Brownian Bridge Interpolation for Human Mobility?

John Krumm
jckrumm@microsoft.com
Microsoft Research
Redmond, Washington, USA

ABSTRACT
The Brownian bridge is a method for probabilistically interpolating the location of a moving person, animal, or object between two measured points. This type of probabilistic interpolation is useful, because it represents the uncertainty of the interpolated points. It can be used to infer the probability of having visited a certain location, including possible exposure to disease. In the class of probabilistic interpolators, the Brownian bridge is attractive, because it has only a single adjustable parameter, the diffusion coefficient. This paper investigates the suitability of the Brownian bridge for interpolating human locations using mobility data from over 12 million people. One section looks at the consistency of the diffusion coefficient from person to person. As part of this, the paper presents, for the first time, a closed form solution for the maximum likelihood estimate of this parameter. The paper also presents statistical tests aimed at evaluating the accuracy of the Brownian bridge for interpolating human location.

CCS CONCEPTS
• Human-centered computing → Ubiquitous and mobile computing; • Information systems → Geographic information systems; Global positioning systems; Location based services.

KEYWORDS
location interpolation, brownian bridge, probabilistic location, human mobility

ACM Reference Format:

1 INTRODUCTION
Everyday location trajectories from people, animals, and vehicles are necessarily sampled discretely in time, sometimes with long, sporadic time intervals. This may be due to circumstances under our control, such as preserving the battery of the location sensor or the opportunistic sampling strategy of the sensing device. Other factors are beyond our control, such as where location sensing signals are too weak or blocked by buildings and terrain.

When we have two successive 2D location measurements \( x_a \) and \( x_b \), from a person, it is often important to infer something about the person’s location between the two measurements. We may want to know if the person was exposed to a communicable disease or if they could have spread a disease to others. Businesses may be interested if a person passed near their store. Traffic engineers might need to infer which roads a person traveled on. Inferring these points is a way to account for missing data.

![Brownian Bridge with Time Integrated Out](image)

Figure 1: Brownian bridge with time integrated out. The two spikes represent the start and end points.

One way to infer a person’s location between two measured points is the Brownian bridge. This model is based on the assumption that the person moves according to Brownian motion, but both the starting and ending points are known. Applying the Brownian bridge to everyday location data appears to have originated in the animal tracking literature, with the 1991 master’s thesis of Bullard [3].

The Brownian bridge gives a way to interpolate between two points. For each time instant between the two points, the Brownian bridge says that the interpolated location \( x \) is described by a 2D Gaussian probability distribution \( f(x|t) \). If the first 2D location point \( x_a \) was measured at time \( t = 0 \), and the second location point \( x_b \) was measured at time \( t = T \), then the Brownian bridge describing an in-between point \( x \) is:

\[
f(x|t) = \frac{1}{2\pi\sqrt{|\Sigma(t)|}} e^{-\frac{1}{2}(x-\mu(t))^T\Sigma^{-1}(t)(x-\mu(t))}
\]

(1)

\[
= \frac{1}{2\pi\sigma(t)^2} e^{-\frac{1}{2\sigma(t)^2}(x-\mu(t))^2}
\] for \( 0 < t < T \)

(2)
where
\[ \mu(t) = x_a + \frac{t}{T}(x_b - x_a) \text{ for } 0 < t < T \] (3)
and
\[ \Sigma(t) = \begin{bmatrix} \sigma^2(t) & 0 \\ 0 & \sigma^2(t) \end{bmatrix} \text{ for } 0 < t < T \] (4)
and
\[ \sigma^2(t) = \frac{t(T-t)}{T} \sigma^2_m \text{ for } 0 < t < T \] (5)

Equation 2 is a time-dependent Gaussian distribution, where both the mean \( \mu(t) \) and variance \( \sigma^2(t) \) depend on time. The mean
in Equation 3 is simply a straight line segment from \( x_a \) to \( x_b \) with
constant speed over the time interval \([0, T]\). The nonzero elements
of the diagonal covariance matrix in Equation 4 are \( \sigma^2(t) \), which
is given by Equation 5. The variance is a parabola as a function
of time. It is zero at \( t = 0 \) and \( t = T \), and it rises to a maximum of \( \frac{T}{4} \sigma^2_m \)
at \( t = \frac{T}{2} \). The parameter \( \sigma^2_m \) is called the diffusion coefficient [9],
and it has units of length^2/time. We derive how to compute the
diffusion coefficient in closed form from data in Section 3.2.

The Brownian bridge is commonly drawn as shown in Figure 1.
This is the time-integrated Brownian bridge (TIBB), where time
has been integrated out, i.e.
\[ f(x) = \frac{1}{T} \int_0^T f(x(t))dt. \] (6)

This is a difficult integral that has been worked out in [22]. The
two peaks in \( f(x) \) occur at the two measurement points \( x_a \) and
\( x_b \). The TIBB is the mean over time of the Gaussian distributions
of the Brownian bridge, and it is customarily considered as the
time-independent, spatial probability distribution associated with
the Brownian bridge. In the animal tracking literature, e.g. [3, 9],
researchers build one TIBB for each pair of time-adjacent
points and sum them to estimate an animal’s home range. A larger
diffusion coefficient \( \sigma^2_m \) results in a TIBB that is more spread out in
space.

The main advantage of the Brownian bridge for interpolation is
that it acknowledges and precisely specifies the uncertainty of the
person’s location between measurements. In fact, Bullard [3] shows
how to incorporate measurement uncertainty for the measured
points \( x_a \) and \( x_b \), which we do not include in this paper. This
representation of uncertainty is a significant advantage over, say,
linear interpolation, which gives a deterministic assertion of the
in-between locations.

However, despite the representational advantage, it is not clear
that the Brownian bridge is an accurate model of the location
of a person between measurements. This is the main question
we attempt to answer in this paper. The Brownian bridge has been
applied to human location data. For instance, Venek et al. used
home range estimation similar to animal studies on human locations
to detect regular visits to certain places [23]. Lin and Hsu developed
an algorithm for short term prediction of human motion based on
matching real time measurements with historic trajectories. They
used the Brownian bridge to estimate the locations of in-between
points in the historic trajectories that temporally correspond to
points in the real time trajectories [17]. You and Krumm used the
Brownian bridge for inferring roads from sparse mobility data [24],
Palmer and Bartumeus used the model to understand environmental
exposure of tracked people [20], and Jongeling used it to segment
human trajectories. It has also been applied to military ships [5].
Alsolami et al. motivated their Brownian Bridge modeling using
human disease spread [1]. Hyland and Smith used the model for
pedestrian flow [10]. These applications build on the probabilistic
interpolation of the Brownian bridge. We are not aware of any
formal study testing the fundamental ability of the Brownian bridge
model to accurately estimate in-between locations.

There are several other methods for interpolating locations be-
tween two measured points, all of which fit into a simple taxonomy
with two dimensions. The first dimension distinguishes between
techniques that respect time and those that do not. Simple linear
interpolation, or higher order interpolation, is an example tech-
nique that respects time, giving a deterministic \( x(t) \) for the object’s
position. This approach is used frequently, such as for inferring
the path of a vehicle on a road between measured points [4]. A
category of techniques that does not respect time include those
that impose a maximum assumed speed on the object. If the object
travels at or below this maximum speed between \( x_a \) and \( x_b \), then
its in-between locations are confined to an ellipse whose foci are
the two endpoints, as shown by Pfoser and Jensen [21]. Similar
arguments are spelled out by Burns [16] and Lenntorp [16].

In addition to time, the other taxonomic dimension is probability.
The two techniques mentioned above, deterministic interpolation
and the ellipse constraint, do not give a probability distribution for
the in-between points. This contrasts with another popular method,
kernel-density estimation (KDE), that replaces each measured point
with a kernel function and then sums the kernel functions. For
example, Biagioni and Eriksson use symmetric Gaussian kernels on
measured GPS data as a first step toward making a road map from
vehicle trajectories [2]. Note that KDEs are independent of time, so
the approach gives no inference about the in-between location as a
function of time.

There are several methods that give probabilistic interpolations
as a function of time, including the Brownian bridge. Others are the
familiar Kalman filter [11], particle filter [6], Gaussian process [25],
and Markov model [7].

<table>
<thead>
<tr>
<th>probabilistic</th>
<th>time sensitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>ellipse</td>
</tr>
<tr>
<td>yes</td>
<td>KDE</td>
</tr>
<tr>
<td>yes</td>
<td>Kalman filter</td>
</tr>
<tr>
<td>yes</td>
<td>particle filter</td>
</tr>
<tr>
<td>yes</td>
<td>Gaussian process</td>
</tr>
<tr>
<td>yes</td>
<td>Markov process</td>
</tr>
<tr>
<td>yes</td>
<td>Brownian bridge</td>
</tr>
</tbody>
</table>

Table 1: Taxonomy of interpolation methods for two spatial points.

It is this last category of interpolation, both time-sensitive and
probabilistic, that offer the most flexibility and potential for accu-
recision. This investigation concentrates on the Brownian bridge in
an effort to determine its suitability for interpolation for human loca-
tion data. There is no theoretical reason that the Brownian bridge
should describe human motion, but it has been used as such. In this

category of time-sensitive, probabilistic interpolants, the Brownian bridge is attractive because of its simplicity: its only adjustable parameter is the diffusion coefficient $\sigma_m^2$. However, it is important to understand its accuracy before taking advantage of its simplicity. In this paper, we assess the Brownian bridge for interpolating human location in two basic ways. One is a look at the consistency of the diffusion coefficient. We look at whether or not it is similar between people or similar between different spatiotemporal arrangements of the measured points. As part of this first set of tests, we develop a closed form solution for the maximum likelihood estimate of the diffusion coefficient. Our second assessment applies statistical tests to the distributional model of the Brownian bridge to see if it is supported by data. We perform our tests on location data from over 12 million individuals, described next.

2 DATA

This section describes the mobility data for our experiments. The raw mobility data consists of time stamped latitude/longitude points, as explained in Subsection 2.1. For analyzing the Brownian bridge model, we grouped the raw points into triples of points, explained in Subsection 2.2.

2.1 Raw Mobility Data

The data for our experiments came from SafeGraph\textsuperscript{1}, a company that aggregates cellphone location data. Each data point comes with an anonymous user identifier, a time stamp, and a latitude/longitude pair. While cellphone data has its own characteristics of accuracy and sampling frequency, it is the most ubiquitous location data available for humans, and thus serves as an important basis for testing the Brownian bridge on people. Our version of SafeGraph’s data is mostly concentrated in the U.S., thus we extracted coordinates inside a bounding box surrounding the American lower 48 states, shown in Figure 2. The southwest corner of this bounding box is at $(\text{latitude}, \text{longitude}) = (22.0^\circ, -125.0^\circ)$, and the northeast corner is at $(50.0^\circ, -66.0^\circ)$. We extracted data for the first three days of November 2020 measured in UTC. We took only users with at least three points to ensure enough data for at least one triple. In order to limit computation time, we took only users with at most 100 points over the three-day period. This resulted in about 12 million (12,394,326) distinct users with a total of about 314 million (313,673,781) latitude/longitude points. We denote a single time-stamped point as $(t, x)$, where $t$ is the time stamp and $x$ is the two-dimensional ordered pair giving the location.

2.2 Triples

Estimating the diffusion coefficient and other statistical computations require a set of time-stamped location triples. A triple is a set of three time-ordered points, with an example in Figure 3. We assume there are $n$ points to choose from for a moving object (e.g. a person), and the points are arranged in ascending temporal order. Our use of these triples is explained in Section 4.

In order to limit computation time, we took only triples whose endpoints, $(t_a, x_a)$ and $(t_b, x_b)$, are a maximum of 50 kilometers apart and whose time gap $t_b - t_a$ is a maximum of 48 hours. The point between the two endpoints is called $(t_c, x_c)$. It is this point that the Brownian bridge would model if the point were missing.

In the Brownian bridge research literature, the standard way of making triples from a sequence of an object’s time stamped location measurements was introduced by Horne et al. [9], illustrated in Figure 4. For $n$ points, there are $\binom{n}{3}$ possible Horne triples, and each point is used in at most one triple.

\textsuperscript{1}https://www.safegraph.com/
There are other ways to make triples, including overlapping groups of adjacent threesomes by one or two points, where points are reused in multiple triples. Taking reuse to the extreme, the maximum number of triples comes from taking every distinct combination of three points, maintaining their time ordering. This gives \( \binom{n}{3} = \frac{n(n-1)(n-2)}{6} \) possible triples, which we call the "all triples" case. We perform our tests in the next section on both Horne triples and all triples. From our data, we had a total of about 287 million (286,541,325) Horne triples and 174 billion (173,844,732,847) all triples.

2.3 Split by Users and (Distance,Time)
In deploying a Brownian bridge model to a group of moving objects, e.g. people, it would be most convenient to use the same Brownian bridge model for each object. Since the Brownian bridge is parameterized by only a single scalar diffusion coefficient \( \sigma_m^2 \), this would mean using the same diffusion coefficient for everyone. One of our experiments looks at how well the Brownian bridge model fits each individual in our data. Figure 5 shows the number of Horne triples and all triples for our subjects. We call these 'individual' triples, since the triples are grouped by person.

![Triple Count Histograms](image)

Figure 5: Histograms show the number of individuals with a given number of triples, for both Horne triples and all triples.

In case a single diffusion coefficient did not fit everyone, we also looked at fitting different diffusion coefficients to groups of triples with similar separations in space and time, mixing triples from different users. The separation distance of a triple is \( D = |x_b - x_a| \), and the time separation is \( T = t_b - t_a \). We split the \((D, T)\) space into a grid of cells of size \((\Delta D, \Delta T) = (50 \text{ meters}, 5 \text{ minutes})\), looking at groups of triples in each cell, with \((0, 0) \leq (D, T) \leq (50 \text{ km}, 48 \text{ hours})\). This gives 1000 different \( D \) values and 576 different \( T \) values for a total of 576,000 different separations. The number of triples for each separation, for both Horne and all triples, is shown in Figures 6a and 6b. We call these triples "collective" triples, since each \((D, T)\) separation mixes triples from different individuals. Through the remainder of this paper, we will perform experiments on Horne triples and all triples for both individual and collective triples in our search for a statistically accurate version of the Brownian bridge for human mobility.

3 DIFFUSION COEFFICIENT

The Brownian bridge has only one free parameter, the diffusion coefficient \( \sigma_m^2 \) whose units are \( \text{length}^2/\text{time} \). The diffusion coefficient controls the variance of the Gaussian distribution, and a larger \( \sigma_m^2 \) allows for more random deviations from the straight line path between \( x_a \) and \( x_b \). This section investigates the consistency of the diffusion coefficient from person to person. While researchers have not claimed that all people have the same diffusion coefficient, it would be convenient if it were true, because then we could apply the same Brownian bridge model to all human location data. It would also be interesting to discover such a universal constant for human mobility.

We also investigate in this section if there is some consistent variation in the diffusion coefficient as a function of the time and distance between \( x_a \) and \( x_b \). It may be that points separated by short times and small distances have consistent diffusion coefficients that are reliably different from points with larger separations.

For all our tests, we look at both Horne triples and all triples. Because these tests require the computation of millions of diffusion coefficients, we first develop a more efficient technique for computing \( \sigma_m^2 \) that does not require a numerical search, which has been the standard approach until now.

3.1 Maximum Likelihood for Diffusion Coefficient

The standard practice for finding the diffusion coefficient is maximum likelihood estimation as described in the frequently cited paper by Horne et al. [9]. While [9] proposes a numerical search for the diffusion coefficient, we show in this section that there is a simple, exact, closed form solution for the maximum likelihood estimator, which enables us to use a very large dataset for our experiments.

The input data for fitting a Brownian bridge is a set of \( N \) time-stamped location triples: \( \{(t_{i,a}, x_{i,a}), (t_{i,c}, x_{i,c}), (t_{i,b}, x_{i,b})\} \) for \( i \in [1...N] \) and \( t_{i,a} \leq t_{i,c} \leq t_{i,b} \). Here \( x \) is a two-dimensional location vector and \( t \) is the time at which \( x \) was measured. As shown in Figure 3, a triple is three time-stamped points subscripted with \( a, c \), and \( b \), where \( a \) represents the first point, \( b \) represents the last point, and \( c \) represents a point between \( a \) and \( b \) in time. We think of \( a \) and \( b \) as the two endpoints of a Brownian bridge and \( c \) as a point on the way from \( a \) to \( b \). Then the Brownian bridge says the likelihood of the location of \( x_{i,c} \) is

\[
L_i = \frac{1}{2\pi\sigma^2(t_i)} e^{-\frac{1}{2\sigma^2(t_i)} \| x_{i,c} - \mu(t_i) \|^2} \tag{7}
\]

where

\[
t_i = t_{i,c} - t_{i,a} \\
\bar{t}_i = t_{i,b} - t_{i,a} \\
\mu(t_i) = x_{i,a} + \frac{t_i}{\bar{t}_i} (x_{i,b} - x_{i,a}) \\
\sigma^2(t_i) = \frac{t_i(\bar{t}_i - t_i)}{\bar{t}_i^2} \sigma_m^2
\]
These equations are the same as Equations 2 - 5, but written with subscripted triples \((t_{i,a}, x_{i,a})\), \((t_{i,b}, x_{i,b})\), and \(t_i\) and \(T_i\). The total likelihood for all \(N\) triples is:

\[
L = \prod_{i=1}^{N} L_i.
\]

To get the maximum likelihood solution in closed form, we rewrite the likelihood for point \(i\) as:

\[
L_i = \frac{1}{\alpha_i \sigma_m^2} e^{-\frac{\beta_i}{\sigma_m^2}}
\]

where

\[
\alpha_i = 2\pi t_i(T_i - t_i) / T_i
\]

\[
\beta_i = T_i |x_{i,c} - \mu(t_i)|^2 / 2 t_i(T_i - t_i)
\]

The total log likelihood is then

\[
l = \ln(L) = \ln \left( \prod_{i=1}^{N} L_i \right)
\]

\[
= \sum_{i=1}^{N} \ln(L_i)
\]

\[
= \sum_{i=1}^{N} \ln \left( \frac{1}{\alpha_i \sigma_m^2} e^{-\frac{\beta_i}{\sigma_m^2}} \right)
\]

\[
= - \sum_{i=1}^{N} \ln(\alpha_i \sigma_m^2) - \frac{1}{\sigma_m^2} \sum_{i=1}^{N} \beta_i
\]

We seek a maximum over \(\sigma_m^2\), so we differentiate with respect to \(\sigma_m^2\):

\[
\frac{dl}{d(\sigma_m^2)} = - \sum_{i=1}^{N} \frac{1}{\sigma_m^2} + \frac{1}{\sigma_m^2} \sum_{i=1}^{N} \beta_i
\]

\[
= - \frac{N}{\sigma_m^2} + \frac{1}{\sigma_m^2} \sum_{i=1}^{N} \beta_i
\]

Setting the derivative, Equation 12, to zero and solving for \(\sigma_m^2\) shows a single extremum at

\[
\hat{\sigma_m}^2 = \frac{1}{N} \sum_{i=1}^{N} \beta_i
\]

where \(\beta_i\) comes from Equation 10. This gives a convenient, closed-form solution for the maximum likelihood estimate of the diffusion coefficient.

To verify that the estimate from Equation 13 occurs at a maximum (not minimum) value of the log likelihood, we show that the second derivative is negative. Differentiating Equation 12, we have

\[
\frac{d^2l}{d(\sigma_m^2)^2} = \frac{N}{\sigma_m^4} - \frac{2N}{\sigma_m^4} \sum_{i=1}^{N} \beta_i
\]

\[
= \frac{N}{\sigma_m^4} - \frac{2N^2 \hat{\sigma_m}^2}{\sigma_m^4}
\]

where the simplification comes from Equation 13. Evaluating the second derivative at \(\sigma_m^2 = \hat{\sigma_m}^2\), we have

\[
\frac{d^2l}{d(\sigma_m^2)^2} \bigg|_{\sigma_m^2 = \hat{\sigma_m}^2} = \frac{N}{\sigma_m^4} (1 - 2N)
\]

This is always negative for \(N \geq 1\), thus \(\hat{\sigma_m}^2\) is the maximum log likelihood estimate.

We recommend using Equation 13 for estimating the diffusion coefficient instead of a numerical search. Equation 13 requires only one
We investigated the behavior of the computed diffusion coefficients with small diffusion coefficients, there is also a relatively long tail because small separations may be associated with more directed that smaller separations might have smaller diffusion coefficients, of individuals. We are curious if there is any systematic variation at the same triples as the individual case, but here the triples are 3.2.2 Collective Diffusion Coefficients.

Horne triples or all triples with the more robust median statistic. For both the untrimmed and trimmed data. But there is much less the histograms are similar but not identical. The mean values, from rather than just Horne triples (red bars in Figure 7a), we see that 3,354,072,073 m4 respectively. This large difference between the mean and median, along with the large standard deviation, implies a long tail or outliers. Thus we also trimmed the diffusion coefficient values, eliminating the top 5%, giving a mean, median, and standard deviation of 810.5, 0.126, and 2896 m2/s respectively. We can conclude that the diffusion coefficients vary widely from person to person, so it is best to compute a separate coefficient for each person rather than use the same value for everyone. However, if it is necessary to use one diffusion coefficient for everyone in a study, the trimmed median value for Horne triples of 0.126 m2/s is a reasonable choice.

Using instead all triples for individuals (green bars in Figure 7a), rather than just Horne triples (red bars in Figure 7a), we see that the histograms are similar but not identical. The mean values, from Table 2 are quite different between the Horne and all triples cases for both the untrimmed and trimmed data. But there is much less difference in the medians, implying little difference in using just the Horne triples or all triples with the more robust median statistic.

3.2.2 Collective Diffusion Coefficients. The collective case looks at the same triples as the individual case, but here the triples are grouped by separations into a grid of different (D,T) groups, where (ΔD,ΔT) = (50 meters, 5 minutes). Each (D,T) group has a mix of individuals. We are curious if there is any systematic variation in the diffusion coefficient as a function of (D,T). It is plausible that smaller separations might have smaller diffusion coefficients, because small separations may be associated with more directed routes between the endpoints.

Figure 7b shows histograms of the diffusion coefficient for the collective case, where each count in the histogram comes from one (D,T) separation. Here the values have a much wider range than for individuals, borne out by the statistics in Table 2.

To check for systematic variation as a function of separation, we plotted the diffusion coefficients as a function of (D,T) as shown in Figures 8a and 8b for the Horne and all triples cases respectively. For both cases, there is no apparent consistent variation in diffusion coefficients across the range of separation values. Even nearby separations show no consistency, implying there is no simple, systematic relationship between the diffusion coefficient and the separation of the triple.

<table>
<thead>
<tr>
<th>Triples</th>
<th>Mean</th>
<th>Median</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>individual Horne</td>
<td>9.867·10^6</td>
<td>1.717·10⁻¹</td>
<td>3.354·10⁹</td>
</tr>
<tr>
<td>individual all</td>
<td>4.541·10^6</td>
<td>1.748·10⁻¹</td>
<td>1.946·10⁹</td>
</tr>
<tr>
<td>collective Horne</td>
<td>5.389·10⁶</td>
<td>3.810·10⁻³</td>
<td>7.975·10⁸</td>
</tr>
<tr>
<td>collective all</td>
<td>2.435·10⁻⁵</td>
<td>3.426·10⁻⁴</td>
<td>5.872·10⁶</td>
</tr>
<tr>
<td>trimmed individual Horne</td>
<td>8.105·10⁻²</td>
<td>1.256·10⁻¹</td>
<td>2.896·10³</td>
</tr>
<tr>
<td>trimmed individual all</td>
<td>9.764·10⁻⁶</td>
<td>1.717·10⁻¹</td>
<td>3.410·10⁹</td>
</tr>
<tr>
<td>trimmed collective Horne</td>
<td>7.007·10⁻⁴</td>
<td>3.127·10⁻³</td>
<td>2.286·10⁵</td>
</tr>
<tr>
<td>trimmed collective all</td>
<td>5.331·10⁻⁶</td>
<td>3.809·10⁻³</td>
<td>8.085·10⁸</td>
</tr>
</tbody>
</table>

Table 2: Statistics of diffusion coefficients with different sets of triples. The trimmed values are based on data with the top 5% of diffusion coefficient values deleted. All values are in units of m²/s.

3.2.3 Diffusion Coefficients Summary. We conclude that people can have very different diffusion coefficients from each other. Using Horne triples or all triples of an individual does not make a large difference in the distribution of diffusion coefficients. We do not see a consistent variation in diffusion coefficient as a function of the spatiotemporal triple separation. Overall, it appears there is not a simple way to characterize the diffusion coefficient for human location due to person-to-person variation and variation over triple separations.

4 TESTING BROWNIAN BRIDGE MODEL

While the Brownian bridge model has been applied to human mobility data, it appears the model has not yet been tested for applicability in a rigorous way. This section describes our statistical tests of the Brownian bridge on human mobility data. We first test the model’s prediction for the mean of the in-between point, which the model predicts as a linear interpolation between the endpoints. Then we test for the normality of the in-between point. For both these tests, we test on both individualized mobility data and collective data containing all the individuals.

4.1 Mean

Recall that the Brownian bridge model says the mean of an in-between point is given by

\[ \mu(t) = x_a + \frac{t}{T}(x_b - x_a) \]  (14)
If this is true, then for an in-between point \((t_c, x_c)\), we would expect \(E[\mu(t_c) - x_c] = 0\), where \(x_c\) is the actual in-between point and \(\mu(t_c)\) is the predicted in-between point from the Brownian bridge model. Formulated for statistical testing, the null hypothesis is that the mean of \(\mu(t_c) - x_c\) is the zero vector. An appropriate test for multivariate means is based on the Hotelling’s T-square test statistic, which we use. A larger p-value means the data is more likely to conform to the mean prediction of the Brownian bridge.

The p-values of the Hotelling’s test are shown in Figure 9. The curves show the fraction of tests whose p-value is greater than or equal to the p-value on the horizontal axis. The curves for individuals (“individual Horne triples” and “individual all triples”) show the results of testing on all 12 million users. The curves for collective data (“collective Horne triples” and “collective all triples”) show results of testing on all the triple separations \((D, T)\), i.e. 1000 different separations in distance and 576 separations in time, for a total of 576,000 separations, as described in Section 2.3.

As an example curve in Figure 9, we will examine the blue curve, which comes from testing the Brownian bridge mean prediction on individuals using Horne triples. At \(p = 0.05\) on the horizontal axis, it is apparent that a fraction of 0.983 of individuals had \(p \geq 0.05\), meaning we fail reject the null hypothesis that the mean is zero in these cases, lending support to the Brownian bridge’s mean prediction. (Here \(p\) is the probability of observing the data if the
null hypothesis is true. So a low p-value indicates the probability of observing the data is unlikely if the mean is actually zero. The same is true for the orange curve, which shows p-values for the collective data using Horne triples. The specific fractions associated with the mean test at \( p \geq 0.05 \) are shown in Table 3.

The two lower curves of Figure 9 show results for the individual and collective tests using all triples rather than Horne triples. In both of these cases, the null hypothesis is usually rejected at \( p = 0.05 \). This means that this data tends to reject the mean prediction.

We conclude from the statistical test of means that the Brownian bridge model for means appears reasonable for Horne triples for both individual and collective data. It is not reasonable when using all possible triples. This may be because Horne triples are generally separated less in space and time compared to all possible triples. These more closely spaced triples give less opportunity for the in-between point to veer from a straight line segment connecting the two endpoints.

### 4.2 Normality

Just as the Brownian bridge model makes a prediction for the mean of the in-between point, it also gives a distribution for the in-between point. This distribution is a Gaussian with a time-varying mean vector and covariance matrix as in Equation 2. We would like to test if the in-between point actually has a Gaussian distribution. It would be convenient to use a standard normality test, but the time-varying mean and variance pose a problem for such tests. However, we can transform the in-between point into a 2D standard normal distribution and apply a traditional normality test. For an in-between point \((t_c, x_c)\), the transformation is:

$$z_c = \frac{1}{\sigma(t_c)} \begin{pmatrix} x_c - \mu(t_c) \\ \frac{1}{\sigma(t_c)} \end{pmatrix}$$

which gives \( z_c \sim N(0, I) \). Since we do not know the value of \( \sigma_m \), we set it to one. This changes the covariance of \( z_c \), but the distribution still remains normal, which is what we test for.

One highly recommended test for multivariate normality is the Henze-Zirkler test \([8, 12, 18]\), which is what we use here. Similar to our mean tests, we test the normality of each individual’s data as well as the collective data in each group of triple separations. The null hypothesis is that the data are distributed normally. The resulting p-values are shown in Figure 10. At \( p = 0.05 \), the fraction of cases failing to reject the null hypothesis are given in Table 3, which shows the majority of cases are not normal. This indicates, for our tests on human mobility data, that the Brownian bridge model does not accurately reflect the distribution of the in-between points.

### 5 CONCLUSION

The goal of this work was to test the suitability of the Brownian bridge for interpolating human location data. There are several methods for such interpolation, but the Brownian bridge is attractive due to its simplicity and its frequent use in animal tracking. This sort of probabilistic interpolation is useful for inferring where a person may have been between location measurements, including their possible exposure to disease. The Brownian bridge model also underpins other higher-level inference processes, so it is important to assess the model as a foundation.

Our tests are all based on location tracking data from over 12 million people. The fundamental unit for testing is a trip of time-ordered location points. We made triples in the traditional way (Horne triples) and also by taking all possible combinations of three points for each person. In addition to looking at each individual’s triples, we also grouped the triples by their separation in distance and time to look for systematic variations with these two quantities.

Our first investigation looked at the diffusion coefficient, which is the model’s only adjustable parameter. After introducing the first closed form solution for the maximum likelihood estimate of the diffusion coefficient, we computed diffusion coefficients for

<table>
<thead>
<tr>
<th>Data</th>
<th>Fraction Fail Reject Null p = 0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>null: zero mean</td>
<td>collective all triples 0.238</td>
</tr>
<tr>
<td>null: normality</td>
<td>individual all triples 0.393</td>
</tr>
<tr>
<td></td>
<td>collective Horne triples 0.994</td>
</tr>
<tr>
<td></td>
<td>individual Horne triples 0.983</td>
</tr>
</tbody>
</table>

Table 3: At the \( \rho = 0.05 \) level, this table shows the fraction of instances that fail to reject the null hypothesis of the Brownian bridge prediction for the mean and normality of the in-between point. These are the fractions of cases that support the Brownian bridge model.
each individual and each triple separation. For both the individual and collective data, we found a large variation in the diffusion coefficient. This means there is not a single instantiation of the Brownian bridge that fits all people nor all separations. However, we did suggest a median value to use for all individuals if necessary.

Our second investigation applied statistical tests for the mean and normality of the model. We found that the model’s prediction for the mean was statistically valid for Horne triples for both individuals and for collective separations. Our tests for normality showed that the data rarely supports the Gaussian distribution of the in-between points. This is the main reason that we conclude that the Brownian bridge is not an accurate representation of the distribution of in-between points for human mobility.

Future work on the suitability of the Brownian bridge should concentrate on its variations, such as the bivariate [14] and dynamic [13] Brownian bridge models. It would be valuable to apply studies like the present to other forms of human mobility interpolation, such as those in Table 1. Another worthwhile effort would be to compute an empirical, probabilistic interpolation bridge from data.

REFERENCES


