Statistical Methods for the Forensic Analysis of Geolocated Event Data

By:
Christopher Galbraith (Department of Statistics, University of California, Irvine), Padhraic Smyth (Department of Computer Science, University of California, Irvine), and Hal S. Stern (Department of Statistics, University of California, Irvine)

From the proceedings of
The Digital Forensic Research Conference
DFRWS USA 2020
July 20 - 24, 2020

Forensic Science International: Digital Investigation 33 (2020) 301009

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

Statistical Methods for the Forensic Analysis of Geolocated Event Data

Christopher Galbraith [a, *], Padhraic Smyth [b], Hal S. Stern [a]

[a] Department of Statistics, University of California, Irvine, Bren Hall 2019, Irvine, CA, 92697, USA
[b] Department of Computer Science, University of California, Irvine, Bren Hall 3019, Irvine, CA, 92697, USA

ARTICLE INFO

Article history:

Keywords:
Event data
Spatial point pattern
Geolocation data
Geoparcel data
Likelihood ratio
Score-based likelihood ratio

ABSTRACT

A common question in forensic analysis is whether two observed data sets originated from the same source or from different sources. Statistical approaches to addressing this question have been widely adopted within the forensics community, particularly for DNA evidence. Here we investigate the application of statistical approaches to same-source forensic questions for spatial event data, such as determining the likelihood that two sets of observed GPS locations were generated by the same individual. We develop two approaches to quantify the strength of evidence in this setting. The first is a likelihood ratio approach based on modeling the spatial event data directly. The second approach is to instead measure the similarity of the two observed data sets via a score function and then assess the strength of the observed score resulting in the score-based likelihood ratio. A comparative evaluation using geolocated Twitter event data from two large metropolitan areas shows the potential efficacy of such techniques.

1. Introduction

There is a growing need for the development of quantitative statistical methodologies in digital forensics. The OSAC Task Group on Digital/Multimedia Science recently issued a recommendation for the development of "systematic and coherent methods for studying the principles of digital/multimedia evidence to assess the causes and meaning of traces in the context of forensic questions, as well as any associated probabilities" (Pollitt et al., 2019). In addition, as stated recently in Casey (2018), there is "a growing expectation that forensic practitioners treat digital traces in a manner that is becoming widely accepted in forensic science: evaluating and expressing the relative probabilities of the forensic findings given at least two mutually exclusive hypotheses." Existing forensic tools for digital evidence, however, are often focused on supporting the process of information extraction from digital devices followed by exploratory analysis (e.g., see Roussev, 2016; Årnes, 2017; SWGDE, 2019), with relatively little support for statistical quantification.

In particular, logs of geolocation data are now routinely available on modern mobile devices. This type of data is typically associated with events generated on the device, such as actions taken by a user in a software application. Such data can be collected in a variety of ways—from the device itself, from servers that store the locations based on IP addresses, from cellular towers, and so on. Given the general prevalence of mobile devices, this type of spatial event data is now encountered with increasing regularity during forensic investigations. For instance, an investigator might wish to determine if two sets of events with geolocations, corresponding to different accounts or devices, were in fact generated by the same individual.

The forensic problem of identification of source from observed evidence has been well-studied. Statistical techniques have played a key role in forensic analysis, providing investigators with tools that allow them to make robust inferences from limited and noisy data. The best-known example is the use of likelihood ratio techniques for determining if a DNA sample from a crime scene is a match to a suspect's DNA sample. For other types of evidence—including fingerprints, shoeprints, and bullet casing impressions—the development of quantitative methodologies is more challenging (Stern, 2017). In particular, there are significant challenges in developing realistic statistical models, both for capturing the process by which the evidential data is produced and for modeling the inherent variability of such data from a relevant population.

The primary contribution of this paper is the development of quantitative techniques for forensic analysis of geolocated event data. In particular we investigate two types of approaches to obtain strength of evidence: a likelihood ratio approach based on modeling the evidential data directly and a score-based likelihood

* Corresponding author.
E-mail address: galbraic@uci.edu (C. Galbraith).

https://doi.org/10.1016/j.fsidi.2020.301009

ratio that instead models a summary measure of the similarity of the evidence.

2. Motivating example

Suppose that a forensic investigator is given a set of GPS coordinates associated with criminal activity and is tasked with finding the most likely suspect from a set of individuals for whom reference location data is available. The GPS coordinates could be the locations of crime scenes (e.g., in the case of serial crime) or data gathered from a device of unknown origin (e.g., a burner phone recovered from a crime scene). In either case, we do not know who generated this location data and will refer to it as the *unknown source data*.

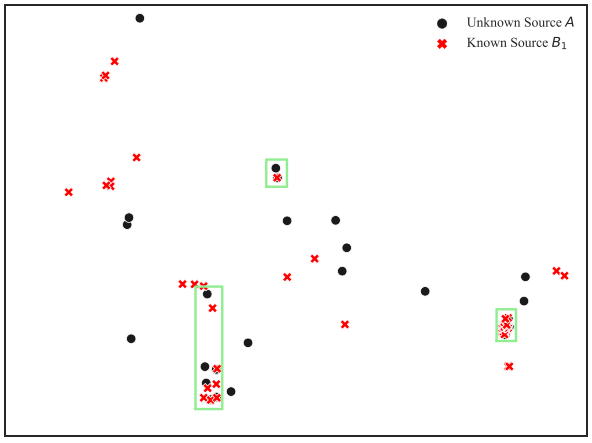
One investigative approach in this context is to gather location data for a set of potential suspects via a geofence warrant (e.g., Valentino-DeVries (2019)). A geofence warrant refers to a situation where a "fence" or bounding box is constructed around a set of locations, such as locations associated with a crime. A law enforcement agency then requests data from a service provider (such as Google or Twitter) for any individuals whose devices were within the geofence during a time-period of interest (e.g., in a window of time around which the criminal activity occurred). For individuals who match the geofence (i.e., potential suspects), their geolocation data is given an anonymous identifier and their data is sent to law enforcement to aid in the investigation. We will refer to the location data for these individuals as the *known source data* because, once persons of interest have been identified, the service provider can reveal their identities.

Fig. 1 provides an illustrative example of such a geofence situation. The data points are geolocated events, with colors and shapes indicating different accounts. Here we treat A (black points) as the unknown source data, where each point has an associated geofence surrounding it whose size and shape was determined based on the land parcel data described in Appendix C. Fig. 1a and b show geolocation events from two different known source accounts B_1 (red crosses) and B_2 (blue triangles) with at least one GPS coordinate inside a geofence (highlighted by green boxes).

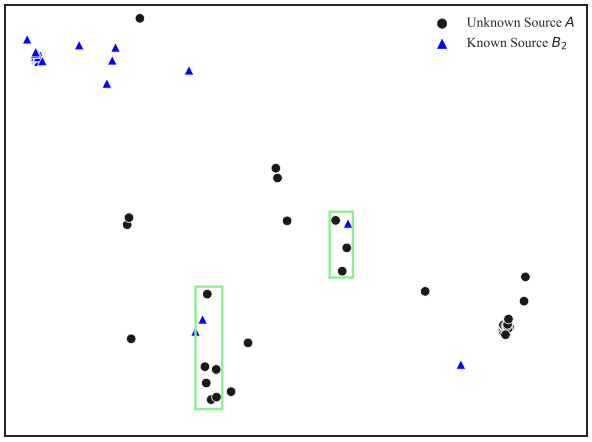
An investigator looking at this data would need to infer how likely it is that the locations in each panel of the figure match to the same source (e.g., were generated by the same individual). In this example, we have selected the data so that the points in Fig. 1a are from the same account (over different time-periods) and the points in 1b are from different accounts. Determining if sets of locations "match" can be a difficult task due to many factors including variability in human behavior and the typicality of locations of interest (e.g., how common they are in the population in general). To address this problem we propose a technique that, given two sets of locations, produces an objective measure of their probative evidential value. Using our method for the data in Fig. 1, an investigator would be able to conclude that there is strong support for the hypothesis that the two sets of locations in Fig. 1a were generated by the same individual. She would also be able to conclude that the individual that generated the known source data (blue triangles) in Fig. 1b can likely be excluded as the source of the unknown source data (black points). In the remainder of this paper, we will show how to produce such conclusions given this type of location data.

3. Related work

In prior work we have developed statistical methods for same-source questions involving temporal user-generated event data (Galbraith and Smyth, 2017; Galbraith et al., 2020). In this paper we extend these approaches to the spatial domain.



(a) Same source



(b) Different sources

Fig. 1. Location data (taken from Section 9) in a 3.5 square mile region of Orange County, CA. Green boxes represent geofences with events in both sets. (a) Both the unknown and known source data were generated by the same individual; (b) the unknown and known source data were generated by different individuals. The unknown source data is the same in both panels.

Evaluating location-related mobile device evidence and expressing probative conclusions in the forensic setting is challenging due to both technological and circumstantial subtleties that can be present in the data. Casey et al. (2020) discuss these challenges and present a structured framework for the evaluation of geolocation data. However, the hypotheses considered in that work are focused on specific locations of interest rather than comparing sets of spatial patterns (which is the focus here).

The recent work of Bosma et al. (2019) is similar in spirit to our work in that they address same-source problems using mobile geolocation data. They develop a method that uses the location and time of cellular tower registrations of mobile phones to assess the strength of evidence that a pair of phones were used by the same person. Their approach creates features from the cell tower data and makes parametric modeling assumptions via logistic regression in how those features indicate same- and different-source phone usage patterns. The methods that we propose in this paper differ in that we make no such parametric assumptions, and no data has to be held out to estimate model parameters (although we do require a reference set of data in order to estimate the typicality of locations, e.g., how frequently-visited they are by the population in general).

From a statistical perspective, there is also a general line of work known as spatial point patterns, which focuses on the development of methodologies for modeling and evaluation of dependence between spatial sets of locations (e.g., Berman, 1986; Schlather et al., 2004). Much of this type of work relies on assumptions such as spatial homogeneity that are not well-suited to the type of bursty and non-stationary human-generated event data that is often of interest in a forensics setting. Nonetheless this prior work in spatial point processes can provide a useful starting point for analyzing spatial event data in a systematic manner.

4. Notation & problem statement

To formally define the question of interest, we adopt notation and terminology from the forensic statistics literature. A common problem in forensic science is that of determining the degree to which two samples of pattern evidence "match," or have the same generative mechanism (e.g., Aitken and Taroni, 2004). The evidence corresponds to observed data and can take different forms such as measurements related to DNA, fingerprints, or shoe prints. Denote the *evidence* as (A, B), where in general.

A: set of observations for a sample from an unknown source (e.g., a sample recovered from a crime scene),
B: set of observations for a reference sample from a known source (e.g., a sample from a suspect).

For geolocated event data, sets A and B could consist of locations at which actions were taken on two different devices, e.g., locations where phone calls were made. The forensic question of interest in this scenario would be to determine how likely it is that the events on the different devices were generated by the same individual. Alternatively, sets A and B could consist of locations at which events were generated from a single account (e.g., accounts on a social media platform such as Twitter) or locations from the same device but over two different time periods. The forensic question of interest would be to determine if the same individual was responsible for generating both sets of events. This scenario is relevant for example when the person of interest invokes the "it wasn't me" defense, with A corresponding to events for which the individual claims they are not responsible and B corresponding to a sample of his or her typical activity.

In the scenarios above, A and B refer to sets of (longitude, latitude) coordinates at which events occurred. Fig. 2 provides an example of such geolocated event data. In this specific example $A = \{(-73.984, 40.754), (-73.977, 40.761), \ldots, (-73.987, 40.727)\}$ for a total of $n_a = 42$ events in A, and $B = \{(-73.988, 40.742), (-74.009, 40.711), \ldots, (-73.995, 40.718)\}$ for a total of $n_b = 39$ events in B.[1]

The goal of a forensic examination is to assess the likelihood of observing the evidence (A, B) under two hypotheses

$H_s : (A, B)$ came from the same source,

$H_d : (A, B)$ came from different sources.

In the context of the geolocation data we will be focusing on in this paper, the term "source" refers to a specific individual or user account, and the term "came from" can be interpreted as meaning "generated by." Thus, H_s is the proposition that the sample from the unknown source A was generated by the same individual or user account as the sample from the known source B. H_d is the proposition that the sample from the unknown source A was *not* generated by the specific source of B, but instead from another individual among an alternative source population. Ommen and Saunders (2018) provide an in-depth discussion of the competing propositions.

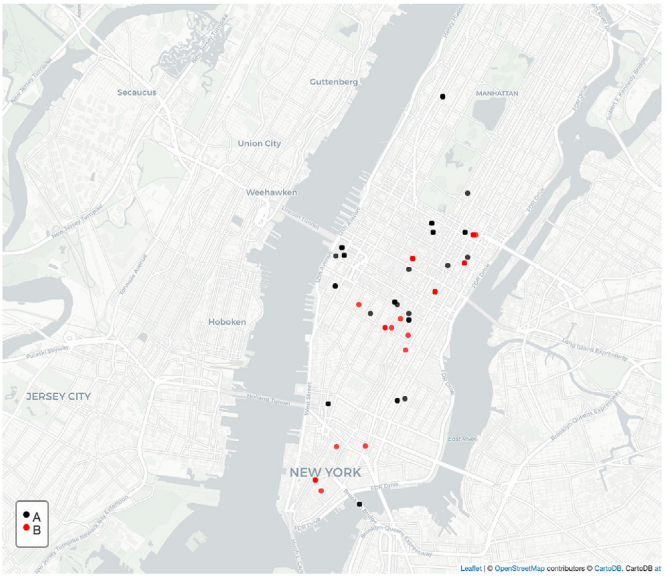


Fig. 2. Example of sets of locations for Twitter data from New York. The patterns correspond to geolocatons of tweets from the same account over two different months, with month 1 corresponding to A (red) and month 2 corresponding to B (black).

In this paper, we propose and investigate two approaches for assessing the *strength of evidence* in this context. The first is a likelihood ratio approach that uses kernel density estimation techniques to estimate the relative likelihood of the observed location evidence under each proposition, H_s and H_d. The second approach is to instead measure the similarity of the two sets of locations via a score function and then assess the strength of the observed score resulting in the score-based likelihood ratio.

Sections 5–7 discuss the technical details of our proposed approaches for computing the probative value of location evidence. Readers who would like to skip these details can go directly to Section 8, which shows how to form conclusions from the numeric values computed, and Sections 9–11, which provide a case study and discussion of the results.

5. The likelihood ratio

The likelihood ratio (LR) is widely accepted in the forensic science community as "a logically defensible way" to assess the strength of evidence (Willis et al., 2016). It has been applied in a variety of forensic disciplines, including fingerprints (Champod and Evett, 2001) and DNA (Evett and Weir, 1998). The LR arises naturally in the application of Bayes' Theorem to updating the relative likelihoods (odds) of the two competing hypotheses (same- and different-source) given the evidence (A, B). Bayes' Theorem in the forensic context is

$$\underbrace{\frac{Pr(H_s|A,B)}{Pr(H_d|A,B)}}_{\textit{posterior}\ \text{odds}} = \overbrace{\frac{Pr(A,B|H_s)}{Pr(A,B|H_d)}}^{\text{likelihood ratio}} \underbrace{\frac{Pr(H_s)}{Pr(H_d)}}_{\textit{prior}\ \text{odds}} \tag{1}$$

[1] The latitude and longitude values presented in the text were rounded. Generally much higher precision is available, e.g., for coordinates provided by GPS.

where $Pr(\cdot)$ refers to the appropriate probability distribution. For the likelihood ratio term these are probability distributions for the evidence A and B (i.e., either a probability mass function or probability density function) and for the prior and posterior odds these are probabilities assigned to the hypotheses.[2]

The likelihood ratio measures the relative probability of observing the evidence (A, B) under each of the two competing hypotheses. A large likelihood ratio means the observed evidence is much more likely under the same-source hypothesis H_s than the different-source hypothesis H_d. A small LR means that the observed evidence is much less likely under the same-source hypothesis. Equation (1) tells the evaluator of the evidence (e.g., a member of the jury) how to modify his or her prior odds given the evidence to obtain posterior odds of the two hypotheses. One common view is that the goal of the forensic examination is to supply the LR to said evaluator. See Stern (2017) for a thorough discussion of the likelihood ratio and its application in forensic science.

The likelihood ratio in Equation (1) requires probabilistic generative models $Pr(\cdot)$ for the evidence (A, B). Specifying such models can be extremely difficult in practice. One needs to construct two models that not only specify the distribution of the locations of the events in A and B but also the correlation between those locations under the same- and different-source hypotheses. For that reason, we pursue two different approaches that avoid the complexities of specifying such distributions. The first is a likelihood ratio that conditions on one set of events (rather than modeling the joint probability of both sets), and the second is a score-based likelihood ratio that computes a likelihood ratio based on some similarity function defined on the two sets of events. These general approaches have been proposed in the statistical forensics literature in the past but have not previously been applied to spatial event data.

Finally, note that in this paper we treat the event locations in A and B as real-valued numbers in the two-dimensional plane, and thus the numerator and denominator terms in the likelihood ratio are modeled via probability densities, henceforth referred to by $f(\cdot)$.

6. Computing the likelihood ratio

A well-known way (Stern, 2017) to simplify the likelihood ratio is to factor the joint distribution of (A, B) under each model such that

$$\frac{f(A,B|H_s)}{f(A,B|H_d)} = \frac{f(B|A,H_s)f(A|H_s)}{f(B|A,H_d)f(A|H_d)}. \tag{2}$$

In this scenario we can simplify $f(A|H_s) = f(A|H_d) = f(A)$ because the distribution of the locations A do not depend on the same- or different-source hypothesis. Furthermore it is natural to assume that the distribution of locations B is independent of the distribution of locations A under the different-source hypothesis, which results in the simplification $f(B|A, H_d) = f(B|H_d)$. Given these assumptions, the *likelihood ratio (LR)* for the same source problem can be written as

$$LR = \frac{f(B|A,H_s)}{f(B|H_d)}. \tag{3}$$

Traditional parametric models for the conditional densities $f(B|A, H_s)$ and $f(B|H_d)$ above, such as spatial Poisson point process models, are often insufficient to capture the typical characteristics of user-generated geolocated event data that tends to be bursty and inhomogeneous. For that reason, we focus on non-parametric kernel density estimation techniques for modeling sets of locations A and B.

To estimate the likelihood ratio, we first define a *reference population* E of geolocated events, denoted $E = \{e_k : k = 1, \ldots, n_p\}$ where e_k is the (longitude, latitude) coordinate of the kth event and n_p is the total number of population events in E. For a particular set of geolocated events B, the probability density function in the denominator of Equation (3), $f(B|H_d)$, can be estimated by

$$\widehat{f}\left(B|H_d\right) = \prod_{j=1}^{n_b} f_{KD}\left(e_j^b \middle| E\right) \tag{4}$$

where $f_{KD}(\cdot|E)$ is a kernel density built on the population data E, and e_j^b is the location of the jth event in B. See Appendix A for a more detailed discussion of kernel density estimation and a formal definition of f_{KD}. The term $f_{KD}(e|E)$ is the likelihood that a randomly selected event from the population will occur at some particular location e. Thus, Equation (4) is the likelihood of observing the set of locations B in the reference population, under the assumption of conditional independence of events given the model. Locations that are often-visited in the population (e.g., airports, shopping malls, etc) receive high probability in this model, while rare locations (e.g., individual homes, areas without cellular service, etc) receive low probability. Fig. 3a provides an illustration of a population model of this type, using the Twitter geolocation event data that we describe later in the paper in more detail.

The numerator of Equation (3), $f(B|A, H_s)$, is the probability of observing new location data B given that we have already seen location data A and under the hypothesis that A and B came from the same source. Effectively, it is a predictive density for geolocated events from A. We model this as a mixture of two densities where the first density corresponds to an individual component based on the locations of events in A, and the second density corresponds to the population component defined in Equation (4). See Fig. 3 for an example of such a mixture model using the data presented in the motivating example of Section 2. This addresses two potential problems. First, if A has very little data this model will appear similar to the population model resulting in LR values near 1 and proper calibration. Second, it allows for the possibility that an individual would visit new locations in a second sample.

We use a non-parametric kernel density approach for the mixture model components in $f(B|A, H_s)$, defined as

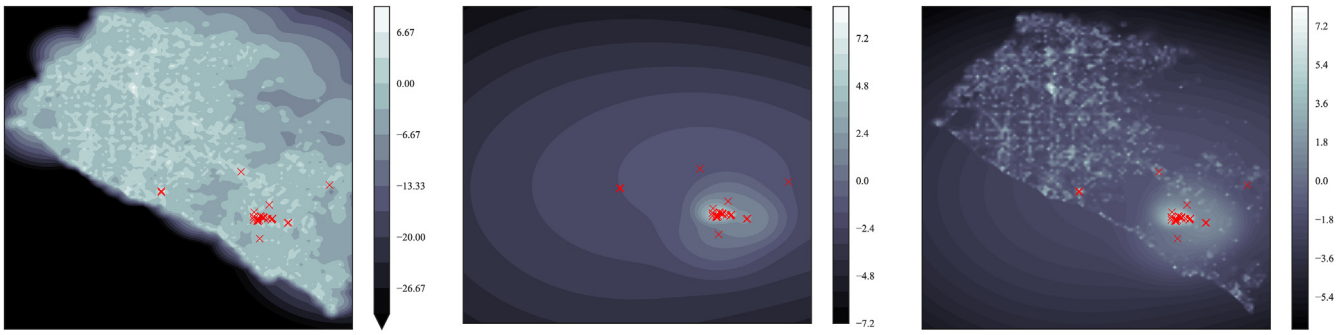
$$\widehat{f}\left(B|A,H_s\right) = \prod_{j=1}^{n_b} f_{MKD}\left(e_j^b \middle| A, E, \alpha\right). \tag{5}$$

Here $f_{MKD}(\cdot|A, E, \alpha)$ refers to a mixture of kernel densities (e.g., see Lichman and Smyth, 2014), defined as

$$f_{MKD}\left(e_j^b \middle| A, E, \alpha\right) = \alpha f_{KD}\left(e_j^b \middle| A\right) + (1-\alpha) f_{KD}\left(e_j^b \middle| E\right) \tag{6}$$

where $f_{KD}(\cdot|A)$ is a kernel density built on the unknown source data A, which we refer to as the individual component. The parameter $\alpha \in [0, 1]$ determines how much weight to put on the individual component $f_{KD}(\cdot|A)$ of the model relative to the population component $f_{KD}(\cdot|E)$. If the set of events B contains locations nearby to those in A, $\widehat{f}(B|A, H_s)$ will be large relative to $\widehat{f}(B|H_d)$ and the LR will have a value greater than 1 which indicates that A and B are

[2] More generally, the probability distributions in the likelihood ratio can be conditioned upon additional information that should be considered in evaluating the evidence. For convenience, we suppress notation regarding the additional information. For instance, this could be population data relevant to (A, B) as shown in Section 6.



(a) Population Component (b) Individual Component (c) Mixture

Fig. 3. Example of the KDE models used to estimate the likelihood ratio for Twitter events in Orange County, CA, from the experimental results in this paper. Overlaid on each panel are the set of points A from the motivating example in Section 2. (a) Population component used to estimate the denominator of the LR $f(B|H_d)$; (b) individual component built using the overlaid points; (c) mixture model with $\alpha = 0.8$ used to estimate the numerator of the LR $f(B|A, H_s)$.

likely to have been generated by the same individual.

7. The score-based likelihood ratio

Instead of specifying a generative model for the observed data, an alternative approach is to instead measure the similarity between sets of locations A and B via a *score function* $\Delta(A, B)$ that is usually univariate and continuous. Typically, low scores indicate the samples are similar, while high scores indicate considerable differences.

A natural approach to assess the strength of evidence via score functions is the *score-based likelihood ratio (SLR)*, which has been gaining popularity in forensic science (e.g., Bolck et al., 2015; Meuwly et al., 2017; Galbraith et al., 2020). Given an observed set of evidence (A, B) related to a forensic investigation and the value of the score function for that evidence $\Delta(A, B) = \delta$, the SLR is defined as

$$SLR_\Delta = \frac{g(\Delta(A,B) = \delta | H_s)}{g(\Delta(A,B) = \delta | H_d)} \tag{7}$$

where $g(\cdot)$ denotes the conditional probability density function of $\Delta(A, B)$ given one of the two propositions (H_s for same-source or H_d for different-source). These conditional densities are typically straightforward to estimate via standard parametric or non-parametric techniques given data for a large number of instances or exemplars (A, B) under H_s and H_d. The numerator of the SLR can be interpreted as the likelihood of observing the score $\Delta(A, B) = \delta$ if A and B came from the same source. The interpretation of the denominator is the likelihood of observing this score if A and B came from different sources. The interpretation of the SLR is similar to that of the LR, with values greater than 1 favoring the same-source proposition. See Fig. 4 for an example of how the SLR approach might be applied.

In order to compute the SLR for a particular pair of sets of locations (A, B), we need a reference sample of exemplars from a relevant population. Assume that we have a sample of N accounts for a given spatial region, then the relevant data consist of the pairs (A_i, B_i) for $i = 1, \ldots, N$. Define a reference data set of all N^2 possible pairwise combinations constructed from these sets of locations, denoted $\mathscr{D} = \{(A_j, B_k) : j, k = 1, \ldots, N\}$. Assuming that (A, B) is not an element of $\mathscr{D}$,[3] we can use the scores of all of the N same source pairs, $\mathscr{D}_s = \{(A_i, B_i)\}$, to estimate the probability density function in the numerator of Equation (7) and the scores of all $N^2 - N$ pairs with different sources, $\mathscr{D}_d = \{(A_j, B_k) : j \neq k\}$, to estimate the probability density function in the denominator of Equation (7).

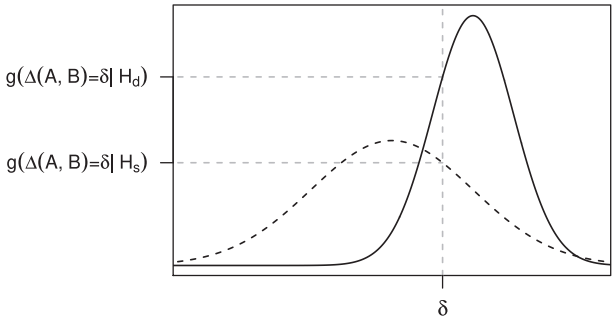


Fig. 4. Hypothetical illustration of the densities of the score function Δ under the hypotheses that the samples are from the same source (H_s, dashed line) and that the samples are from different sources (H_d, solid line). The score-based likelihood ratio SLR_Δ is the ratio of the conditional density functions g evaluated at δ.

Given the observed score $\Delta(A, B) = \delta$, we estimate the score-based likelihood ratio via

$$\widehat{SLR}_\Delta = \frac{\hat{g}(\Delta(A,B) = \delta | \mathscr{D}_s)}{\hat{g}(\Delta(A,B) = \delta | \mathscr{D}_d)} \tag{8}$$

where $\hat{g}$ is a kernel density estimator with a Gaussian kernel and rule-of-thumb bandwidth (Scott, 1992). We explicitly condition on the reference sets $\mathscr{D}_s$ and $\mathscr{D}_d$ because the score values for the point patterns in these sets along with the kernel density parameters fully specify the estimated density.

7.1. Score functions for geolocation data

In terms of defining a suitable score $\Delta(A, B)$ for sets of locations A and B, there are a number of techniques that can be borrowed from the statistics literature on spatial point patterns. In general, they fall into two categories: distance-based and area-based techniques (Haggett et al., 1977). Distance-based techniques use information on the spacing of points to characterize the pattern (typically, mean distance to the nearest neighboring point). Area-based techniques rely on characteristics of the frequency of observed points in subregions of the region under consideration. In this paper we investigate two different distance-based score functions $\Delta(A, B)$ to quantify the similarity of the points within the sets A and B and

[3] See Appendix E for the construction of the reference sets when (A, B) is an element of $\mathscr{D}$, as is the case for the results in Section 10.

incorporate area-based information via various event-weighting strategies. Full details on the score functions are provided in Appendix B, and a discussion of the motivation for using weights and definitions of the various weighting strategies used in this paper are provided in Appendices C and D respectively.

The two score functions we use are the mean nearest neighbor distance (denoted $\overline{D}_{min}$) and the earth mover's distance (denoted *EMD*), which both rely on computing the distance from each event in B to the nearest neighboring event in A. Intuitively, we expect same-source pairs to contain events at locations nearby each other in the spatial region as individuals tend to be self-consistent (repeatedly generating events from the same locations over time). If events in B are spatially clustered among (i.e., "close to") events in A, then the score functions considered tend to be smaller than if the A and B events are generated independently and do not spatially cluster together.

8. Drawing conclusions from the LR or SLR

After computing the likelihood ratio,[4] the forensic investigator can then come to a conclusion about the two propositions under consideration. This conclusion should express the degree of support provided by the evidence for the same-source hypothesis H_s versus the different-source hypothesis H_d depending on the magnitude of the LR. See Willis et al. (2016) for practical guidelines.

When the $LR = 1$ the conclusion should be that the evidence provides no assistance in distinguishing between the two hypotheses. For $LR > 1$ the conclusion should be that the evidence is more probable if the two sets of locations were generated by the same source. For $LR < 1$ the conclusion should be that the evidence is more probable if the alternative is true, i.e., that the two sets of locations were generated by different sources.

To aid in interpretability (e.g., for presentation to a jury), the likelihood ratio may be expressed by a verbal equivalent according to a scale of conclusions (Nordgaard and Rasmusson, 2012). Table 1 provides an example of such a verbal equivalent. For a more thorough discussion on expressing the probative value of forensic evidence in a clear and consistent manner, see Thompson (2017).

9. Data

Collecting data directly from a sufficiently large number of mobile devices for research purposes is difficult. For this reason, we used geolocation datasets of Twitter events to evaluate our proposed approaches. Twitter, a popular social media and microblogging service, provides a useful publicly accessible[5] source of user-event data that, given certain account configurations, exposes the geolocation of each event generated by that account. This data can be thought of as a subset of data collected from a given mobile device during a forensic investigation[6] and is sufficient for illustrating our methods.

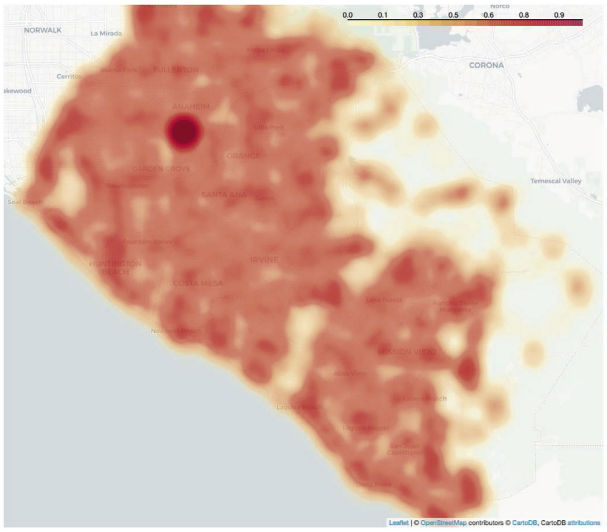
We consider two spatial regions: Orange County, California, and the Manhattan borough of New York City. The data was collected from May 2015 to February 2016, selecting only events (tweets) with GPS coordinates from public accounts. Each event is composed of tuples of the following form:

`<account_id, longitude, latitude, timestamp>`.

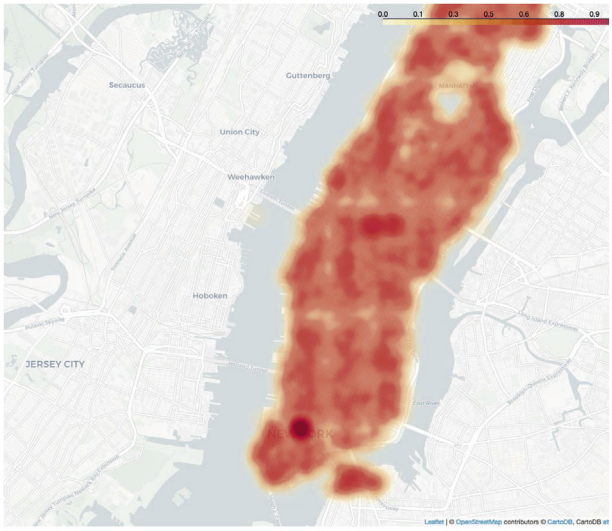
Thus, for any given account we have a set of geolocated events occurring in some bounded region. See Fig. 5 for the background event rates in both spatial regions.

As the focus of our analysis is on the unique locations a user visits (and not his or her rate of events at those locations), we define a *visit* as a set of events occurring within the same hour and within 50 meters of each other and treat the visit as a single effective event e (the first event from each visit is kept). Table 2 provides summary statistics for the Twitter data before and after filtering for visits. The visit data in this table is referred to as the population data, and was used for constructing the reference population E discussed in Section 6.

To generate the spatial event data for our experiments we filtered the data based on sequential time periods of activity. Users with at least 1 visit per month in each of the first 2 months were considered. For a given user, we define the sets of locations A and B to be all geolocated events in the first or second month, respectively. Table 3 contains summary statistics for the Twitter data used



(a) Orange County



(b) New York

Fig. 5. Population distribution of Twitter events used in the experimental results in this paper. (a) In Orange County, CA, note that the area of high density in Anaheim is the Disneyland Resort. (b) New York, NY.

[4] The score-based likelihood ratio may be used interchangeably with the LR in this section.

[5] Note that while the data is publicly available via Twitter's API (https://developer.twitter.com/en/docs/tweets/filter-realtime/overview), the terms of use require that collected data sets cannot be shared amongst researchers.

[6] We make the simplifying assumption that all Twitter events for a given account occur on the same device.

Table 1
Association of Forensic Science Providers (2009) verbal scale for presenting conclusions from the LR (or SLR).

LR Value	Verbal Expression
1–10	Weak or limited support
10–100	Moderate support
100-1000	Moderately strong support
1000–10,000	Strong support
10,000–100,000	Very strong support
$> 100,000$	Extremely strong support

Table 2
Number of observed days, accounts, events and visits for the Twitter data sets. Average number per account denoted in parentheses.

	Days	Accounts	Events	Visits
OC	240	103,271	655,917 (6.4)	545,697 (5.3)
NY	239	194,224	1,162,871 (6.0)	989,494 (5.1)

in the analysis.

10. Results

For both the Orange County and New York regions, we compared the likelihood-ratio and the score-based likelihood-ratio techniques in terms of their effectiveness in quantifying the strength of evidence for pairs of sets of locations A and B. For computational efficiency we included all same-source pairs (6,714 in OC and 13,523 in NY) and a stratified random sample of different-source pairs. The sampling was stratified by the number of visits in each pattern, n_a and n_b, because the data is highly skewed towards a small number of visit events per individual and we wanted to assess performance of the methods under varying amounts of data. The strata correspond to all $3 \times 3 = 9$ combinations of 1 visit, between 2 and 19 visits, and 20 or more visits for n_a and n_b. 1,000 different-source pairs in each strata were randomly sampled, resulting in 9,000 total different-source pairs in each region.

For the likelihood ratio approach, two choices for the mixing parameter α were used. The first was a constant $\alpha = 0.80$ for all pairs, and the second was a function of the number of visits in A, $\alpha = f(n_a)$, defined by the following

$$f(n_a) = \begin{cases} 0.05, & \text{for } n_a \leq 5 \\ 0.15, & \text{for } n_a \in (5, 10] \\ 0.40, & \text{for } n_a \in (10, 20] \\ 0.55, & \text{for } n_a \in (20, 50] \\ 0.70, & \text{for } n_a \in (50, 100] \\ 0.85, & \text{for } n_a > 100. \end{cases} \tag{9}$$

Alternative choices are also possible for the function defining the mixing parameter. The score-based likelihood ratio approach was estimated for both the mean inter-event distance and earth mover's distance score functions under all weighting strategies discussed in Appendix D.

Table 3
Number of observed accounts and visits for the Twitter data sets used in the analysis. Average number per account denoted in parentheses.

Region	Accounts	Visits in A	Visits in B
OC	6,714	44,310 (6.6)	38,697 (5.8)
NY	13,523	72,799 (5.4)	65,852 (4.9)

10.1. Motivating example

We begin the exploration of the results by re-visiting the motivating example in Fig. 1 of Section 2. Recall that the investigator was given one set of locations from an unknown source, A, as well as sets of locations from two known sources, B_1 and B_2. She was tasked with assessing the probative value of each pair of evidence—(A, B_1) and (A, B_2)—in order to determine the likelihood that either pair was generated by the same source. Using the likelihood ratio approach with fixed mixing weights, the LR for (A, B_1) was approximately 1137. Following the verbal equivalents provided in Section 8, the investigator would conclude that there is strong support that A and B_1 were generated by the same individual. For the second pair, (A, B_2), the LR was approximately 2.8e-28 which would lead the investigator to conclude that the individual that generated B_2 could be excluded as the source of A.

10.2. Overall results

The resulting LR and SLR values were thresholded to obtain binary decisions of same- or different-source, and these binary decisions were compared to the known ground truth to compute true and false positive rates. We then varied the threshold to achieve different trade-offs in terms of sensitivity and specificity. The area under the receiver operating characteristic (ROC) curve, abbreviated as AUC, can be used to summarize this trade-off. AUC is a measure of goodness of fit and can be thought of as the probability that the method will result in a larger LR or SLR for a randomly chosen same-source pair than for a randomly chosen different-source pair (e.g., Fawcett, 2006; Krzanowski and Hand, 2009). Higher AUC values are indicative of better detection performance.

Using likelihood ratios with a threshold of 1, corresponding to the data being equally likely to have been generated under either hypothesis, we classify pairs with LR greater than 1 as same-source and those with LR less than 1 as different-source. We can then compare the true and false positive rates for each choice of α. Table 4 provides these rates (listed as TP@1 and FP@1, respectively) along with the AUC. In both spatial regions the LR had similar performance, with the highest true positive rate and AUC belonging to the varying mixing weight approach and the lowest false positive rate for fixed α.

Similarly, using SLRs with a threshold of 1 we can compare the true and false positive rates for each score function. Table 5 provides these rates (listed as TP@1 and FP@1, respectively) along with the AUC. In both spatial regions, the SLR built on the EMD score function tends to outperform that using $\overline{D}_{min}$ within a given weighting scheme across TP, FP and AUC. Uniform weights tend to out-perform both the account and visit weighting schemes in terms of TP and AUC, but not FP. In Orange County account weights yield the lowest FP rate, while in NY both account and visit weights yield similarly low FP rates within a given score function.

Regardless of the region considered and choice of α, Δ and weighting scheme used, the likelihood ratio approach outperforms

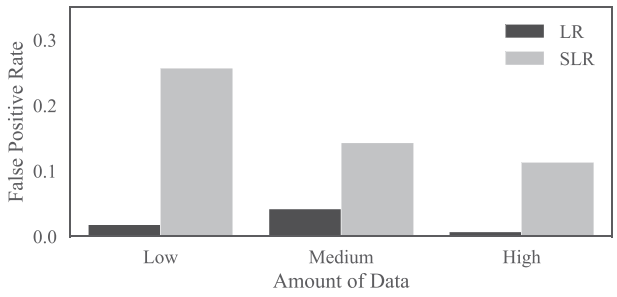
Table 4
Performance of a classifier based on *LR*.

Region	α	TP@1	FP@1	AUC
OC	0.80	0.340	**0.026**	0.787
	$f(n_a)$	**0.380**	0.038	**0.845**
NY	0.80	0.251	**0.067**	0.712
	$f(n_a)$	**0.285**	0.090	**0.768**

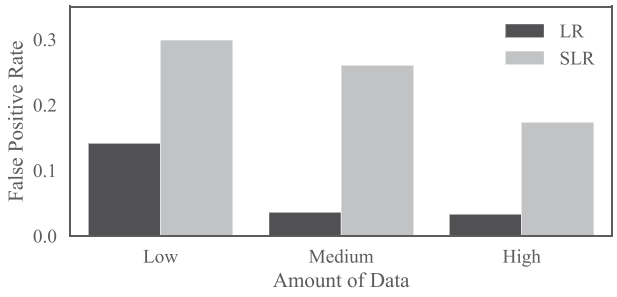
Table 5
Performance of a classifier based on SLR_{Δ}.

Region	Δ	Weights	TP@1	FP@1	AUC
OC	$\overline{D}_{min}$	Uniform	0.628	0.202	0.768
	$\overline{D}_{min}$	Account	0.610	0.171	0.774
	$\overline{D}_{min}$	Visit	0.611	0.180	0.768
	EMD	Uniform	**0.654**	0.197	**0.790**
	EMD	Account	0.614	**0.162**	0.783
	EMD	Visit	0.602	0.169	0.774
NY	$\overline{D}_{min}$	Uniform	0.508	0.287	0.656
	$\overline{D}_{min}$	Account	0.494	0.254	0.666
	$\overline{D}_{min}$	Visit	0.493	0.257	0.663
	EMD	Uniform	**0.530**	0.253	**0.686**
	EMD	Account	0.511	0.235	0.685
	EMD	Visit	0.504	**0.234**	0.679

the score-based likelihood ratio approach in terms of AUC and false positive rate. While the SLR has a larger true positive rate than the LR, the cost is a FP rate that is typically an order of magnitude larger. This phenomenon not only appears in the overall results, but also when condsidering performance of the techniques within the strata. Fig. 6 depicts the FP rate of the two approaches versus the amount of data in the sets *A* and *B* (corresponding to a selection of 3 of the 9 strata used in sampling) for both spatial regions. For both approaches, as the amount of data increases the false positive rate decreases. The SLR has much higher FP rate than LR across all data regimes.



(a) Orange County



(b) New York

Fig. 6. False positive rate of each method under different data regimes in (a) Orange County, and (b) New York. Low corresponds to 1 event in each of *A* and *B*, medium is between 2 and 19 events, and high is 20 or more events. Showing results for fixed α in the LR approach and the account weighted EMD for the SLR approach each thresholded at 1. Trends are similar for other score functions and threshold choices.

11. Discussion

It is worth noting that the manner in which we defined the sets *A* and *B* for the Twitter data (via time) is just one approach and the techniques we propose are not dependent on how the events in *A* and *B* are defined. For example, other ways of defining the sets of locations could include events from two different devices (e.g., mobile phones) collected over the same time period where an investigator is interested if they are associated with the same individual.

For the datasets investigated, we found that the methods showed promise in terms of being able to separate same-source pairs of spatial patterns from different-source pairs. This observation leads us to believe that these methods could be useful for discovery, e.g., as a method to rank the similarity of multiple different sets of locations from known sources to a single set of locations from an unknown source (similar to the motivating example in Section 2).

There are two main areas that impact the behavior of the techniques: the characteristics of the spatial region under consideration and amount of evidential data available.

11.1. Region characteristics

The spatial regions considered in this paper have very different characteristics. Orange County is largely suburban, while New York is the most densely populated city in the United States. As a result, the characteristics of the locations and how they are used tend to be quite different in each of these regions. In Orange County land parcels are typically single-use with one business or home at each location. However, in New York the parcels are mostly high rise buildings that contain many residences and businesses. We found that the different characteristics of the spatial regions manifest in different performance of the LR and SLR. In general the classification problem for Orange County is easier than it is for New York. The AUC illustrates this phenomenon, with each method having a larger AUC in OC than NY. This suggests that an analyst may need to take into account his or her knowledge of the region under consideration when presenting error rates of the method.

11.2. Amount of evidential data

Varying the number of events in *A* and *B* can significantly impact the behavior of our approaches. The score-based methods tend to be sensitive to the amount of evidential data available because the variance of the underlying score functions is high when the number of events is low. The high variance in the score function would be expected under both the same- and different-source distributions, making them more similar and generally leading to smaller SLR values for same-source pairs and larger values for different-source pairs. There is no natural way to alter behavior of the score functions when the number of observations is low. The LR approach is less sensitive to the amount of data, which makes intuitive sense as the likelihoods in both the numerator and denominator have no explicit reliance on the number of observed events.

12. Conclusion

Analysis of user-generated spatial event data is likely to become increasingly important in the forensic investigation of digital evidence. However, few methods have been developed to date that use statistical techniques for analysis of such data. In this paper we have

taken a step towards the development of such techniques, focusing on the problem of investigating whether two sets of user-generated geolocated events were generated by the same source or by different sources. Given a reference population, we proposed two approaches to quantify the strength of evidence in this setting. The first is a likelihood ratio approach based on modeling the location data directly. The second is to instead measure the similarity of the two sets of locations via a score function and then assess the strength of the score resulting in the score-based likelihood ratio. Experimental results, based on analysis of Twitter data in two spatial regions, indicate that the proposed methodology provides a useful starting point for forensic investigation of geolocated event data.

Acknowledgements

This research was partially funded through Cooperative Agreement #70NANB15H176 between the National Institute of Standards and Technology and Iowa State University, which includes activities carried out at Carnegie Mellon University, University of California, Irvine, and University of Virginia.

Appendix A. Kernel Density Estimation

In general, we follow the notation of Lichman and Smyth (2014) for our definition of kernel densities and mixtures of kernel densities. Assume that we are given a set of 2-dimensional points $e = (x, y)$ that represent the location of an event, denoted $E = \{e_i : i = 1, \ldots, n\}$. Kernel density estimation (KDE) is a common choice for the non-parametric estimation of a bivariate probability density function f using this data. Given the bivariate Gaussian kernel function K and a bandwidth parameter h, we get the following bivariate KDE

$$f_{KD}\left(e|E\right) = \frac{1}{n}\sum_{i=1}^{n} K(e, e_i|h)\Bigg) \tag{A.1}$$

$$K(e, e_i|h) = \frac{1}{2\pi h}\exp\Bigg(-\frac{1}{2}(e - e_i)^T \Sigma_h^{-1}(e - e_i)\Bigg) \tag{A.2}$$

$$\Sigma_h = \begin{pmatrix} h & 0 \\ 0 & h \end{pmatrix} \tag{A.3}$$

Thus the estimated density at e is the average of the kernels centered at the observations e_i and scaled by h across all n observations. KDEs are essentially a local smoothing method.

The choice of the kernel itself is not as important as the selection of the bandwidth h. As h decreases, the height of the peak at each observation increases resulting in undersmoothing. As h increases, the height of the peak at each observation decreases and probability mass is pushed away from the observation resulting in oversmoothing. Geolocated event data is hard to model via a homogeneous bandwidth given the high density of events in urban areas and low density in sparsely populated areas. More appropriate for this data is an adaptive bandwidth method where h is replaced with a bandwidth that depends on the observation e_i

$$f_{KD}\left(e|E\right) = \frac{1}{n}\sum_{i=1}^{n} K(e, e_i|h = h(e_i))\Bigg). \tag{A.4}$$

Lichman and Smyth (2014) showed that using an adaptive bandwidth $h(e_i)$ determined from the geodesic distance from e_i to its 5th nearest neighbor works well for modeling geolocated Twitter data, so the KDE estimates in the LR use these values. The minimum bandwidth was set to 50 meters to prevent issues with points occurring at the exact same location.

Appendix B. Score Functions for Geolocation Data

To define the score functions, we first construct an inter-event distance matrix by measuring the geodesic distance (Karney, 2013) from each event in B to each event in A. Let $\boldsymbol{D} = [d_{jk}]$ represent the $n_b \times n_a$ distance matrix where each element $d_{jk} = d(e_j^b, e_k^a)$ denotes the geodesic distance between the position of the jth event in set B and the position of the kth event in set A.

Appendix B.1. Nearest Neighbor Distances

Treating each point in set B as the focus, we can compute the inter-event distance to its nearest neighbor in A. Let D_{min} represent the collection of the n_b nearest neighbor distances from B to A, and define it as follows

$$\begin{aligned} &D_{min} \equiv \left\{d_j^{min} : j = 1, \ldots, n_b\right\} \\ &\text{where } d_j^{min} = \min_{k \in \{1,\ldots,n_a\}} d_{jk} \end{aligned} \tag{B.1}$$

If events of type B are spatially clustered among events of type A, then the nearest neighbor distances D_{min} tend to be smaller than if A and B events are generated independently and do not cluster together. A variety of characteristics of the distribution of nearest neighbor distances can be used as score functions $\Delta(A, B)$. In this paper we consider variants of the weighted arithmetic average nearest neighbor distance from B to A, defined in general as

$$\overline{D}_{min}\left(B, A\middle|\Omega^b\right) = \frac{\sum_{j=1}^{n_b} \omega_j^b d_j^{min}}{\sum_{j=1}^{n_b} \omega_j^b} \tag{B.2}$$

where $\Omega^b = \{\omega_j^b : j = 1, \ldots, n_b\}$ are weights assigned to each of the events in B. A discussion of the motivation for using weights and definitions of the various weighting strategies used here are provided in Appendices C and D.

Note that it is also possible to define a nearest neighbor distance from A to B. That distance would compute the nearest neighbor for each event of type A and weight these according to weights Ω^a. The asymmetry of the nearest neighbor distance is one motivation for seeking an alternative.

Appendix B.2. Earth Mover's Distance

The *earth mover's distance (EMD)*, or Wasserstein metric, is a measure of the distance between two probability distributions. To gain an intuition for the EMD, consider the problem of having multiple piles of earth of different sizes spread over some region that you wish to move into a collection of holes of different volumes in that same region. The EMD measures the least amount of "work" it takes to fill the holes with earth, where a unit of work consists of transporting a unit of earth by a unit of ground distance. For the problem at hand, we can think of the piles of earth as one point pattern (B) and the holes as the other (A). EMD has been widely used as a general approach for measuring distances between two sets as a function of the distance between elements of the sets (e.g., Rubner et al., 1998; Cohen, 1999). We develop the use of EMD in the context of measuring the similarity of spatial point patterns.

Computing the EMD is based on a solution to the transportation problem (Hitchcock, 1941). The first step is to find a flow $\boldsymbol{F}' = [f'_{jk}]$,

where f'_{jk} is the flow (or amount of mass) moved from e^b_j to e^a_k, that minimizes the overall cost

$$\boldsymbol{F}' = \arg\min_{[f_{jk}]} f_{jk} \sum_{j=1}^{n_b} \sum_{k=1}^{n_a} f_{jk} d_{jk} \tag{B.3}$$

subject to the following constraints

$$f_{jk} \geq 0 \quad j \in \{1, \ldots, n_b\}, k \in \{1, \ldots, n_a\} \tag{B.4}$$

$$\sum_{k=1}^{na} n_a f_{jk} \leq \omega^b_j \quad j \in \{1, \ldots, n_b\} \tag{B.5}$$

$$\sum_{j=1}^{nb} n_b f_{jk} \leq \omega^a_k \quad k \in \{1, \ldots, n_a\} \tag{B.6}$$

$$\sum_{j=1}^{nb} n_b \sum_{k=1}^{na} n_a f_{jk} = \min\left(\sum_{j=1}^{nb} n_b \omega^b_j, \sum_{k=1}^{na} n_a \omega^a_k \right). \tag{B.7}$$

where in principle the weights Ω^a and Ω^b are the same as those used in Equation B.2. The first constraint (B.4) restricts the flow of mass from B to A and not vice versa. The next two constraints (B.5, B.6) limit the amount of mass that can be sent from points in B to their weights, and the points in A receive no more mass than their corresponding weights. The last constraint (B.7) ensures the total amount of mass moved is equal to that of the lighter distribution, and is referred to as the total flow. Given the solution $\boldsymbol{F}'$ that minimizes (B.3), define the score function $\Delta(A, B)$ based on the earth mover's distance as the cost normalized by the total flow

$$EMD(B, A|\Omega) = \frac{\sum_{j=1}^{n_b} \sum_{k=1}^{n_a} f'_{jk} d_{jk}}{\sum_{j=1}^{n_b} \sum_{k=1}^{n_a} f'_{jk}} \tag{B.8}$$

where $\Omega = \{\Omega^a, \Omega^b\}$.

Note that the earth mover's distance is a metric when the distance between the points is a metric and the total weights of the point patterns are equal. Since geodesic distance is a metric, the first property is satisfied. We enforce that the weights sum to 1 for both sets A and B. Therefore, the earth mover's distance considered in this paper is a metric which implies that $EMD(B, A|\Omega) = EMD(A, B|\Omega)$. This simplifies computation and results in the same conclusions being drawn regardless of which pattern you consider as the focus of analysis.

Appendix C. Geoparcel Data

Geolocated event data is quite useful, but additional information can be incorporated if we also consider spatial properties of locations at which the events occur. High-traffic locations like shopping malls, theme parks and stadiums will have a high likelihood of appearing in any randomly selected point pattern and thus make patterns generated by different individuals look alike. Conversely, less common locations such as homes are highly unlikely to appear in multiple point patterns unless those patterns were generated by the same individual or someone close to him or her.

One option for incorporating spatial information is to partition the spatial region into a regular grid of disjoint cells, and compute population frequencies of events in each grid cell. However, defining the grid is a difficult problem as the result can be highly arbitrary since locations very rarely fall perfectly into a grid. Further, the spatial resolution of the grid is proportional to the amount of events in each cell—too small of a grid size results in highly sparse data. Given these limitations, we chose to use geoparcel information. Geoparcels are disjoint polygons (or parcels) that partition a spatial region where each individual parcel represents a specific property. The parcels vary in size and shape depending on the function of the property, solving the issues posed by using a grid. Within each parcel, we can measure the rarity of visits to that particular location. See Figure C.1 for an example of a parcel and a comparison to a grid-based approach.

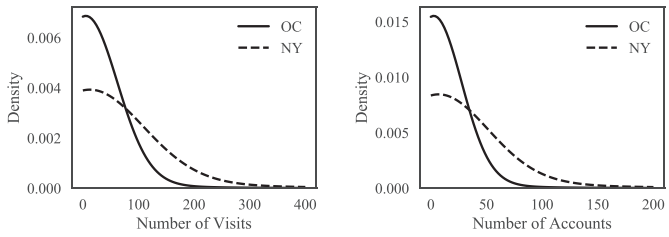
Figure C.1. Area around John Wayne Airport (SNA) in Orange County, California, highlighting the parcel corresponding to the airport and Twitter events in the region. Figure credit Lichman (2017).

We use the same publicly available geoparcel data as Kotzias et al. (2018). The 32,978 parcels for Orange County were collected from the Southern California Association of Government website.[7] The 21,312 parcels for New York were collected via the OpenStreetMap API.[8] Both the OC and NY data sets exhibit long-tailed distributions for the number of visits and number of unique accounts with at least one visit in each parcel, as shown in Table C.1 and Figure C.2. On average, parcels in New York have more visits and unique accounts than parcels in Orange County.

Table C.1
Summary statistics for the distribution of number of visits (Type "Visits") and unique accounts with at least one visit (Type "Accounts") in each parcel computed from the full population data in Table 2. The minimum and 25th percentile are 1 for all cases.

Region	Type	Mean	Med.	75th%ile	Max
OC	Visits	16.5	2	5	72,290
OC	Accounts	7.9	1	2	30,874
NY	Visits	46.4	4	16	77,760
NY	Accounts	26.8	3	10	25,775

[7] https://www.scag.ca.gov/.
[8] https://wiki.openstreetmap.org/wiki/API.



(a) Visits (b) Accounts

Fig. C.2. Density estimate of the number of parcels versus (a) the number of visits in the parcel, and (b) the number of unique accounts in that parcel. Note that both figures are right-truncated due to the extremely long tails.

Appendix D. Weighting Events

In our definitions of score functions $\Delta(A,B)$ for spatial point patterns in Equations B.2 and B.8, we require weights for each event. We consider three different weighting schemes that rely upon the geoparcel data discussed in the previous section. The weights are defined for events in point pattern B, but similar definitions hold for events in A. All weights are normalized for each point pattern, i.e., $\sum_{j=1}^{n_b}\omega_j^b = 1$.

1. *Uniform.* Let $\omega_j^b = n_b^{-1}$ for $j = 1,\ldots,n_b$. Under uniform weighting, Equation B.2 simplifies to the unweighted mean nearest neighbor distance. Furthermore uniform weights result in the empirical distribution for each point pattern being used as the relevant distribution in the earth mover's distance calculation.
2. *Location Visits.* Define the weight for each event as a function of the number of visits occurring at the location (geoparcel) of that event across the reference population. Namely,

$$\omega_j^b \propto \left[n_{vis}\left(\ell\left(e_j^b\right)\right)\right]^{-1} \tag{D.1}$$

where $n_{vis}(\ell)$ is the number of visits at location ℓ, in this case the geoparcel in which the jth event in B occurred.

3. *Location Accounts.* Define the weight for each event as a function of the number of unique accounts in the reference population with at least one visit at the location of that event. Namely,

$$\omega_j^b \propto \left[n_{acc}\left(\ell\left(e_j^b\right)\right)\right]^{-1} \tag{D.2}$$

where $n_{acc}(\ell)$ is the number of unique accounts with at least one visit at location ℓ, in this case the geoparcel in which the jth event in B occurred.

The uniform weighting scheme is the most naive method, and requires no geoparcel data. Both location weighting schemes attempt to solve what we refer to as the "Disneyland Problem." Specifically, in some spatial regions, a small subset of parcels can be responsible for a large fraction of the Twitter activity. At such locations, it is highly likely that any randomly-selected account will generate an event there. For Orange County, one of these parcels corresponds to the Disneyland Resort, as is evidenced in Fig. 5a. The location-based weighting schemes above down-weights events from such parcels, placing more weight on events at rarer locations such as homes.

Appendix E. Leave-pairs-out Cross Validation

The results in this paper use a slight variant of the set construction for $\mathscr{D}_s$ and $\mathscr{D}_d$ discussed in Section 7 because the point patterns of interest are elements of $\mathscr{D}$. To evaluate the out-of-sample performance of the techniques we use *leave-pairs-out cross-validation* to construct the reference data sets used to estimate the score-based likelihood ratio. Let $(A,B) = (A_\ell, B_m)$ be an arbitrary pair from $\mathscr{D}$, where ℓ and m may or may not be equal. Given (A_ℓ, B_m) let $\mathscr{D}_s = \{(A_j,B_j) : j \in \{1,\ldots,N\}\backslash\{\ell,m\}\}$ and $\mathscr{D}_d = \{(A_j,B_k) : j,k \in \{1,\ldots,N\}\backslash\{\ell,m\}, j \neq k\}$ be the sets used in the results of Section 10. Essentially, we remove any pair with a point pattern from either account currently being evaluated.

References

Aitken, C., Taroni, F., 2004. Statistics and the Evaluation of Evidence for Forensic Scientists, second ed. John Wiley & Sons.

Årnes, A., 2017. Digital Forensics. John Wiley & Sons.

Association of Forensic Science Providers, 2009. Standards for the formulation of evaluative forensic science expert opinion. Sci. Justice 49, 161–164.

Berman, M., 1986. Testing for spatial association between a point process and another stochastic process. J. Roy. Stat. Soc.: Series C (Applied Statistics) 35 (1), 54–62.

Bolck, A., Ni, H., Lopatka, M., 2015. Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison. Law Probab. Risk 14 (3), 243–266.

Bosma, W., Dalm, S., van Eijk, E., el Harchaoui, R., Rijgersberg, E., Tops, H.T., Veenstra, A., Ypma, R., 2019. Establishing phone-pair co-usage by comparing mobility patterns. Sci. Justice.

Casey, E., 2018. Clearly conveying digital forensic results. Digit. Invest. 24, 1–3.

Casey, E., Jaquet-Chiffelle, D.-O., Spichiger, H., Ryser, E., Souvignet, T., 2020. Structuring the evaluation of location-related mobile device evidence. Forensic Sci. Int.: Digit. Invest.

Champod, C., Evett, I.W., 2001. A probabilistic approach to fingerprint evidence. J. Forensic Ident. 51 (2), 101–122.

Cohen, S., 1999. Finding Color and Shape Patterns in Images. Stanford University, Department of Computer Science. No. 1620.

Evett, I.W., Weir, B.S., 1998. Presenting evidence. In: Fagerberg, J., Mowery, D.C., Nelson, R.R. (Eds.), Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists. Sinauer, pp. 235–266. Ch. 9.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27 (8), 861–874.

Galbraith, C., Smyth, P., 2017. Analyzing user-event data using score-based likelihood ratios with marked point processes. Digit. Invest. 22, S106–S114.

Galbraith, C., Smyth, P., Stern, H.S., 2020. Quantifying the association between discrete event time series with applications to digital forensics. J. Roy. Stat. Soc. A (in press).

Haggett, P., Cliff, A.D., Frey, A., 1977. Locational analysis in human geography. Tijdschr. Econ. Soc. Geogr. 68 (6).

Hitchcock, F.L., 1941. The distribution of a product from several sources to numerous localities. J. Math. Phys. 20 (1-4), 224–230.

Karney, C.F.F., Jan 2013. Algorithms for geodesics. J. Geodes. 87 (1), 43–55.

Kotzias, D., Lichman, M., Smyth, P., 2018. Predicting consumption patterns with repeated and novel events. IEEE Trans. Knowl. Data Eng. 31 (2), 371–384.

Krzanowski, W.J., Hand, D.J., 2009. ROC Curves for Continuous Data. Chapman and Hall/CRC.

Lichman, M., 2017. Context-based Smoothing for Personalized Prediction Models. Ph.D. thesis, UC Irvine.

Lichman, M., Smyth, P., 2014. Modeling human location data with mixtures of kernel densities. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 35–44.

Meuwly, D., Ramos, D., Haraksim, R., 2017. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. Forensic Sci. Int. 276, 142–153.

Nordgaard, A., Rasmusson, B., 2012. The likelihood ratio as value of evidence—more than a question of numbers. Law Probab. Risk 11 (4), 303–315.

Ommen, D.M., Saunders, C.P., 2018. Building a unified statistical framework for the forensic identification of source problems. Law Probab. Risk 17 (2), 179–197.

Pollitt, M., Casey, E., Jaquet-Chiffelle, D.-O., Gladyshev, P., 2019. A Framework for Harmonizing Forensic Science Practices and Digital/multimedia Evidence. OSAC Task Group on Digital/Multimedia Science. OSAC Technical Series 0002R1, OSAC/NIST.

Roussev, V., 2016. Digital forensic science: issues, methods, and challenges. In: Synthesis Lectures on Information Security, Privacy, & Trust, 8, pp. 1–155, 5.

Rubner, Y., Tomasi, C., Guibas, L.J., 1998. A metric for distributions with applications to image databases. In: Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271). IEEE, pp. 59–66.

Schlather, M., Ribeiro Jr., P.J., Diggle, P.J., 2004. Detecting dependence between marks and locations of marked point processes. J. Roy. Stat. Soc. B 66 (1), 79–93.

Scott, D.W., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley.

Stern, H.S., 2017. Statistical issues in forensic science. Ann. Rev. Statist. Its Appl. 4 (1), 225–244.

Swgde, 2019. Best Practices for Mobile Device Evidence Collection & Preservation,

Handling, and Acquisition. Tech. rep., Scientific Working Group on Digital Evidence.
Thompson, W.C., 2017. How should forensic scientists present source conclusions? Seton Hall Law Rev. 48, 773.
Valentino-DeVries, J., 2019. Tracking Phones, Google Is a Dragnet for the Police. The New York Times. https://www.nytimes.com/interactive/2019/04/13/us/google-location-tracking-police.html.
Willis, S., McKenna, L., McDermott, S., O'Donell, G., Barrett, A., Rasmusson, B., Nordgaard, A., Berger, C., Sjerps, M., Molina, J.L., Zadora, G., Aitken, C., Lunt, L., Champod, C., Biedermann, A., T.N. Hicks, F.T., 2016. ENFSI Guideline for Evaluative Reporting in Forensic Science. European Network of Forensic Science Institutes.