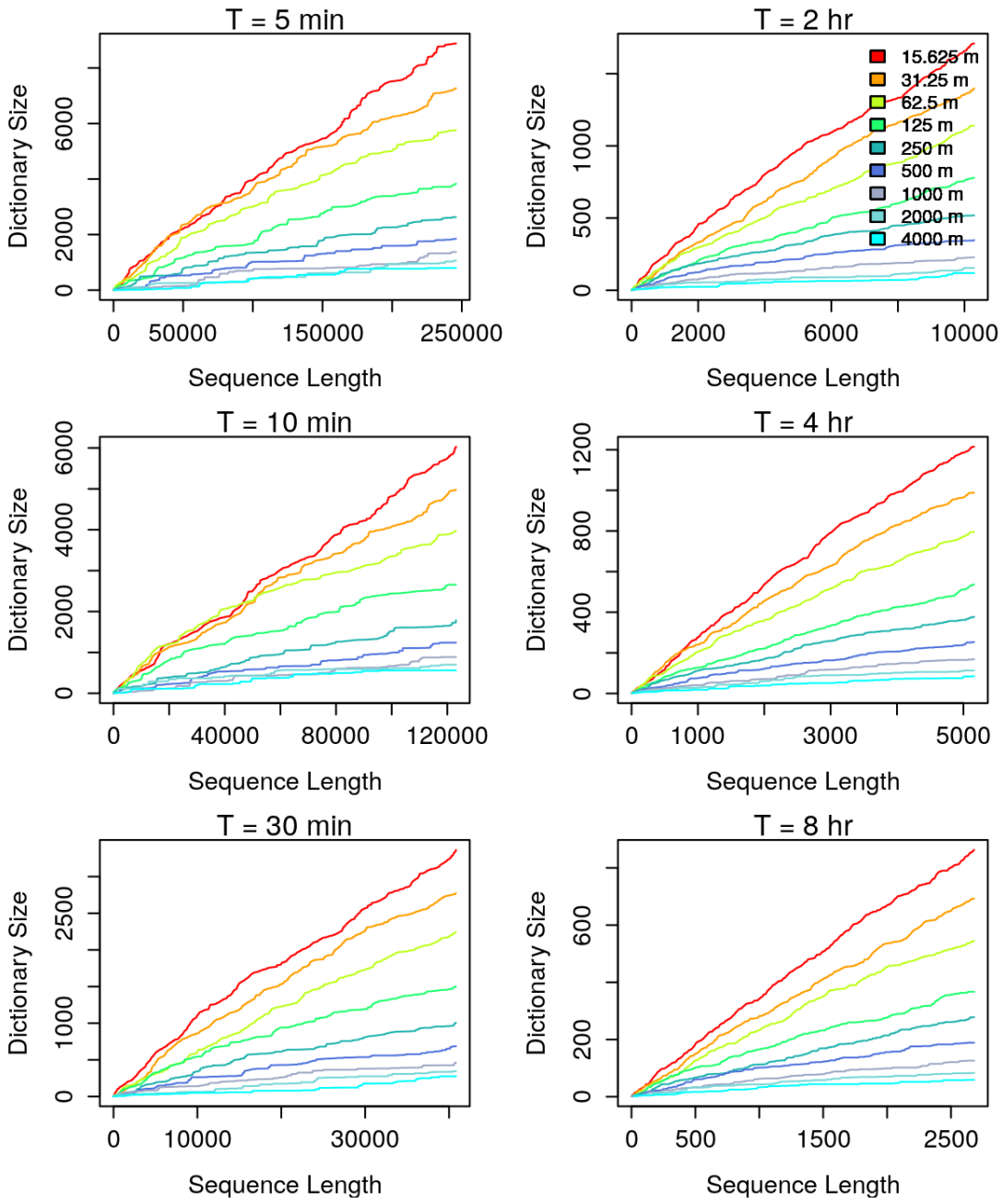


Fig 5.24: Growth of dictionary size in SHED 8

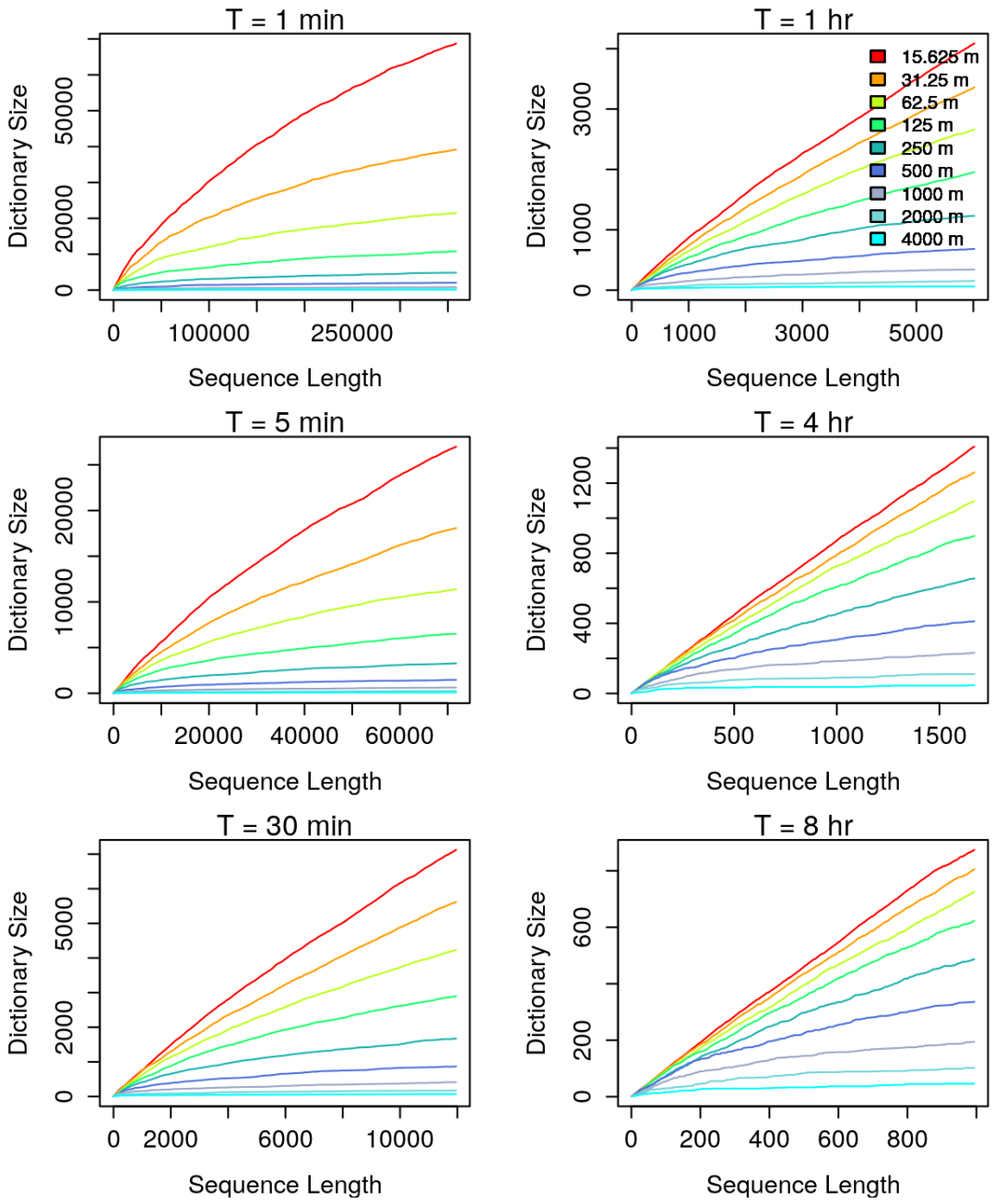


Fig 5.25: Growth of dictionary size in Taxi Cab Dataset

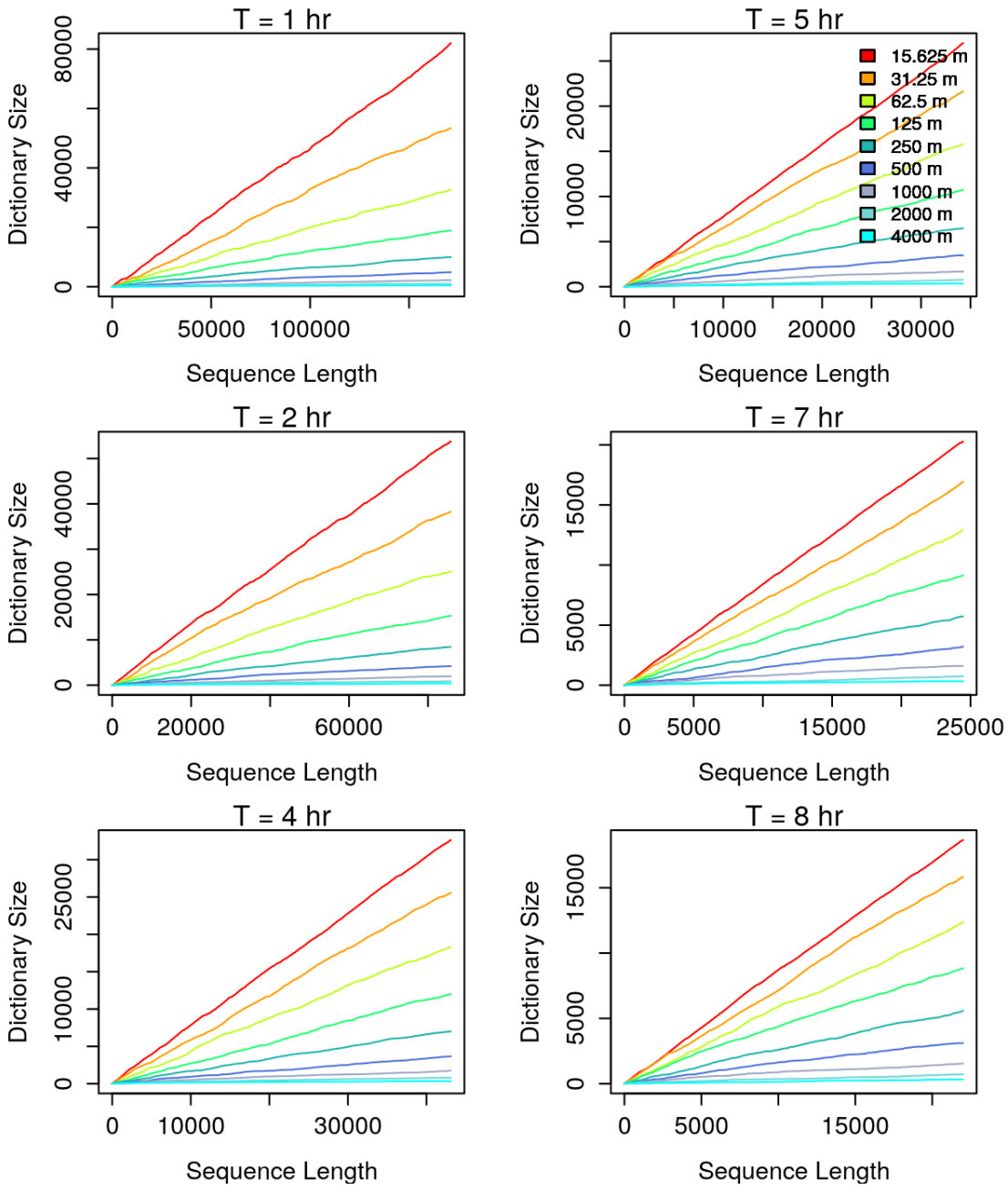


Fig 5.26: Growth of dictionary size in Moose Dataset

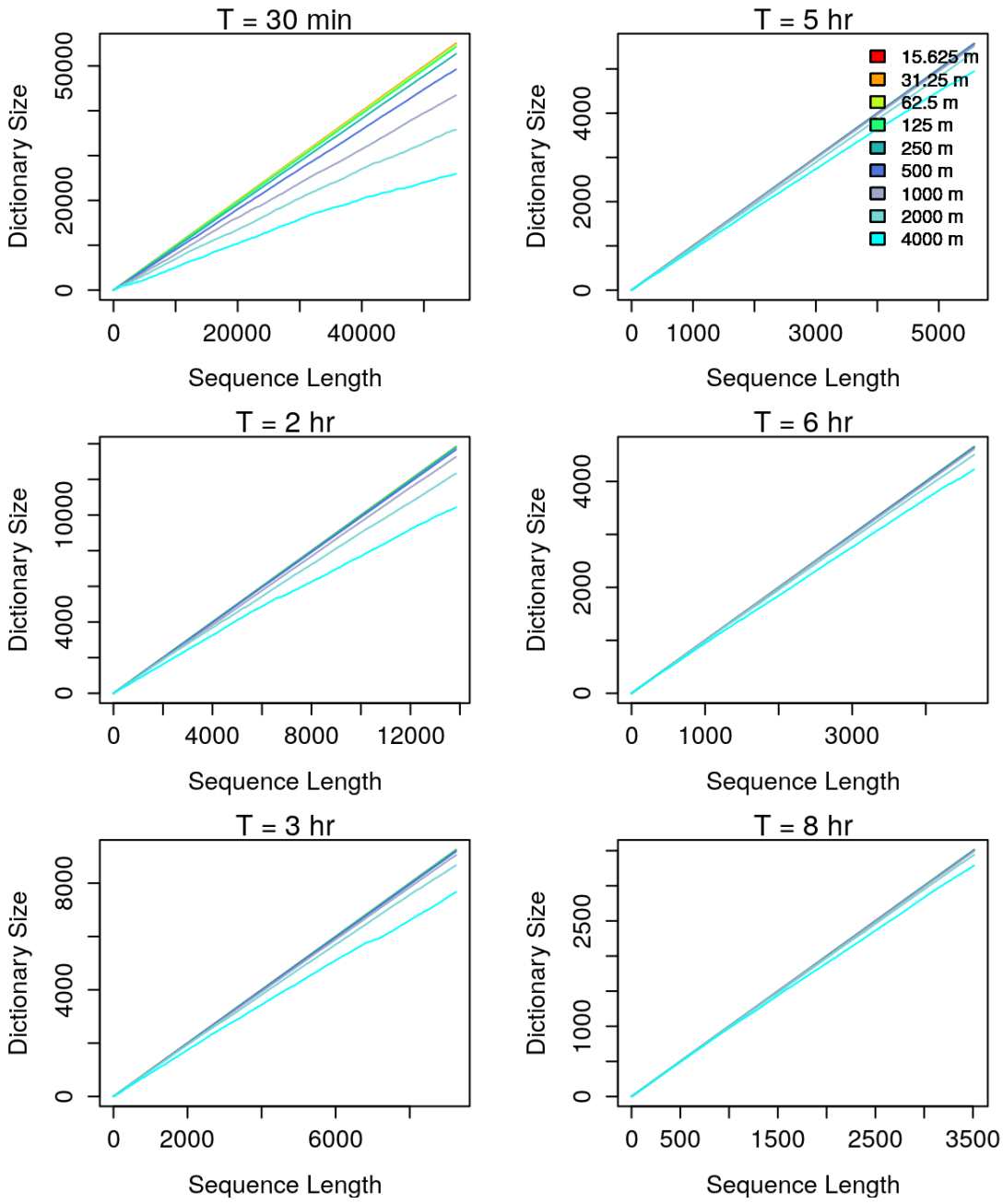


Fig 5.27: Growth of dictionary size in Antarctic Petrel Dataset

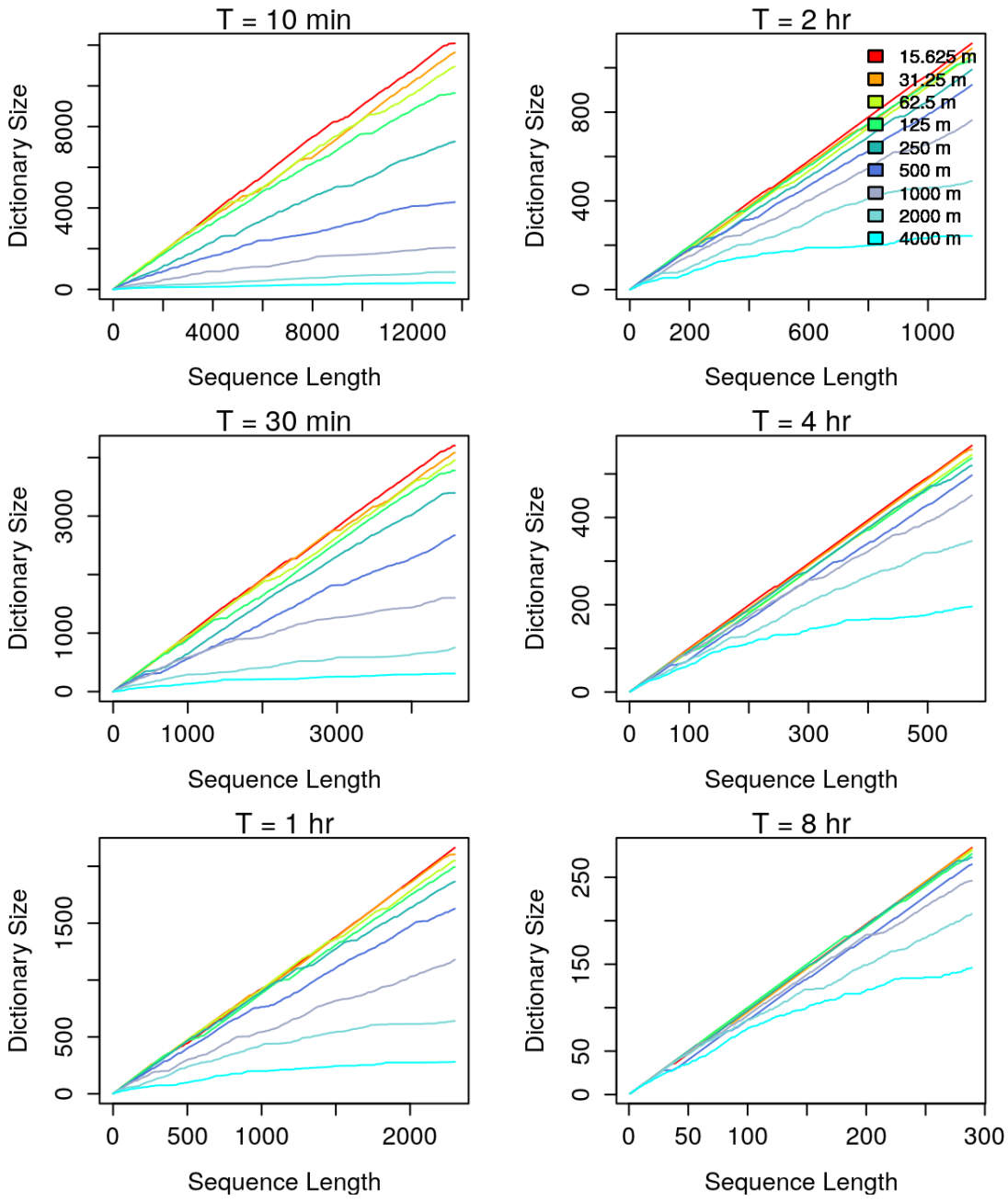


Fig 5.28: Growth of dictionary size in Ocean Drifter Dataset

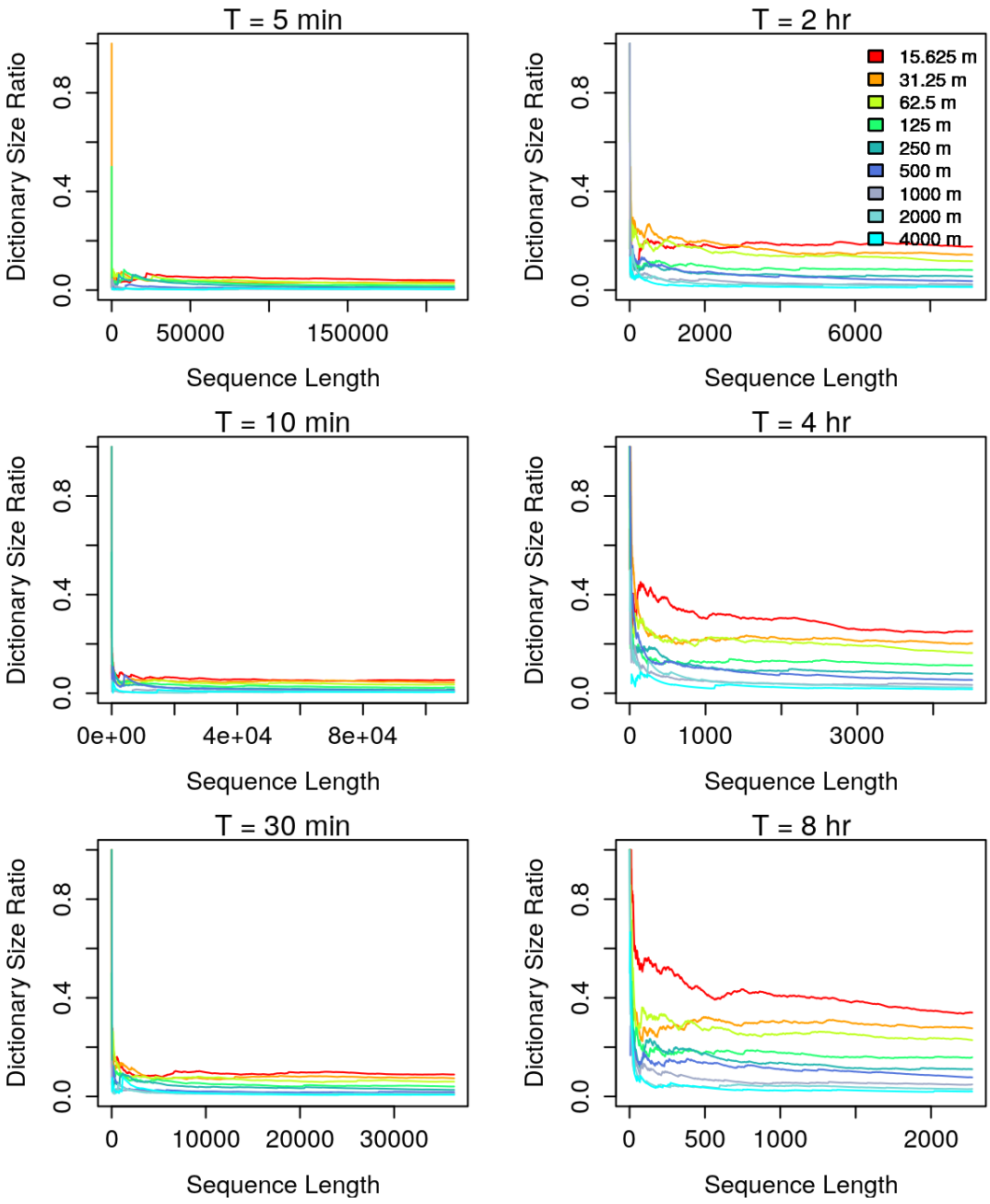


Fig 5.29: Growth of dictionary size ratio in SHED 7

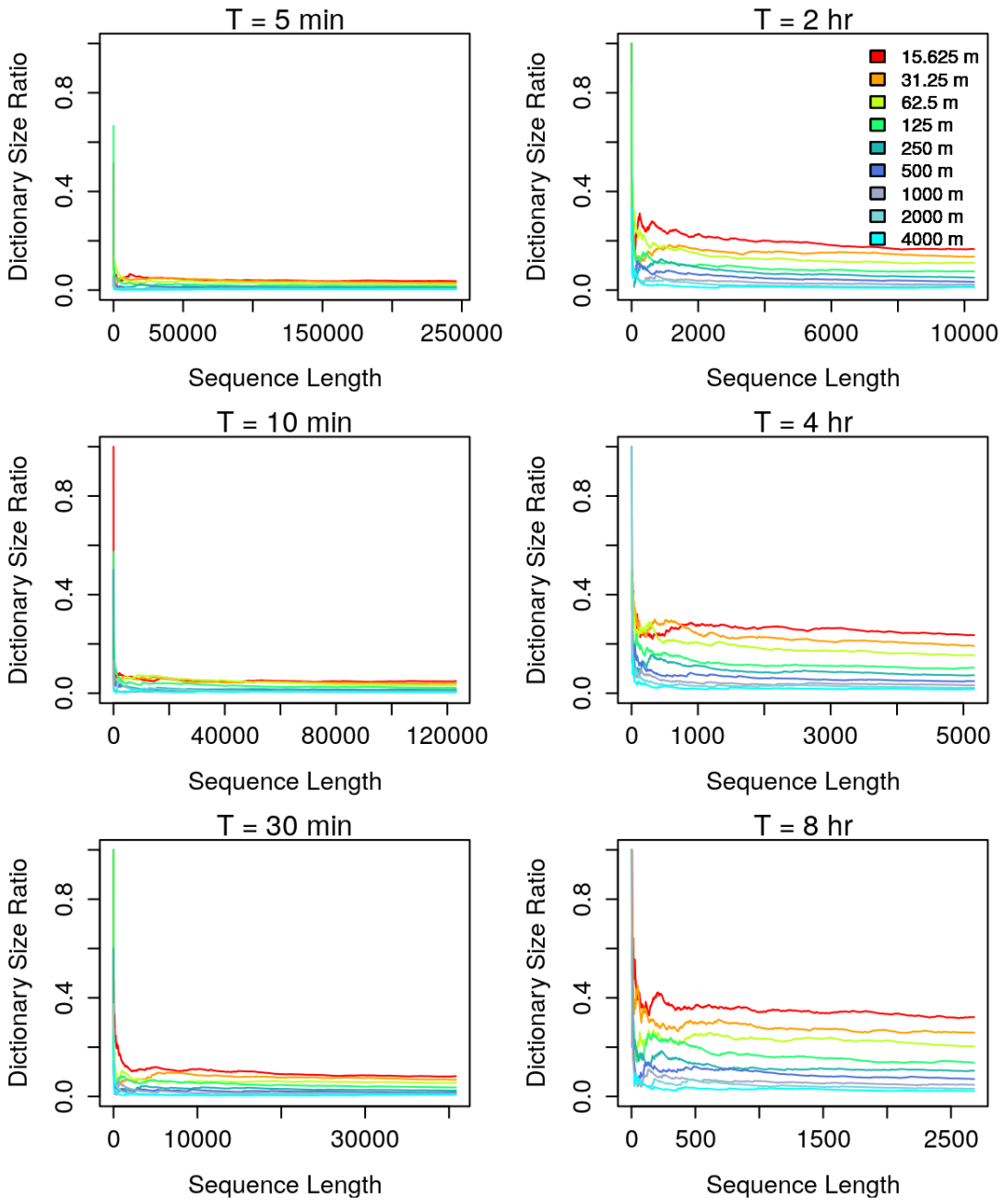


Fig 5.30: Growth of dictionary size ratio in SHED 8

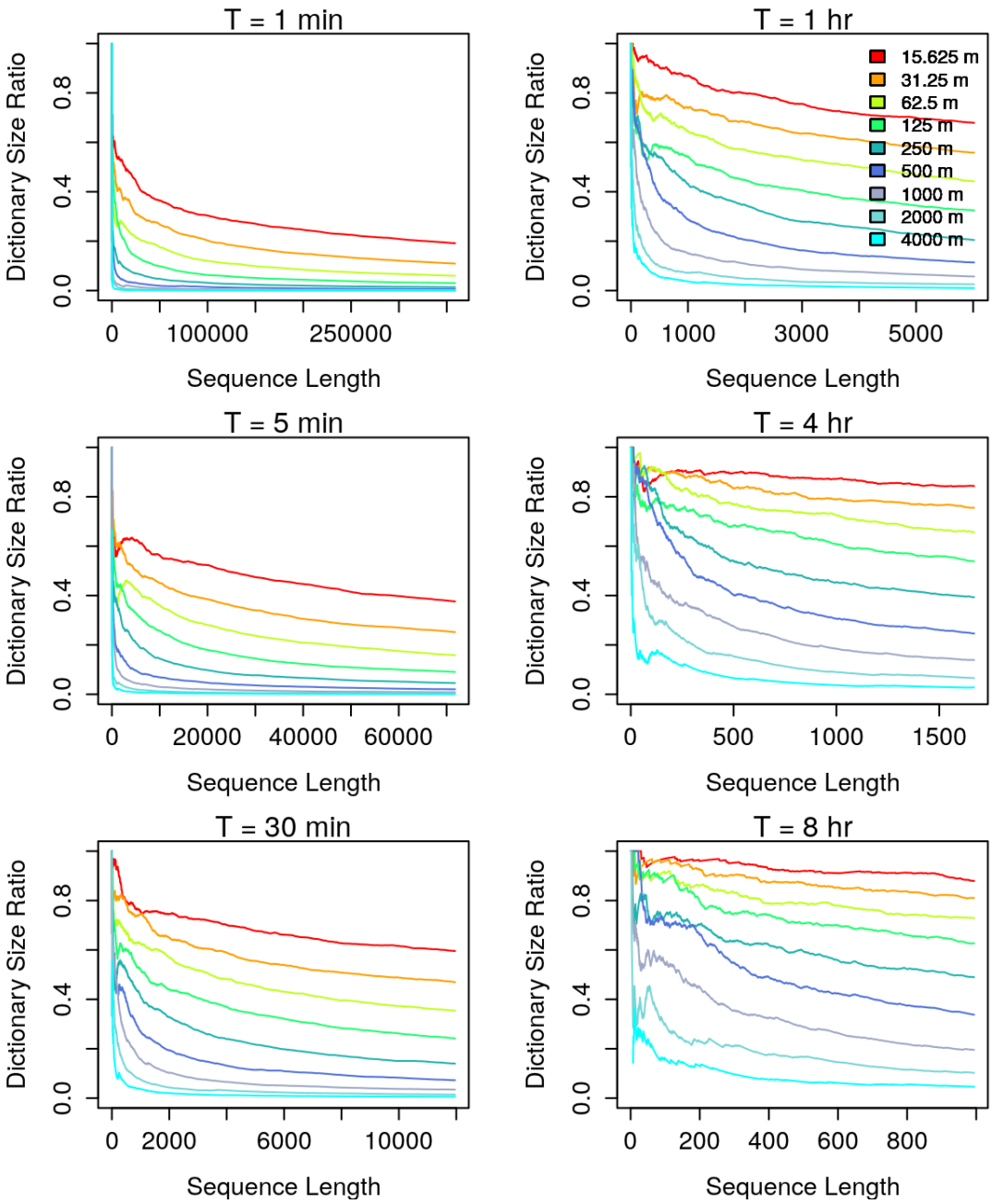


Fig 5.31: Growth of dictionary size ratio in Taxi Cab Dataset

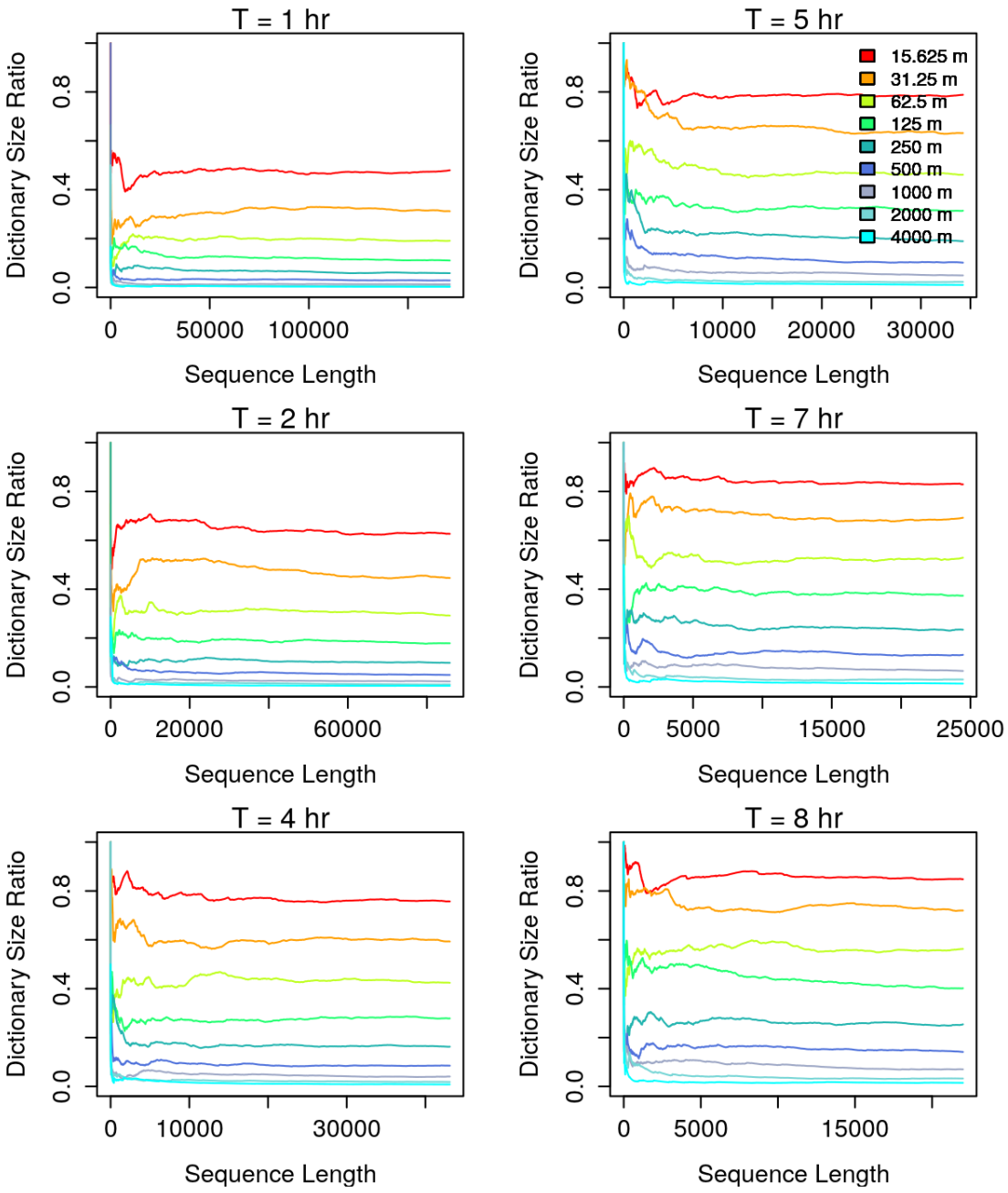


Fig 5.32: Growth of dictionary size ratio in Moose Dataset

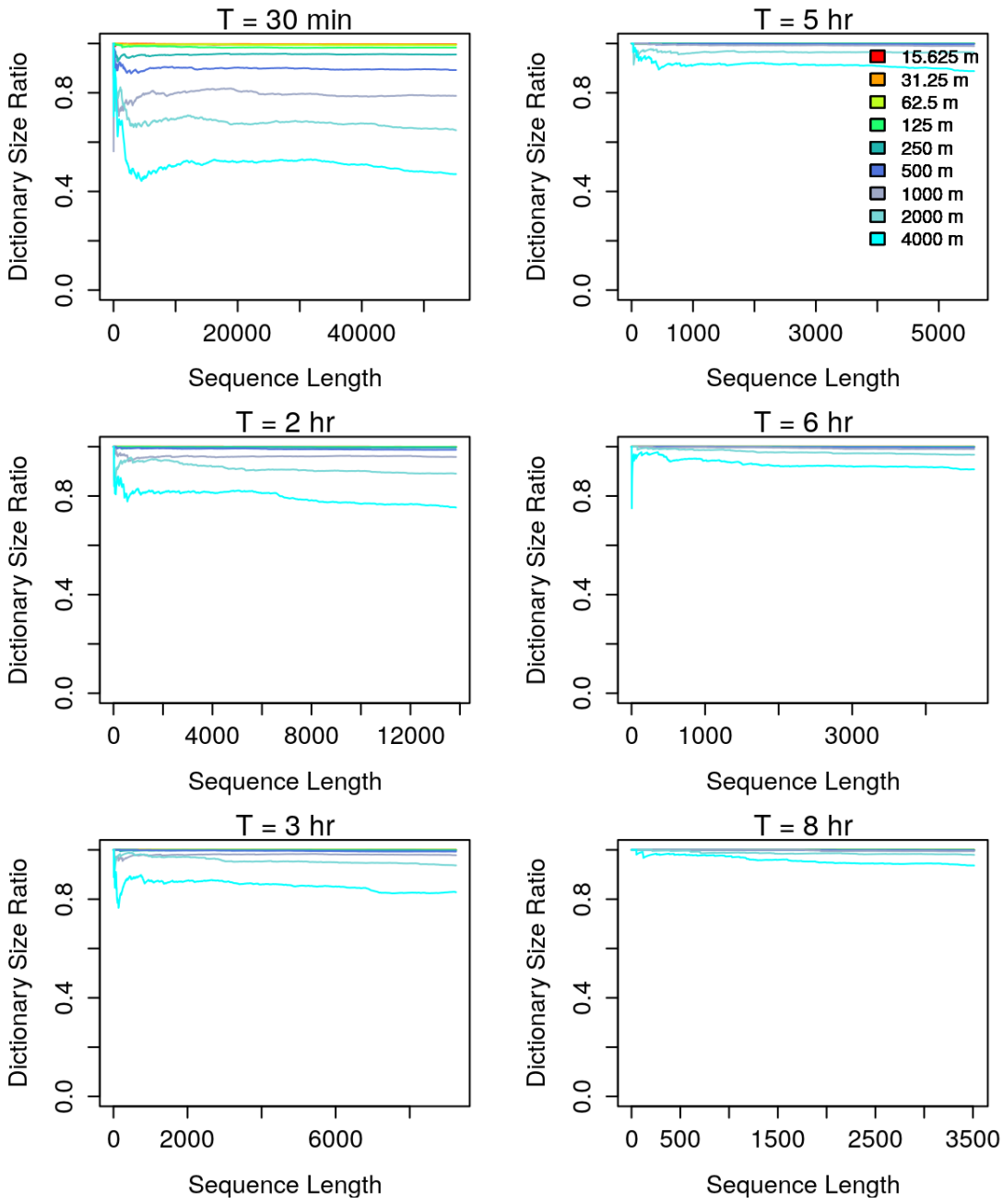


Fig 5.33: Growth of dictionary size ratio in Antarctic Petrel Dataset

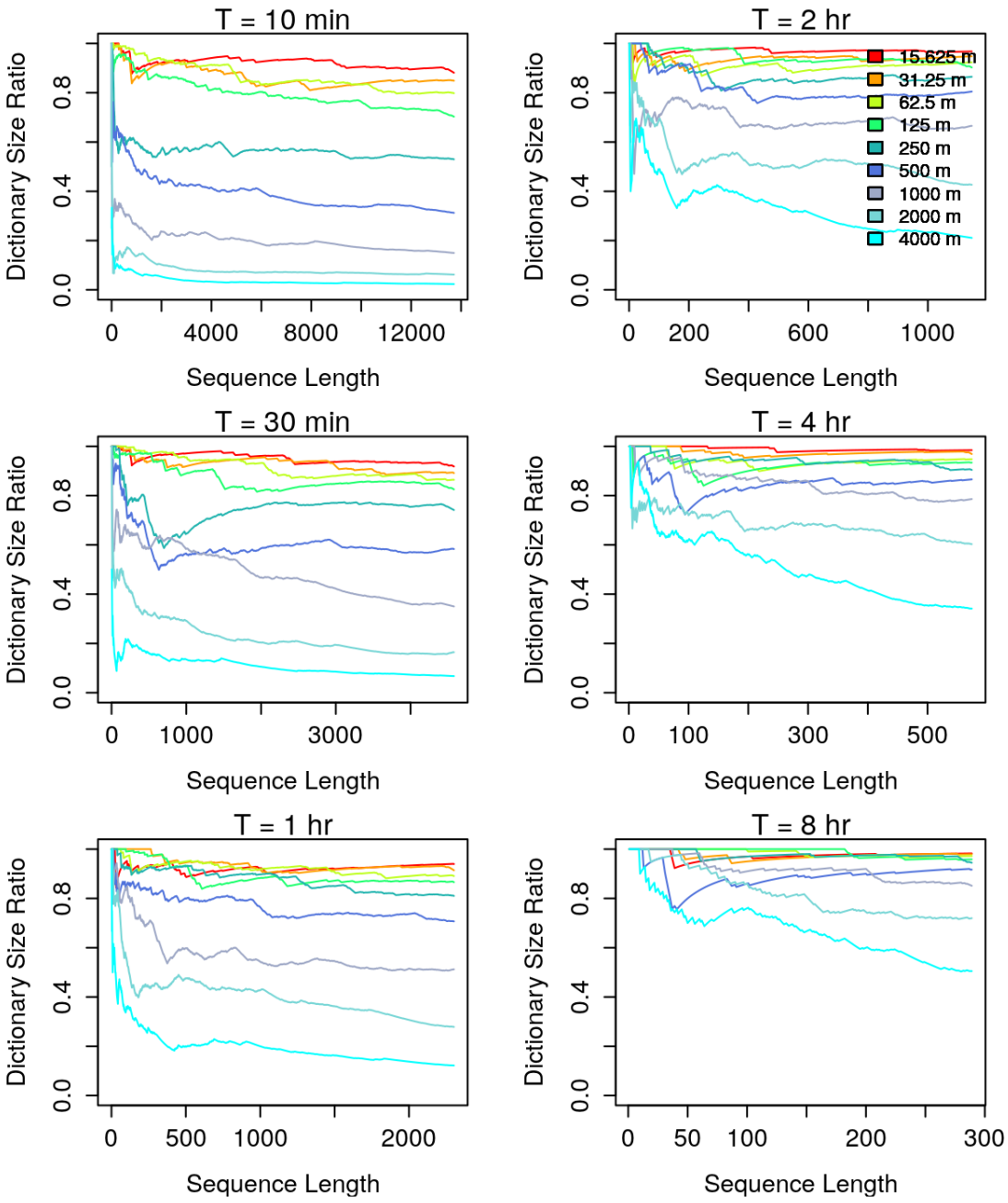


Fig 5.34: Growth of dictionary size ratio in Ocean Drifter Dataset

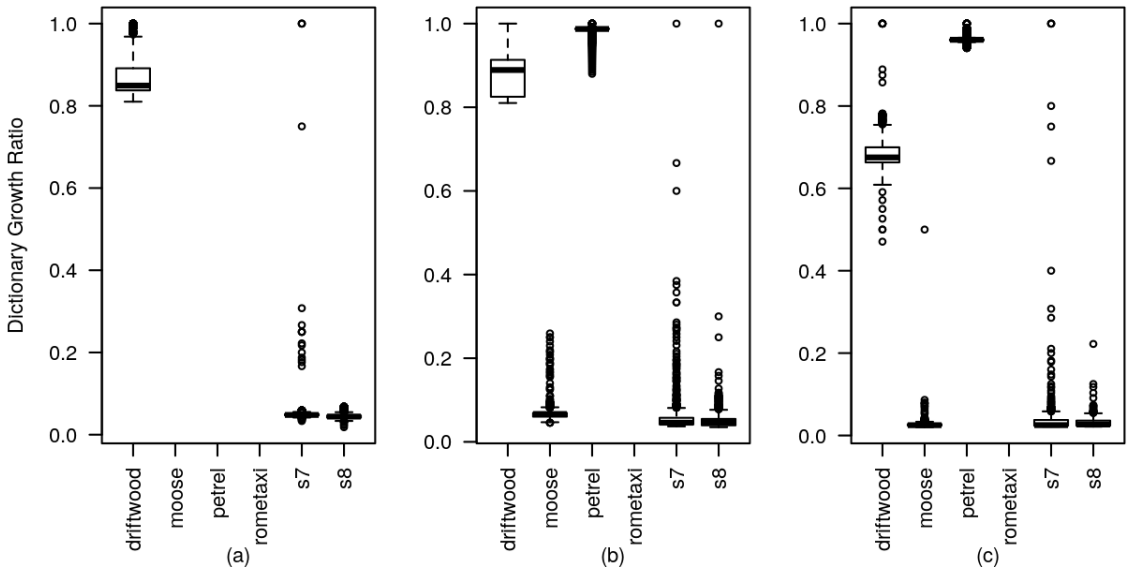


Fig 5.35: Comparison of the distributions of aggregate dictionary growth ratios across datasets: **(a)** $(T, d) = (10min, 62.5m)$ **(b)** $(T, d) = (60min, 250m)$ **(c)** $(T, d) = (4hrs, 1km)$

5.9 Addendum

The manuscript in this chapter has been reformatted, and some paragraphs/sentences of the submitted version have been modified/added/deleted, based on edits proposed by the examining committee, for inclusion in the dissertation. No substantial changes to the results/findings were made.

Chapter 6

Conclusions

6.1 Summary

The analysis of mobility is a fundamental step in many research areas. Measures such as the dispersion of people, predictability of daily movements, and the amount of time spent in the most popular locations is commonly used in aggregate form to gain insight into the behavior of a population. The aggregate features of interest depend on spatial and temporal scales of data collection, which is intrinsically limited to the amount of information available at a given resolution for a particular path.

The first contribution of this dissertation is the analysis of the scaling properties of mobility data features. Some features, as shown in Chapter 3, demonstrate repetitive behavior with changing spatial scales and sampling intervals. Moreover, the order of the metrics are generally preserved across datasets as spatio-temporal resolutions change, but different features show different degrees of sensitivity to changes to spatio-temporal resolutions. For example, in Chapter 3, results showed that RoG was not well described by power law distribution although it has been reported to follow a heavy tailed distribution. This anomaly may have resulted from the sampling setup of studies and/or participant behavior. In the literature, different distribution parameters are found for the same feature in different studies. As the distributions do not generalize across resolution, data from empirical studies may not be broadly generalizable in mathematical and computational models which employ these parameters.

When these aggregate features vary, the predictability of the moving agent varies as well. Therefore, entropy rate, which is a common metric of predictability, should vary with

spatial and temporal quantization. Entropy rate is also an important metric in mobility studies, which facilitates predictability analysis and applications that make decisions based on probable next location of movement (e.g., routing in a mobile ad-hoc or vehicular network). Entropy rate is calculated from the past history of human movement, as described in Section 1.6. The mathematical background required to understand entropy rate is provided in Chpater 1.

In the literature, researchers have worked on the bounds of the predictability of human movement, which is quantified using entropy rate [7, 21]. However, these results are dependent on the underlying spatio-temporal resolutions. Predictability also has a human behavioral element beyond the impact of the spatio-temporal scaling. This dissertation develops a model from first principles, presented in Manuscript 2, to express mobility entropy rate as a function of spatio-temporal quantization and mobility characteristics. Although constrained by simplifying assumptions for mathematical tractability, this model is important because it explains how spatio-temporal resolution and mobility parameters influence the entropy rate. Despite the assumptions, the model showed agreement with Lempel-Ziv-based compression of simulated paths, over a wide range of spatial and temporal sampling scales. The approximation broke down at small spatial scales, which can be explained by a violation of the assumptions, and at larger scales due to the loss of specificity at coarse spatial resolutions (kilometer level) or large sampling intervals (e.g., a few samples a day). Apart from providing the means for inter-mobility-study comparison, this model has some other notable features: it is generic in the number of dimensions of a place, subject to the assumptions; and it demonstrate that the maximal entropy rate for a certain spatial scale may be achieved at an optimal sampling interval, which signifies that samples should be taken neither too frequently nor at large intervals. Researchers may leverage this result in designing efficient data collection studies.

Different populations and environments show different sensitivity to resampling. The theoretical model also verifies that the scaling model of entropy rate is not purely a mathematical artifact; the entropy rate is affected by both the agent behavior and the sampled realization of that path. This finding motivates the potential to separate the two components of entropy rate scaling, which would allow for the isolation of the behavioral fingerprint

present in the data.

Chapter 5 relaxes the key assumptions of the previous model: notably that of a continuously moving agent. The new model considers dwell time distribution of the agents and incorporate that into the model. The model is not limited to humans, and is relevant to naturally complex paths and the movement of actors having some degree of agency. The model was successfully validated with mobility datasets of humans, moose, petrels, taxicabs, and ocean drifters. The analysis presented in the work yielded novel findings related to the behavioral similarity between dataset populations. The model can represent mobility entropy rate well within a constrained spatio-temporal scaling regime. At spatial scales on the order of $15m$, the model deviates because of unmodeled noise associated with GPS records impacts the structure of visit strings. At coarse spatial scales ($\geq$ 2 km) and large sampling intervals, the observable mean velocities and dwell times of an agent are not guaranteed to reflect the underlying path properties of the agent. Within the modeled range, the scaling law helps researchers perform meta-analysis on mobility studies collected at different spatio-temporal scales. With the help of this model, researchers can make actionable conclusions about the relative movement behaviors of two populations by rescaling samples or directly comparing scaling behaviors.

Different applications that depend on the results of mobility studies, e.g., containing the spread of contagious pathogens, urban planning, and network routing have some commonality in the mobility features of interest. In the literature, independent studies are known to have reported different parameters for the common mobility features, which are partly due to the study design, and partly due to the differences of the populations [5,93]. The mathematical model derived and validated in this dissertation enables researchers to port the results of one study to another, making cross comparisons regardless of the spatio-temporal resolution of the data. The model also explicitly identifies and segregates the mobility parameters from the results to analyze their impact in policy and decision making.

6.2 Conclusions

This dissertation provides an entropy rate scaling model, which separates dataset specific parameters from the spatio-temporal parameter, and allows researchers to extract scale-agnostic subjective parameters of a population from the results of a mobility study. The scaling law allows for inter-study comparisons. The evaluation of the model shows substantial agreement with the empirical results. To summarize, the contributions of this dissertation to the literature are as follow:

- Research communities across various disciplines are interested in mobility features such as dwell time and trip duration. Different populations exhibit different distributions of these features, asserting their behavioral differences. This dissertation shows that these features also depend on the scales of measurement.

- Since the predictability of a moving object is measured by observing its mobility traces, the measure of predictability is contingent on mobility features like speed and dwelling habits. This dissertation progressively develops a scaling model to relate entropy rate, which is a key to measuring predictability, to spatio-temporal scales and behavioral factors of the agent. At the primary stage, the model shows how entropy rate scales with the spatio-temporal resolution; and demonstrate that velocity is an integral part of predictability. Although the model at this stage was developed based on strong assumptions about the path and velocity of the agent, it acted as the foundation stone of the final scaling law, which was validated with a diverse set of empirical datasets.

- Objects with a degree of agency intermittently dwell at places instead of moving constantly. The relative amount of dwelling, in conjugation with moving speed, influences predictability; and is incorporated into the final model. The model is normalizable to spatio-temporal resolutions different from the study resolutions, making it a tool to compare results from two independent studies. Comparison of the model parameters for two datasets may reveal behavioral differences between the underlying populations.

The goodness of fit data show that the entropy rate scaling model demonstrates significant agreement with the empirical data within a constrained spatio-temporal range. Development

of location tracking technologies have made it possible to collect frequent location samples. Although the study on a large population is limited by cost and management challenges, data collected with commodity location-enabled devices provide far better precision and accuracy than mobile cell tower data. The entropy rate scaling law model developed in this dissertation deviates when T or d is large because discretizing movement feature distributions at coarse resolutions results in measurements that deviate from the true distributions. Re-binning data to a coarser resolution causes loss of information. However, if the base data are collected at a relatively coarser resolution (e.g., cell tower logs), fine-grained path properties are already unavailable; and projecting the data to a finer resolution will be far from providing reliable conclusions. The level of imprecision of the base data will have an impact on the conclusions made. The parameters of the model, described in Chapter 5 of this thesis, can be used for scale-free entropy rate comparison. This is desirable because it allows different apparatus to be used; but for the scaling analysis to be valid and to make generalizable conclusions, high fidelity traces are required because we have noted that low spatial and temporal fidelity regimes are often characterized by slowly varying entropy surfaces.

Although the dissertation was motivated from the analysis of human mobility entropy rate, the final model is generally applicable to the movements of objects exhibiting agency. The model was successfully validated with mobility datasets from a variety of sources, including humans, taxicabs, seabirds, moose, and ocean drifters. The scaling model exploits the interaction of spatio-temporal scales, dwelling habits, and movement speeds of the underlying agents to estimate entropy rates at spatial scales consistent with the underlying mobility. The model enables researchers to re-normalize results to a different resolution to analyze relative mobility behaviors of different populations. The model also gives insight into the characteristic spatial scales of the underlying population.

The final model of this dissertation was fitted against and validated with population-wide aggregated data. It would be worthwhile to assess the model with individual traces. As the interval between samples increase, the number of samples go down. To faithfully compute model parameters for an agent, the agent's movements should be observed for a long enough period. Data should be collected at a reasonably fine spatio-temporal resolution so that marginal path-properties are not lost.

The scaling model can provide entropy rate estimates at a spatio-temporal resolution different from the one at which the study was performed. The dissertation shows that distributions of mobility features are sensitive to spatio-temporal scales, and the degree of sensitivity varies from population to population. Further research may be directed at quantifying the sensitivity of mobility features to changes in scales. Further research works could incorporate the knowledge on sensitivity of mobility features to scaling, and the entropy rate scaling model into location prediction methodologies. A closely related direction for future research will be to instrument a synthetic mobility model exhibiting the feature distributions and sensitivity reported in this dissertation.

References

1. Bhadra S, Majumder SK. Entropy Optimization and Its Application to Regional & Urban Planning. Int J Pure Appl Sci Technol. 2013;14(2):50–60.

2. Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A. Multiscale Mobility Networks and the Spatial Spreading of Infectious Diseases. National Academy of Sciences. 2009;106(51):21484–21489.

3. Shin R, Hong S, Lee K, Chong S. On the Levy-walk Nature of Human Mobility: Do Humans Walk Like Monkeys? In: IEEE INFOCOM; 2008. p. 924–932.

4. Hashemian M, Stanley K, Osgood N. Leveraging H1N1 Infection Transmission Modeling with Proximity Sensor Microdata. BMC Medical Informatics and Decision Making. 2012;12(1):35:1–35:15.

5. Song C, Koren T, Wang P, Barabási AL. Modelling the Scaling Properties of Human Mobility. Nature Physics. 2010;6(10):818–823.

6. Kwan MP. Algorithmic Geographies: Big Data, Algorithmic Uncertainty, and the Production of Geographic Knowledge. Annals of the American Association of Geographers. 2016;106(2):274–282.

7. Song C, Qu Z, Blumm N, Barabási AL. Limits of Predictability in Human Mobility. Science. 2010;327(5968):1018–1021.

8. Rodriguez-Carrion A, Garcia-Rubio C, Campo C, Das SK. Analysis of a Fast LZ-Based Entropy Estimator for Mobility Data. In: IEEE International Conference on Pervasive Computing and Communication Workshops; 2015. p. 451–456.

9. Schürmann T, Grassberger P. Entropy Estimation of Symbol Sequences. Chaos: An Interdisciplinary Journal of Nonlinear Science. 1996;6(3):414–427.

10. Ziv J, Lempel A. Compression of Individual Sequences via Variable-Rate Coding. IEEE Transactions on Information Theory. 1978;24(5):530–536.

11. Wyner AD, Ziv J. Some Asymptotic Properties of the Entropy of a Stationary Ergodic Data Source with Applications to Data Compression. IEEE Transactions on Information Theory. 1989;35(6):1250–1258.

12. Scholkmann F. Power-Law Scaling of the Impact Crater Size-Frequency Distribution on Pluto: A Preliminary Analysis Based on First Images from New Horizons' Flyby. Progress in Physics. 2016;12(1):26–29.

13. Lee K, Hong S, Kim SJ, Rhee I, Chong S. SLAW: A New Mobility Model for Human Walks. In: IEEE INFOCOM; 2009. p. 855–863.

14. Newman ME. Power Laws, Pareto Distributions and Zipf's Law. Contemporary Physics. 2005;46(5):323–351.

15. Clauset A. Power Law Distributions; 2011. Last accessed: 2017-02-24. `http://tuvalu.santafe.edu/~aaronc/courses/7000/csci7000-001_2011_L2.pdf`.

16. Adamic LA. Zipf, Power-laws, and Pareto-a ranking tutorial; 2000. Last accessed: 2017-02-24. Xerox Palo Alto Research Center, Palo Alto, CA, `http://ginger.hpl.hp.com/shl/papers/ranking/ranking.html`.

17. Asmussen S. Applied Probability and Queues. vol. 51. Springer Science & Business Media; 2008.

18. Foss S, Korshunov D, Zachary S. An Introduction to Heavy-Tailed and Subexponential Distributions. vol. 6. Springer; 2011.

19. Haas M, Pigorsch C. Financial Economics, Fat-tailed Distributions. In: Complex Systems in Finance and Econometrics; 2009. p. 308–339.

20. Song C, Qu Z, Blumm N, Barabási AL. Supporting Online Material for Limits of Predictability in Human Mobility. Science. 2010;327(5968):1–20.

21. Smith G, Wieser R, Goulding J, Barrack D. A Refined Limit on the Predictability of Human Mobility. In: IEEE International Conference on Pervasive Computing and Communications (PerCom); 2014. p. 88–94.

22. Qian W, Stanley KG, Osgood ND. The Impact of Spatial Resolution and Representation on Human Mobility Predictability. In: Web and Wireless Geographical Information Systems; 2013. p. 25–40.

23. Cover TM, Thomas JA. Elements of Information Theory. John Wiley & Sons; 2012.

24. Nason G. Stationary and Non-stationary Time Series. Statistics in Volcanology Special Publications of IAVCEI. 2006;1:11:1–11:29.

25. Jenkins G, Priestley M. The Spectral Analysis of Time-series. Journal of the Royal Statistical Society Series B (Methodological). 1957;p. 1–12.

26. Feller W. An Introduction to Probability Theory and Its Applications. vol. 2. John Wiley & Sons; 2008.

27. Hardy GH. Divergent Series. vol. 334. American Mathematical Society; 2000.

28. Kontoyiannis I, Algoet PH, Suhov YM, Wyner AJ. Nonparametric Entropy Estimation for Stationary Processes and Random Fields, with Applications to English Text. IEEE Transactions on Information Theory. 1998;44(3):1319–1327.

29. Baumann P, Santini S. On the Use of Instantaneous Entropy to Measure the Momentary Predictability of Human Mobility. In: 14th IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC); 2013. p. 535–539.

30. Bhattacharya A, Das SK. LeZi-Update: An Information-Theoretic Approach to Track Mobile Users in PCS Networks. In: 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking; 1999. p. 1–12.

31. Asgari F, Gauthier V, Becker M. A Survey on Human Mobility and Its Applications. arXiv preprint arXiv:13070814. 2013;p. 1–18.

32. Yuan J, Zheng Y, Xie X. Discovering Regions of Different Functions in a City Using Human Mobility and POIs. In: 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2012. p. 186–194.

33. Giannotti F, Pappalardo L, Pedreschi D, Wang D. A Complexity Science Perspective on Human Mobility; 2012.

34. Hashemian M, Qian W, Stanley KG, Osgood ND. Temporal Aggregation Impacts on Epidemiological Simulations Employing Microcontact Data. BMC Medical Informatics and Decision Making. 2012;12(1):132:1–132:15.

35. Hashemian MS, Stanley KG, Knowles DL, Calver J, Osgood ND. Human Network Data Collection in the Wild: the Epidemiological Utility of Micro-contact and Location Data. In: 2nd ACM SIGHIT International Health Informatics Symposium; 2012. p. 255–264.

36. Hashemian M, Knowles D, Calver J, Qian W, Bullock MC, Bell S, Mandryk RL, Osgood N, Stanley KG. IEpi: an End to End Solution for Collecting, Conditioning and Utilizing Epidemiologically Relevant Data. In: 2nd ACM international workshop on Pervasive Wireless Healthcare; 2012. p. 3–8.

37. Jensen BS, Larsen JE, Jensen K, Larsen J, Hansen LK. Estimating Human Predictability from Mobile Sensor Data. In: IEEE International Workshop on Machine Learning for Signal Processing; 2010. p. 196–201.

38. Pu J, Xu P, Qu H, Cui W, Liu S, Ni L. Visual Analysis of People's Mobility Pattern from Mobile Phone Data. In: Visual Information Communication-International Symposium; 2011. p. 13:1–13:10.

39. Zhao M, Mason L, Wang W. Empirical Study on Human Mobility for Mobile Wireless Networks. In: IEEE Military Communications Conference; 2008. p. 1–7.

40. Csáji BC, Browet A, Traag VA, Delvenne JC, Huens E, Van Dooren P, Smoreda Z, Blondel VD. Exploring the Mobility of Mobile Phone Users. Physica A: Statistical Mechanics and its Applications. 2013;392(6):1459–1473.

41. Lenhart A, Ling R, Campbell S, Purcell K. Teens and Mobile Phones: Text Messaging Explodes as Teens Embrace It as the Centerpiece of Their Communication Strategies with Friends. Pew Internet & American Life Project. 2010;p. 1–94.

42. Karamshuk D, Boldrini C, Conti M, Passarella A. Human Mobility Models for Opportunistic Networks. IEEE Communications Magazine. 2011;49(12):157–165.

43. Kos T, Grgic M, Sisul G. Mobile User Positioning in GSM/UMTS Cellular Networks. In: 48th International Symposium on Multimedia Signal Processing and Communications; 2006. p. 185–188.

44. Wu X, Mazurowski M, Chen Z, Meratnia N. Emergency Message Dissemination System for Smartphones During Natural Disasters. In: 11th International Conference on ITS Telecommunications (ITST); 2011. p. 258–263.

45. Becker RA, Cáceres R, Hanson K, Loh JM, Urbanek S, Varshavsky A, Volinsky C. A Tale of One City: Using Cellular Network Data for Urban Planning. IEEE Pervasive Computing. 2011;10(4):18–26.

46. Modsching M, Kramer R, ten Hagen K. Field Trial on GPS Accuracy in a Medium Size City: The Influence of Built-up. In: 3rd Workshop on Positioning, Navigation and Communication; 2006. p. 209–218.

47. Leitl E. Information Technology Issues During and After Katrina and Usefulness of the Internet: How We Mobilized and Utilized Digital Communications Systems. Critical Care. 2005;10(1):110.

48. Lien YN, Jang HC, Tsai TC. P2Pnet: A MANET Based Emergency Communication System for Catastrophic Natural Disasters. Communications. 2010;p. 1–23.

49. Aschenbruck N, Frank M, Martini P, Tolle J. Human Mobility in Manet Disaster Area Simulation - A Realistic Approach. In: 29th Annual IEEE International Conference on Local Computer Networks; 2004. p. 668–675.

50. Foroozani A, Gharib M, Hemmatyar AMA, Movaghar A. A Novel Human Mobility Model for MANETs Based on Real Data. In: 23rd International Conference on Computer Communication and Networks (ICCCN); 2014. p. 1–7.

51. Zhang D, Sterbenz JP. Robustness Analysis of Mobile Ad Hoc Networks Using Human Mobility Traces. In: 11th International Conference on the Design of Reliable Communication Networks (DRCN); 2015. p. 125–132.

52. Medrano-Chávez AG, Pérez-Cortés E, Lopez-Guerrero M. Studying the Effect of Human Mobility on MANET Topology and Routing: Friend Or Foe? In: 13th ACM International Symposium on Mobility Management and Wireless Access; 2015. p. 39–46.

53. Abboud K, Zhuang W. Modeling and Analysis for Emergency Messaging Delay in Vehicular Ad Hoc Networks. In: Global Telecommunications Conference; 2009. p. 1–6.

54. Ota K, Dong M, Zhu H, Chang S, Shen X. Traffic Information Prediction in Urban Vehicular Networks: A Correlation Based Approach. In: 2011 IEEE Wireless Communications and Networking Conference; 2011. p. 1021–1025.

55. Karimi R, Ithnin N, Razak SA, Najafzadeh S. DTN Routing Protocols for VANETs: Issues and Approaches. IJCSI International Journal of Computer Science Issues. 2011;8(6):89–93.

56. Harri J, Filali F, Bonnet C. Mobility Models for Vehicular Ad Hoc Networks: a Survey and Taxonomy. IEEE Communications Surveys & Tutorials. 2009;11(4):19–41.

57. Hou X, Li Y, Jin D, Wu DO, Chen S. Modeling the Impact of Mobility on the Connectivity of Vehicular Networks in Large-Scale Urban Environments. IEEE Transactions on Vehicular Technology. 2016;65(4):2753–2758.

58. Vahdat A, Becker D. Epidemic Routing for Partially Connected Ad Hoc Networks. Technical Report CS-200006, Duke University; 2000.

59. Spyropoulos T, Psounis K, Raghavendra CS. Spray and Wait: an Efficient Routing Scheme for Intermittently Connected Mobile Networks. In: ACM SIGCOMM Workshop on Delay-tolerant Networking; 2005. p. 252–259.

60. Dubois-Ferriere H, Grossglauser M, Vetterli M. Age Matters: Efficient Route Discovery in Mobile Ad Hoc Networks Using Encounter Ages. In: 4th ACM International Symposium on Mobile Ad Hoc Networking & Computing; 2003. p. 257–266.

61. Balasubramanian A, Levine B, Venkataramani A. DTN Routing As a Resource Allocation Problem. In: ACM SIGCOMM Computer Communication Review. vol. 37; 2007. p. 373–384.

62. Zhu Y, Xu B, Shi X, Wang Y. A Survey of Social-based Routing in Delay Tolerant Networks: Positive and Negative Social Effects. Communications Surveys & Tutorials, IEEE. 2013;15(1):387–401.

63. Davis JA, Fagg AH, Levine BN. Wearable Computers As Packet Transport Mechanisms in Highly-Partitioned Ad-Hoc Networks. In: 5th International Symposium on Wearable Computers; 2001. p. 141–148.

64. Juang P, Oki H, Wang Y, Martonosi M, Peh LS, Rubenstein D. Energy-Efficient Computing for Wildlife Tracking: Design Tradeoffs and Early Experiences with ZebraNet. In: ACM Sigplan Notices. vol. 37; 2002. p. 96–107.

65. Lindgren A, Doria A, Schelén O. Probabilistic Routing in Intermittently Connected Networks. ACM SIGMOBILE Mobile Computing and Communications Review. 2003;7(3):19–20.

66. Chatzigiannakis I, Nikoletseas S, Paspallis N, Spirakis P, Zaroliagis C. An Experimental Study of Basic Communication Protocols in Ad-hoc Mobile Networks. In: Algorithm Engineering; 2001. p. 159–171.

67. Sugihara R, Gupta RK. Data Mule Scheduling in Sensor Networks: Scheduling under Location and Time Constraints. UCSD Technical Report. 2007;p. 1–63.

68. Somasundara AA, Ramamoorthy A, Srivastava MB. Mobile Element Scheduling for Efficient Data Collection in Wireless Sensor Networks with Dynamic Deadlines. In: 25th IEEE International Real-Time Systems Symposium; 2004. p. 296–305.

69. Bin Tariq MM, Ammar M, Zegura E. Message Ferry Route Design for Sparse Ad Hoc Networks with Mobile Nodes. In: 7th ACM International Symposium on Mobile Ad Hoc Networking and Computing; 2006. p. 37–48.

70. Liu C, Wu J. An Optimal Probabilistic Forwarding Protocolin Delay Tolerant Networks. In: 10th ACM international symposium on Mobile ad hoc networking and computing; 2009. p. 105–114.

71. Yuan Q, Cardei I, Wu J. Predict and Relay: an Efficient Routing in Disruption-Tolerant Networks. In: 10th ACM International Symposium on Mobile Ad Hoc Networking and Computing; 2009. p. 95–104.

72. Nelson SC, Bakht M, Kravets R, Harris III AF. Encounter-Based Routing in DTNs. ACM SIGMOBILE Mobile Computing and Communications Review. 2009;13(1):56–59.

73. Rasul K, Makaroff D, Stanley KG. Hybrid Community-Based Forwarding: A Complete Energy Efficient Algorithm for Pocket Switched Networks. In: IEEE 40th Local Computer Networks Conference Workshops (LCN Workshops); 2015. p. 760–768.

74. Rasul K, Chowdhury SA, Makaroff D, Stanley K. Community-based Forwarding for Low-capacity Pocket Switched Networks. In: 17th ACM International conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems; 2014. p. 249–257.

75. Zheng Y, Xie X, Ma WY. GeoLife: A Collaborative Social Networking Service Among User, Location and Trajectory. IEEE Data Engineering Bulletin. 2010;33(2):32–39.

76. Ma X, Wu YJ, Wang Y, Chen F, Liu J. Mining Smart Card Data for Transit Riders' Travel Patterns. Transportation Research Part C: Emerging Technologies. 2013;36:1–12.

77. Pelletier MP, Trépanier M, Morency C. Smart Card Data Use in Public Transit: A Literature Review. Transportation Research Part C: Emerging Technologies. 2011;19(4):557–568.

78. Esliger D, Sherar L, Muhajarine N. Smart Cities, Healthy Kids: The Association Between Neighbourhood Design and Children's Physical Activity and Time Spent Sedentary. Can J Public Health. 2012;103(Suppl 3):S22–S28.

79. Eagle N, Pentland A. Reality Mining: Sensing Complex Social Systems. Personal and ubiquitous computing. 2006;10(4):255–268.

80. Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C. Geo-located Twitter As Proxy for Global Mobility Patterns. Cartography and Geographic Information Science. 2014;41(3):260–271.

81. Cho E, Myers SA, Leskovec J. Friendship and Mobility: User Movement in Location-based Social Networks. In: 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2011. p. 1082–1090.

82. Sui D, Goodchild M. The Convergence of GIS and Social Media: Challenges for GIScience. International Journal of Geographical Information Science. 2011;25(11):1737–1748.

83. Simini F, González MC, Maritan A, Barabási AL. A Universal Model for Mobility and Migration Patterns. Nature. 2012;484(7392):96–100.

84. González MC, Hidalgo CA, Barabási AL. Understanding Individual Human Mobility Patterns. Nature. 2009;458(7235):238–238.

85. Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C. A Tale of Many Cities: Universal Patterns in Human Urban Mobility. PLoS ONE. 2012;7(5):e37027.

86. Garske T, Yu H, Peng Z, Ye M, Zhou H, Cheng X, Wu J, Ferguson N. Travel Patterns in China. PLoS ONE. 2011;6(2):e16364.

87. Isaacman S, Becker R, Cáceres R, Martonosi M, Rowland J, Varshavsky A, Willinger W. Human Mobility Modeling at Metropolitan Scales. In: 10th international conference on Mobile systems, applications, and services; 2012. p. 239–252.

88. Thiagarajan A, Ravindranath L, LaCurts K, Madden S, Balakrishnan H, Toledo S, Eriksson J. VTrack: Accurate, Energy-aware Road Traffic Delay Estimation Using Mobile Phones. In: 7th ACM Conference on Embedded Networked Sensor Systems; 2009. p. 85–98.

89. Thiagarajan A, Ravindranath LS, Balakrishnan H, Madden S, Girod L. Accurate, Low-Energy Trajectory Mapping for Mobile Devices. In: 8th USENIX Symposium on Networked Systems Design and Implementation; 2011. p. 1–14.

90. Hunter T, Moldovan T, Zaharia M, Merzgui S, Ma J, Franklin MJ, Abbeel P, Bayen AM. Scaling the Mobile Millennium System in the Cloud. In: 2nd ACM Symposium on Cloud Computing; 2011. p. 28:1–28:8.

91. Fayazbakhsh SK. Modeling Human Mobility and Its Applications in Routing in Delay-Tolerant Networks: a Short Survey. arXiv preprint arXiv:13071926. 2013;p. 1–5.

92. Rhee I, Shin M, Hong S, Lee K, Kim SJ, Chong S. On the Levy-walk Nature of Human Mobility. IEEE/ACM Transactions on Networking (TON). 2011;19(3):630–643.

93. Brockmann D, Hufnagel L, Geisel T. The Scaling Laws of Human Travel. Nature. 2006;439(7075):462–465.

94. Song C, Koren T, Wang P, Barabási AL. Modelling the Scaling Properties of Human Mobility Supplementary Material. Nature Physics. 2010;6(10):1–20.

95. Kim M, Kotz D, Kim S. Extracting a Mobility Model from Real User Traces. In: 25th IEEE International Conference on Computer Communications. vol. 6; 2006. p. 1–13.

96. Karagiannis T, Le Boudec JY, Vojnović M. Power Law and Exponential Decay of Intercontact Times Between Mobile Devices. IEEE Transactions on Mobile Computing. 2010;9(10):1377–1390.

97. Chaintreau A, Hui P, Crowcroft J, Diot C, Gass R, Scott J. Impact of Human Mobility on Opportunistic Forwarding Algorithms. IEEE Transactions on Mobile Computing. 2007;6(6):606–620.

98. Hu T, Wenning BL, Görg C, Toseef U, Guo Z. Statistical Analysis of Contact Patterns Between Human-carried Mobile Devices. In: Mobile Networks and Management; 2012. p. 244–257.

99. Lee K, Hong S, Kim SJ, Rhee I, Chong S. Demystifying Levy Walk Patterns in Human Walks. North Carolina State University, Tech Rep. 2008;p. 1–14.

100. Gaite J. Zipf's Law for Fractal Voids and a New Void-finder. The European Physical Journal B-Condensed Matter and Complex Systems. 2005;47(1):93–98.

101. Mandelbrot BB. The Fractal Geometry of Nature, revised and enlarged edition (of the 1977 edition); 1983.

102. Sharma G, Mazumdar RR. Scaling Laws for Capacity and Delay in Wireless Ad Hoc Networks with Random Mobility. In: IEEE International Conference on Communications. vol. 7; 2004. p. 3869–3873.

103. Hong S, Rhee I, Kim SJ, Lee K, Chong S. Routing Performance Analysis of Human-driven Delay Tolerant Networks Using the Truncated Levy Walk Model. In: 1st ACM SIGMOBILE Workshop on Mobility Models; 2008. p. 25–32.

104. Giannotti F, Nanni M, Pedreschi D, Pinelli F, Renso C, Rinzivillo S, Trasarti R. Unveiling the Complexity of Human Mobility by Querying and Mining Massive Trajectory Data. The VLDB Journal-The International Journal on Very Large Data Bases. 2011;20(5):695–719.

105. Giannotti F, Nanni M, Pinelli F, Pedreschi D. Trajectory Pattern Mining. In: 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2007. p. 330–339.

106. Camp T, Boleng J, Davies V. A Survey of Mobility Models for Ad Hoc Network Research. Wireless Communications and Mobile Computing. 2002;2(5):483–502.

107. Shlesinger M, West B, Klafter J. Lévy Dynamics of Enhanced Diffusion: Application to Turbulence. Physical Review Letters. 1987;58(11):1100.

108. Shlesinger MF, Klafter J, Wong Y. Random Walks with Infinite Spatial and Temporal Moments. Journal of Statistical Physics. 1982;27(3):499–512.

109. Vazquez A, Sotolongo-Costa O, Brouers F. Diffusion Regimes in Levy Flights with Trapping. Physica A: Statistical Mechanics and its Applications. 1999;264(3):424–431.

110. Maruyama Y, Murakami J. Truncated Levy Walk of a Nanocluster Bound Weakly to an Atomically Flat Surface: Crossover from Superdiffusion to Normal Diffusion. Physical Review B. 2003;67(8):085406.

111. Montroll EW, Weiss GH. Random Walks on Lattices. II. Journal of Mathematical Physics. 1965;6(2):167–181.

112. Weiss GH. Aspects and Applications of the Random Walk (Random Materials & Processes); 2005.

113. Shlesinger M, Montroll E. On the Wonderful World of Random Walks. Nonequilibrium phenomena II- From stochastics to hydrodynamics(A 85-43951 21-77) Amsterdam and New York, North-Holland Physics Publishing, 1984,. 1984;p. 1–121.

114. Gillis JE, Weiss GH. Expected Number of Distinct Sites Visited by a Random Walk with an Infinite Variance. Journal of Mathematical Physics. 1970;11(4):1307–1312.

115. Yan XY, Han XP, Wang BH, Zhou T. Diversity of Individual Mobility Patterns and Emergence of Aggregated Scaling Laws. Scientific Reports. 2013;3(2678):1–5.

116. Larralde H, Trunfio P, Havlin S, Stanley HE, Weiss GH. Number of Distinct Sites Visited by N Random Walkers. Physical Review A. 1992;45(10):7128–7139.

117. Larralde H, Trunfio P, Havlin S, Stanley HE, Weiss GH. Territory Covered by N Diffusing Particles. Nature. 1992;355(6359):423–426.

118. Hossmann T, Spyropoulos T, Legendre F. A Complex Network Analysis of Human Mobility. In: IEEE conference on Computer communications workshops; 2011. p. 876–881.

119. Coscia M, Rinzivillo S, Giannotti F, Pedreschi D. Optimal Spatial Resolution for the Analysis of Human Mobility. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM); 2012. p. 248–252.

120. Jo HH, Karsai M, Karikoski J, Kaski K. Spatiotemporal Correlations of Handset-based Service Usages. EPJ Data Science. 2012;1(1):1–18.

121. Pan RK, Saramäki J. Path Lengths, Correlations, and Centrality in Temporal Networks. Physical Review E. 2011;84(1):1–11.

122. Barthélemy M. Spatial Networks. Physics Reports. 2011;499(1):1–101.

123. McInerney J, Stein S, Rogers A, Jennings NR. Exploring Periods of Low Predictability in Daily Life Mobility. Nokia Mobile Data Challenge 2012 Workshop. 2012;p. 1–6.

124. Anagnostopoulos T, Anagnostopoulos C, Hadjiefthymiades S. Efficient Location Prediction in Mobile Cellular Networks. International Journal of Wireless Information Networks. 2012;19(2):97–111.

125. Jeong J, Leconte M, Proutiere A. Human Mobility Prediction Using Non-Parametric Bayesian Model. arXiv preprint arXiv:150703292. 2015;p. 1–12.

126. Bohnert F, Zukerman I. Personalised Pathway Prediction. In: International Conference on User Modeling, Adaptation, and Personalization; 2010. p. 363–368.

127. Chen M, Yu X, Liu Y. Mining Moving Patterns for Predicting Next Location. Information Systems. 2015;54:156–168.

128. Alvarez-Garcia JA, Ortega JA, Gonzalez-Abril L, Velasco F. Trip Destination Prediction Based on Past GPS Log Using a Hidden Markov Model. Expert Systems with Applications. 2010;37(12):8166–8171.

129. Cho SB. Exploiting Machine Learning Techniques for Location Recognition and Prediction with Smartphone Logs. Neurocomputing. 2016;176:98–106.

130. Kim YJ, Cho SB. A HMM-based Location Prediction Framework with Location Recognizer Combining K-nearest Neighbor and Multiple Decision Trees. In: International Conference on Hybrid Artificial Intelligence Systems; 2013. p. 618–628.

131. Mathew W, Raposo R, Martins B. Predicting Future Locations with Hidden Markov Models. In: 2012 ACM Conference on Ubiquitous Computing; 2012. p. 911–918.

132. Ying JJC, Lee WC, Tseng VS. Mining Geographic-Temporal-Semantic Patterns in Trajectories for Location Prediction. ACM Transactions on Intelligent Systems and Technology (TIST). 2013;5(1):1–34.

133. Eagle N, Pentland AS. Eigenbehaviors: Identifying Structure in Routine. Behavioral Ecology and Sociobiology. 2009;63(7):1057–1066.

134. Gong H, Chen C, Bialostozky E, Lawson CT. A GPS/GIS Method for Travel Mode Detection in New York City. Computers, Environment and Urban Systems. 2012;36(2):131–139.

135. Knowles DL, Stanley KG, Osgood ND. A Field-validated Architecture for the Collection of Health-relevant Behavioural Data. In: 2014 IEEE International Conference on Healthcare Informatics (ICHI); 2014. p. 79–88.

136. Versichele M, Neutens T, Claeys Bouuaert M, Van de Weghe N. Time-geographic Derivation of Feasible Co-presence Opportunities from Network-constrained Episodic Movement Data. Transactions in GIS. 2014;18(5):687–703.

137. Openshaw S. The Modifiable Areal Unit Problem. Concepts and Techniques in Modern Geography No38. 1983;p. 1–22.

138. Openshaw S, Taylor PJ. A Million Or So Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem. Statistical applications in the spatial sciences. 1979;21:127–144.

139. Goodchild MF. GIScience, Geography, Form, and Process. Annals of the Association of American Geographers. 2004;94(4):709–714.

140. Bell S, Wilson K, Bissonnette L, Shah T. Access to Primary Health Care: Does Neighborhood of Residence Matter? Annals of the Association of American Geographers. 2013;103(1):85–105.

141. Stanley K, Bell S, Kreuger LK, Bhowmik P, Shojaati N, Elliott A, Osgood ND. Opportunistic Natural Experiments Using Digital Telemetry: a Transit Disruption Case Study. International Journal of Geographical Information Science. 2016;p. 1–20.

142. Barrat A, Fernandez B, Lin KK, Young LS. Modeling Temporal Networks Using Random Itineraries. Physical Review Letters. 2013;110(15):1–5.

143. Miller HJ, Goodchild MF. Data-driven Geography. GeoJournal. 2015;80(4):449–461.

144. Muhajarine N, Katapally TR, Fuller D, Stanley KG, Rainham D. Longitudinal active living research to address physical inactivity and sedentary behaviour in children in transition from preadolescence to adolescence. BMC Public Health. 2015;15(1):1–9.

145. Arribas-Bel D. Accidental, Open and Everywhere: Emerging Data Sources for the Understanding of Cities. Applied Geography. 2014;49:45–53.

146. Dark SJ, Bram D. The Modifiable Areal Unit Problem (MAUP) in Physical Geography. Progress in Physical Geography. 2007;31(5):471–479.

147. Gabriel AK, Goldstein RM, Zebker HA. Mapping Small Elevation Changes Over Large Areas: Differential Radar Interferometry. Journal of Geophysical Research: Solid Earth. 1989;94(B7):9183–9191.

148. Isaacson M, Shoval N. Application of Tracking Technologies to the Study of Pedestrian Spatial Behavior. The Professional Geographer. 2006;58(2):172–183.

149. Brown BB, Werner CM, Tribby CP, Miller HJ, Smith KR. Transit use, physical activity, and body mass index changes: objective measures associated with complete street light-rail construction. American journal of public health. 2015;105(7):1468–1474.

150. Draghici A, Agiali T, Chilipirea C. Visualization System for Human Mobility Analysis. In: RoEduNet International Conference-Networking in Education and Research (RoEduNet NER); 2015. p. 152–157.

151. Paul T, Stanley K, Osgood N, Bell S, Muhajarine N. Scaling Behavior of Human Mobility Distributions. In: International Conference on Geographic Information Science; 2016. p. 145–159.

152. Mathew W, Raposo R, Martins B. Predicting Future Locations with Hidden Markov Models. In: ACM Conference on Ubiquitous Computing; 2012. p. 911–918.

153. Chon Y, Shin H, Talipov E, Cha H. Evaluating Mobility Models for Temporal Prediction with High-granularity Mobility Data. In: IEEE International Conference on Pervasive Computing and Communications (PerCom); 2012. p. 206–212.

154. Xu KS. Predictability of Social Interactions. arXiv preprint arXiv:13061271. 2013;p. 1–6.

155. Roy BN. Fundamentals of Classical and Statistical Thermodynamics. John Wiley & Sons; 2002.

156. Shannon CE. A Mathematical Theory of Communication. The Bell System Technical Journal. 1948;27(3):379–423.

157. Lin M, Hsu WJ, Lee ZQ. Predictability of Individuals' Mobility with High-resolution Positioning Data. In: 2012 ACM Conference on Ubiquitous Computing; 2012. p. 381–390.

158. Osgood ND, Paul T, Stanley KG, Qian W. A Theoretical Basis for Entropy-Scaling Effects in Human Mobility Patterns. PLoS ONE. 2016;11(8):1–21.

159. Shimazaki H, Shinomoto S. A Method for Selecting the Bin Size of a Time Histogram. Neural computation. 2007;19(6):1503–1527.

160. Bracciale L, Bonola M, Loreti P, Bianchi G, Amici R, Rabuffi A. CRAWDAD Dataset Roma/taxi (v. 2014-07-17); 2014. Downloaded from `http://crawdad.org/roma/taxi/20140717`.

161. Laforge MP, Michel NL, Wheeler AL, Brook RK. Habitat Selection by Female Moose in the Canadian Prairie Ecozone. The Journal of Wildlife Management. 2016;80(6):1059–1068.

162. Tarroux A, Weimerskirch H, Wang SH, Bromwich DH, Cherel Y, Kato A, Ropert-Coudert Y, Varpe Ø, Yoccoz NG, Descamps S. Flexible Flight Response to Challenging Wind Conditions in a Commuting Antarctic Seabird: Do You Catch the Drift? Animal Behaviour. 2016;113:99–112.

163. Rosenberger A. Salish Sea Drift Card Study Preliminary Results; 2013. Accessed: 2017-02-23. `http://www.salishseaspillmap.org/files/SalishSeaDriftCardStudyPreliminaryResults.pdf`.

164. Aryadoust V. Application of Evolutionary Algorithm-based Symbolic Regression to Language Assessment: Toward Nonlinear Modeling. Psychological Test and Assessment Modeling. 2015;57(3):301–337.

165. Dubčáková R. Eureqa: Software Review. Genetic programming and evolvable machines. 2011;12(2):173–178.

166. Amici R, Bonola M, Bracciale L, Rabuffi A, Loreti P, Bianchi G. Performance Assessment of an Epidemic Protocol in VANET Using Real Traces. Procedia Computer Science. 2014;40:92–99.

167. Chopde NR, Nichat M. Landmark Based Shortest Path Detection by Using A* and Haversine Formula. International Journal of Innovative Research in Computer and Communication Engineering. 2013;1(2):298–302.