INSIGHTS

Statistical Methods for the Forensic Analysis of Geolocated Event Data

Lead Researchers: Christopher Galbraith, Padhraic Smyth, and Hal S. Stern
Journal: Forensic Science International: Digital Investigation | **Publication Date:** July 2020
Link: forensicstats.link/GeolocatedEvent-DOI

OVERVIEW

Researchers investigated the application of statistical methods to forensic questions involving spatial event-based digital data. A motivating example involves assessing whether or not two sets of GPS locations corresponding to digital events were generated by the same source. The team established two approaches to quantify the strength of evidence concerning this question.

THE GOAL

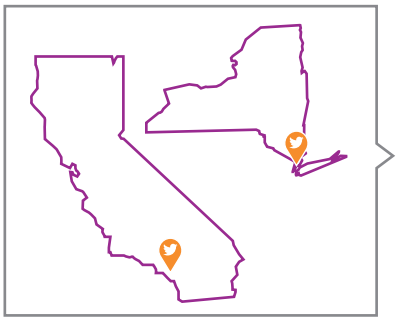
Develop quantitative techniques for the forensic analysis of geolocated event data.

APPROACH AND METHODOLOGY

Key Definitions:

Likelihood Ratio (LR): A comparison of the probability of observing a set of evidence measures under two different theories in order to assess relative support for the theories.

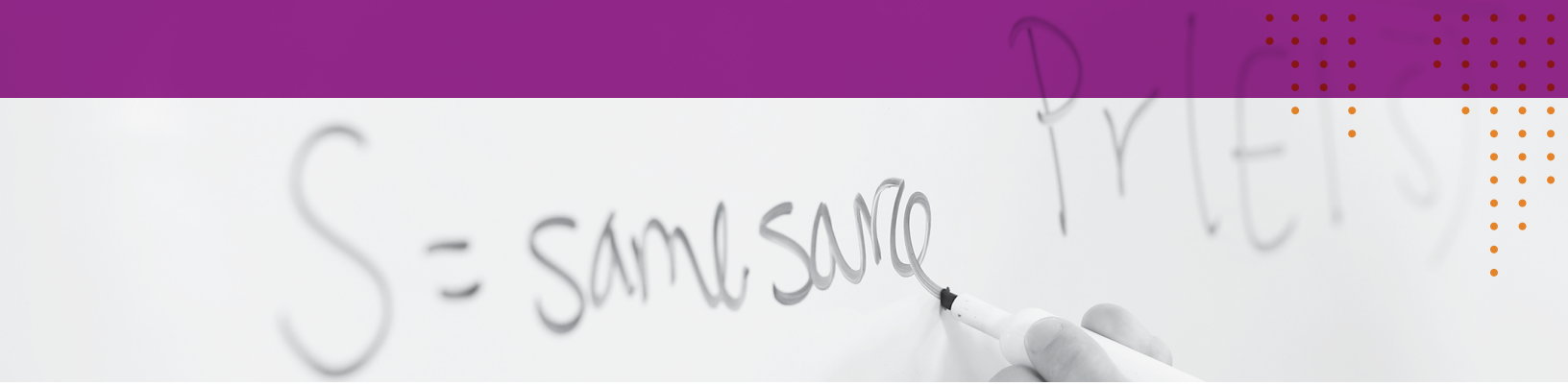
Score-Based Likelihood Ratio (SLR): An approach that summarizes evidence measures by a score function before applying the likelihood ratio approach.



Researchers collected geolocation data from Twitter messages over two spatial regions, Orange County, CA and the borough of Manhattan in New York City, from May 2015 to February 2016. Selecting only tweets from public accounts, they were able to gather GPS data regarding the frequency of geolocated events in each area.

CONTACT

Website: forensicstats.org **Email:** csafe@iastate.edu



This study considered a scenario in which two sets of tweet locations are relevant to then determine the source of the tweets. The tweets could be from different devices *or* from the same device during two different time periods.

The team used kernel density estimation to establish a likelihood ratio approach for observing the tweets under two competing hypotheses: are the tweets from the same source or a different source?

Utilizing this second approach creates a score-based likelihood ratio that summarizes the similarity of the two sets of locations while assessing the strength of the evidence.

Decisions based on both LR and SLR approaches were compared to known ground truth to determine true and false-positive rates.

KEY TAKEAWAYS FOR PRACTITIONERS

1. Both methods show promise in being able to distinguish same-source pairs of spatial event data from different-source pairs.
2. The LR approach outperformed the SLR approach for all dataset sizes considered.
3. The behavior of both approaches can be impacted by the characteristics of the observed region and amount of evidential data available.

FOCUS ON THE FUTURE

- In this study, *time* defined sets of locations gathered from Twitter. But, other methods for defining sets of locations, for example, including *multiple devices over the same time period*, could yield different results.
- The amount of available data (the number of tweets) impacts the score-based approach.

LEARN MORE

Access the full study to learn more at forensicstats.link/GeolocatedEvent.

FUNDING

csafe
Center for Statistics and Applications in Forensic Evidence

CSAFE is a publicly funded organization headquartered at Iowa State University. The National Institute of Standards and Technology (NIST) is one of the center's providers, supporting CSAFE as a nationally recognized Center of Excellence in Forensic Sciences, NIST Award #70NANB15H176 and #70NANB20H019.