A Bayesian Approach to Modeling GPS Errors for Comparing Forensic Evidence

Nika Molan
nm83087@student.uni-lj.si
University of Ljubljana,
Faculty of Computer and
Information Science,
Ljubljana, Slovenia

Ema Leila Grošelj
eg61487@student.uni-lj.si
University of Ljubljana,
Faculty of Computer and
Information Science,
Ljubljana, Slovenia

Klemen Vovk
kv4582@student.uni-lj.si
University of Ljubljana,
Faculty of Computer and
Information Science,
Ljubljana, Slovenia

ABSTRACT

This paper introduces a Bayesian approach to modeling GPS errors for comparing forensic evidence, addressing the challenge of determining the most likely source of a single GPS localization given two proposed locations. We develop a probabilistic model that transforms GPS coordinates into polar coordinates, capturing distance and directional errors. Our method employs Markov chain Monte Carlo (MCMC) sampling to estimate the data-generating processes of GPS measurements, enabling robust comparison of potential device locations while quantifying uncertainty. We apply this approach to three datasets: one from existing literature and two newly collected datasets from Ljubljana and Novo mesto. The result is a posterior distribution of log-likelihood ratios directly comparing the two propositions, which can be transformed into likelihood ratios to comply with current standards in forensic science.

KEYWORDS

device geolocation as evidence, MCMC, digital forensics, likelihood ratio, Stan, Bayesian inference

1 INTRODUCTION

GPS as evidence has been proven problematic in court, as it has often been dismissed or not presented at all [1] due to the fear of wrong judgment or discrediting other evidence due to the uncertainty of GPS measurements. Our goal is to provide a Bayesian approach for evaluating a single GPS localization in light of two proposed locations. To give more context to why such approaches are needed, consider the following problem: a single GPS localization (evidence point *E*) was extracted from a device D found on the crime scene. Since E is a GPS measurement there is inherently some measurement error. Additionally, someone could have moved the device D during or after the crime. Investigators propose two geographical locations (P1 and P2) of where the device could have been when it measured E and we want to identify which of the two propositions is more likely. Since the conclusion is to be presented in court, we need to provide sufficiently precise verbal equivalents of the results while not misleading or misrepresenting the weight of a piece of evidence.

The contributions of this paper are summarized as follows:

- a Bayesian statistics approach utilizing Markov chain Monte Carlo sampling to estimate the data generating processes (DGPs) of GPS measurements taken from different locations to compare which DGP most likely generated a single GPS measurement obtained as forensic evidence,
- Code implementation of the proposed approach along with MCMC diagnostics, results, and visualizations made available at [2],
- two GPS measurement datasets collected in Ljubljana and Novo mesto to aid further research.

2 RELATED WORK

The increased availability of GPS logs from smartphones, activity trackers, navigation, and autonomous vehicles has increased the use of such digital evidence in court[1]. Due to the high risk of misinterpretation and wrongful judgment or discrediting other evidence, statistical methods have been proposed [3] to be able to quantify and reason under uncertainty due to GPS measurement errors.

The magnitude of GPS errors varies between devices and geographical locations. In [8] the authors report a localization error of up to 5 meters in low-cost phones, while others (see [5]) have measurement errors varying up to 100 meters.

The standard in forensic science for evidence source identification is to use likelihood ratios [7]. As different magnitudes of likelihood ratios are not easily explainable in court, forensics standards have been developed to define the verbal equivalent of likelihood ratio ranges to provide in court[4], with the current version of the standard shown in Table 1.

In [5] the authors computed a likelihood ratio to compare two proposed device locations (P1 and P2) in light of the evidence E. They also considered that errors of GPS measurements may not be equal in all directions (the horizontal error is dependent on the direction) resulting in high computational complexity due to sample dependence and brute force distribution fitting.

3 DATASETS

For all used datasets we provide data sources as well as scripts to transform data from other works to the input format for our approach. The scripts used to transform and clean the datasets that were used as inputs for modeling are also provided.

The *University of Lausanne dataset (UNIL)* was obtained from the public implementation of [5]. It consists of 699 GPS measurements that were taken from two proposal points as reference measurements on the University of Lausanne campus. The dataset is visualized in Figure 1.

Our *Ljubljana (LJ)* dataset consists of 4 predefined points (evidence *E*, and three proposal points *P*1, *P*2, and *P*3, each point is specified by latitude and longitude) and a total of 450 images captured while standing on those proposal points along with information

University of Maribor Press
DOI: https://doi.org/10.18690/um.feri.6.2024.10
ISBN 978-961-286-914-4

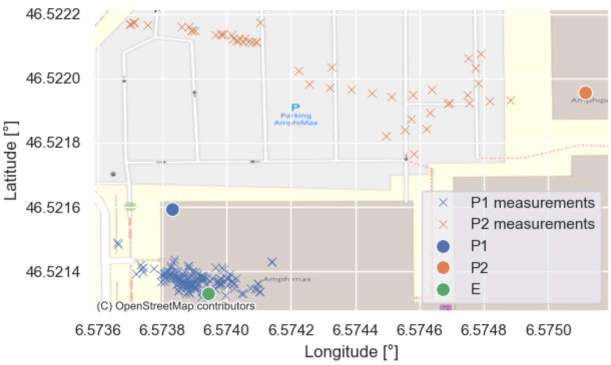

Figure 1: A geographical visualization of the UNIL dataset from [5]. E is the single localization that was recovered, while P1 and P2 are two proposed locations. Only deduplicated reference measurements for both proposals are shown. Note how reference measurements are not even on the same building as the proposal they were measured from.

if the iPhone camera app had permission for precise location for every image. A visualization of the dataset is shown in Figure 2.

Our *Novo mesto (NM)* dataset consists of 4 predefined points (evidence E, and three proposal points $P1$, $P2$, and $P3$, each point is specified by latitude and longitude) and a total of 429 images captured using the same data collection process as the LJ dataset. A visualization of the dataset is shown in Figure 3.

Compared to the UNIL dataset, LJ and NM datasets have significantly less measurement error (on average 15 meters) and a high percentage (90%) of duplicate measurements. This is due to the measurements being done in a very short time interval (30 minutes) which resulted in a lot of cached duplicates.

4 METHODS

Unless specified, everything in this section applies to all used datasets (UNIL, LJ, and NM).

4.1 LJ and NM dataset collection

To more clearly understand what affects the accuracy of GPS evidence as well as error patterns, we collect two additional datasets with the same device. The distance between the predefined points was around 100 meters in Ljubljana and 10 meters in Novo mesto. To take images an iPhone 11 Pro was used with the default camera app. We also note granting and denying permission to precise locations to the camera app for each image. To obtain GPS measurements from images, we extract the latitude, longitude, and time of capture from the EXIF data of each image, and note the corresponding proposal point it was taken from and if precise location permission was granted.

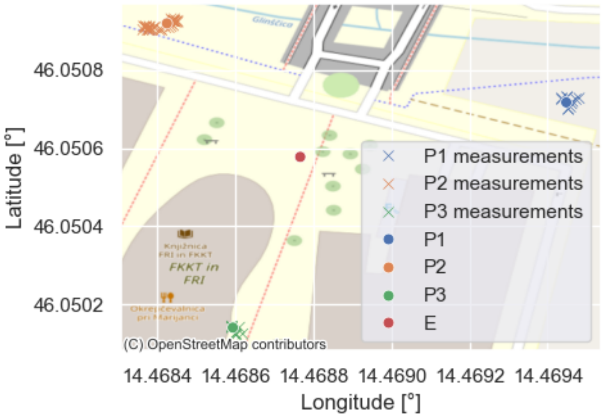

Figure 2: A geographical visualization of our LJ dataset. Only deduplicated measurements are shown.

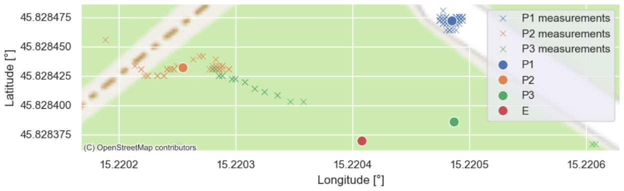

Figure 3: A geographical visualization of our NM dataset. Only deduplicated measurements are shown.

4.2 Dataset preprocessing

Each measurement is defined by time, latitude, longitude, and the label of the proposal it was taken from. For LJ and NM data we keep only measurements that were retrieved when precise location permission was given to the iPhone camera app. We remove consecutive duplicate GPS measurements per proposal by sorting all corresponding measurements ascending by date and time, then removing all consecutive duplicates based on latitude and longitude. This is done because consecutive duplicates could be due to caching and/or rate-limiting to GPS queries. Consequently, if we try to model distance and angle errors of GPS measurements, some angles/distances will have artificially more probability mass due to duplicates, even though these duplicates are obtained from the same GPS measurement.

4.3 Transforming GPS to polar coordinates

To simplify the modeling and representation of GPS errors, each measurement was converted from (latitude, longitude) coordinates to distance (in meters) and angle (azimuth from the north, in radians) from the ground truth point (proposal) it was taken from. To illustrate the concept we show the UNIL dataset transformed to polar coordinates in Figure 4. We aim to model the distance and directionality of GPS errors taken from proposal points (and consequently their DGP) to estimate under which proposal point is the retrieved evidence point E more likely.

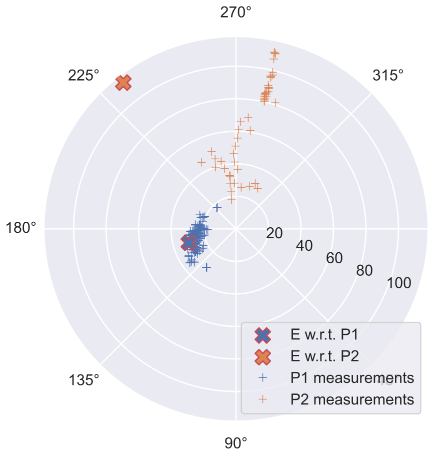

Figure 4: The UNIL dataset [5] after coordinate transformation (distance error in meters and angle - azimuth from north). Note, we only have one evidence point E, we transform it concerning (relative to, as seen from) each proposal point separately. Proposals P1 and P2 are at the center of the polar plot and measurements show the directionality (angle) and magnitude (in meters) of the GPS errors.

4.4 Probabilistic model of GPS errors

To formalize our approach to modeling GPS errors, we define a probabilistic model and estimate its parameters with MCMC sampling. The model is defined in the Stan probabilistic programming language[6] which allows users to define (log) density functions and then perform Bayesian inference with MCMC sampling.

Let $M_{ij} = (d_j, \phi_j)$ be the $i - th$ measurement measured from proposal P_j where d_j is the distance in meters from P_j to M_{ij} and ϕ_j the azimuth (in radians) of the line between P_j and M_{ij}. Analogously, let M_{ik} be the $i - th$ reference measurement measured from proposal P_k. The set of measurements for each proposal was modeled as a bivariate normal distribution:

$$\begin{aligned} M_{ij} &\sim \text{MultiNormal}(\mu_j, \Sigma_j) \\ M_{ik} &\sim \text{MultiNormal}(\mu_k, \Sigma_k) \\ \mu_j &= [\mu_{dist_j}, \mu_{angle_j}] \\ \mu_k &= [\mu_{dist_k}, \mu_{angle_k}] \\ \Sigma_j &\in \mathbb{R}^{2x2} \\ \Sigma_k &\in \mathbb{R}^{2x2} \end{aligned}$$

where μ_j is a mean vector of distance (in meters) and angle (in radians) relative to proposal P_j. Σ_j is the covariance matrix[1]. Analogously for μ_k and Σ_k relative to proposal P_k. Stan's default, non-informative priors were used for all parameters.

To compare under which proposal (P_j or P_k) is the evidence point E more likely, we compute the likelihood of E under each of the models and compute the likelihood ratio. To enhance numerical stability without loss of expressiveness the logarithms of likelihoods were used, which can later be exponentiated back. This was implemented in Stan's *generated quantities* block[2]:

$$\begin{aligned} \log L_j &= \log P(E_j|\mu_j, \Sigma_j) \\ \log L_k &= \log P(E_k|\mu_k, \Sigma_k) \\ \log LR &= \log L_j - \log L_k \end{aligned}$$

where E_j and E_k denote the evidence point E transformed relative to proposals P_j and P_k respectively.

To assess if the models capture the input data (reference measurements) well, a posterior predictive check was performed by randomly sampling points from the estimated bivariate normal models to create replicate datasets (this is also done in the *generated quantities* block in Stan). The idea is that if an estimated model fits input data well, we should be able to generate *similar*, synthetic data by randomly sampling from it. In other words, if the estimated model managed to capture the behavior of distance and angle errors in our reference measurements, it should be able to generate new, synthetic, measurements that resemble the same distance and angle errors. To visualize this, we overlay the generated synthetic data over the reference measurements (input data).

5 RESULTS

Due to the length limit of the paper we only show the full results for the UNIL dataset. However, all results, visualizations, and MCMC diagnostics are available in the provided repository.

MCMC sampling with 4 chains of 4000 samples each was performed to sample from the posterior. Standard MCMC diagnostics (trace plots, effective sample sizes, R-hat values) do not indicate any issues in convergence. Additionally, we visualize a posterior predictive check of the UNIL dataset by overlaying a random replicate dataset over the real measurements in Figure 5.

Figure 6 depicts the posterior distribution of log-likelihood ratios for the UNIL dataset along with 95% highest-density intervals. In

[1] We use the Cholesky parameterization of the Multivariate normal, which is natively implemented in Stan, so $\Sigma_j = L_j L_j'$ for efficiency and numerical stability during MCMC sampling, but omit it here for brevity

[2] Everything in the generated quantities block can be computed outside of Stan (e.g. in Python) as it is performed on the posterior draws after the MCMC sampling is done, we do it in Stan for clarity.

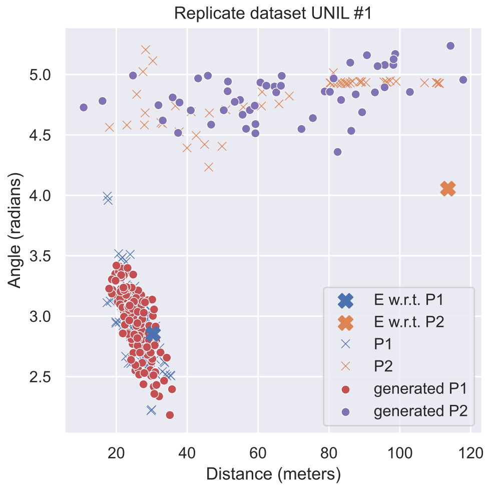

Figure 5: A posterior predictive check for the UNIL dataset in the form of a randomly selected replicate dataset for each proposal. All generated measurements for P1 and P2 are within expected regions, however, the model for P1 is better supported by the real measurements as the dataset heavily favors P1 as the source of E. This is also noted by the authors of the dataset [5].

line with the dataset, the P1 proposal is heavily favored compared to P2 to have generated the evidence point. Log-likelihood ratios can be converted back to likelihood ratios[3] which are currently the standard used in courts as per [4] and [5] to compare evidence for source-identification in forensic science.

While the model is stable and MCMC converges, even the lower-bound of the 95% highest density interval log-likelihood ratio is $\log LR = 11$, which after exponentiation is $LR = \exp 11 = 59874$, which is orders of magnitude above the highest obtainable LR range for a single piece of evidence (see Table 1 and the standard specification in [4]).

6 DISCUSSION

Our method, utilizing MCMC sampling to estimate data-generating processes of GPS measurements, offers direct uncertainty quantification, greater computational efficiency, and numerical stability due to Stan, MCMC, and working with log-likelihood ratios instead of likelihood ratios compared to the seminal method from [5]. The main limitation of our approach is the assumption that reference measurements taken (months) after the original evidence are enough to sufficiently model the DGP of GPS errors. Even if the exact device from the crime scene is used, many other variables are out of our control (GPS satellite visibility, noise and interference of GPS positioning, software updates changing the measuring process, cellular and WiFi networks that are used to improve location). Investigators should always strive to gather more actual evidence (i.e. more GPS logs from the crime scene) to directly model the errors instead of using a proxy such as reference measurements.

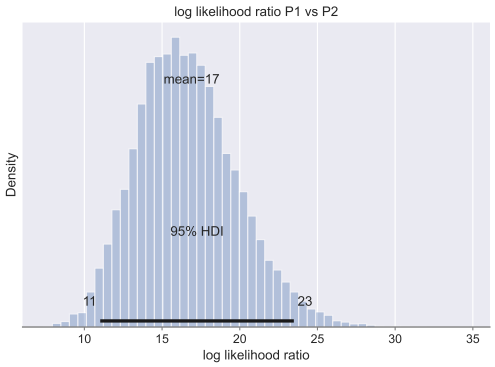

Figure 6: The posterior distribution of log-likelihood ratios for the UNIL dataset along with 95% highest-density intervals to quantify uncertainty.

Table 1: LR verbal equivalents to use in court when comparing two propositions, obtained from [4] page 39.

Range of LR	Verbal Equivalent
1-3	In my opinion the observations are no more probable if [P1] rather than[P2] were true. Therefore, the observations do not assist in addressing which of the two propositions is true.
4-10	In my opinion the observations are slightly more probable if [P1] rather than [P2] were true.
10-100	In my opinion the observations are more probable if [P1] rather than [P2] were true.
100-1000	In my opinion the observations are much more probable if [P1] rather than [P2] were true.

REFERENCES

[1] E. Casey, D.-O. Jaquet-Chiffelle, H. Spichiger, E. Ryser, and T. Souvignet. Structuring the evaluation of location-related mobile device evidence. *Forensic Science International: Digital Investigation*, 32:300928, 2020.

[2] Ema Leila Grošelj, Nika Molan and Klemen Vovk. A Bayesian Approach to Modeling GPS Errors for Comparing Forensic Evidence, 2024. https://github.com/KlemenVovk/gps_evaluation, Last accessed on 2024-08-30.

[3] C. Galbraith, P. Smyth, and H. S. Stern. Statistical methods for the forensic analysis of geolocated event data. *Forensic Science International: Digital Investigation*, 33:301009, 2020.

[4] F. S. Regulator. Development of evaluative opinions. Technical Report FSR-C-118, UK Forensic Science Regulator, Birmingham, 2021.

[5] H. Spichiger. A likelihood ratio approach for the evaluation of single point device locations. *Forensic Science International: Digital Investigation*, 2023.

[6] Stan Development Team. Stan modeling language users guide and reference manual, version 2.35, 2011–2024.

[7] M. M. Vink, M. M. Sjerps, A. A. Boztas, et al. Likelihood ratio method for the interpretation of iphone health app data in digital forensics. *Forensic Science International: Digital Investigation*, 41:301389, 2022.

[8] L. Wang, Z. Li, N. Wang, and Z. Wang. Real-time gnss precise point positioning for low-cost smart devices. *GPS Solutions*, 25:1–13, 2021.

[3] Highest-density intervals are not equal-tailed, this means that when applying transformations, such as exponentiation to the whole distribution, the HDI will change, For such cases we recommend computing equal-tailed credible intervals that are not affected by distribution transformations.