Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
April 28, 2022
A Regional Conference of the Society for Industrial and Systems Engineering

Geospatial Big Data Analytics for Quality Control of Surveys

Benjamin Leehan[1,2] and Nathaniel Bastian[1,3]

[1]Department of Mathematical Sciences
United States Military Academy
West Point, NY 10996

[2]Department of Electrical Engineering and Computer Science
United States Military Academy
West Point, NY 10996

[3]Army Cyber Institute
United States Military Academy
West Point, NY 10996

Corresponding author's Email: nathaniel.bastian@westpoint.edu

Author Note: CDT Benjamin Leehan is a fourth-year cadet at the United States Military Academy, where he double majors in Computer Science and Mathematical Sciences. His thesis advisor, MAJ(P) Nathaniel Bastian, supported this project. We would like to thank members of D3 Systems, particularly Mr. Matt Warshaw, Mr. Tim Van Blarcom, Mr. Jeffry Yan, and Mr. Connor Brassil, for generously sponsoring the project and providing the necessary data to conduct this research.

Abstract: Geospatial big data analytics allows survey quality control analysts to draw important conclusions about survey data quality that otherwise would take excessive time and resources. In this work, we explored two algorithmic methods that can help ensure reliability of survey interviews by detecting geospatial outliers. Focusing on geospatial data collected from surveys, we implemented outlier detection techniques with two different distance metrics to identify statistical anomalies in real-world datasets that may have qualitative interpretations. We found that one algorithm, which considers the local distribution of points in a dataset, identifies a different set of outliers when compared to another method, which considers the global distribution of points. Since there was a small overlap (10-19%) of flagged points between the two algorithms implemented, it may be helpful for analysts to focus on the fewer "outlier" points that are flagged by both methods rather than all the "outlier" points that are flagged by each algorithm. Finally, analysts should consider the computational costs, as the algorithms differ significantly.

Keywords: Outlier Detection, Geospatial Analytics, Survey Quality Control, Data Visualization, Big Data

1. Background

Traditional big data analytics of geospatial data often requires an analyst to rely on data visualizations to draw qualitative conclusions on a quantitative data set. However, this task becomes unfeasible as the data set grows larger and more nuanced. The ability to flag a subset of a large geospatial data set allows an analyst to narrow focus on smaller data sets identified to contain anomalous features. One particular field that benefits from such a capability is survey quality control, where human-collected surveys are examined to identify ill-collected data that might invalidate the survey. In this work, we explore and implement two outlier detection algorithms, with two different distance metrics, that can aid in the identification of geospatial anomalies on two real-world big data sets from surveys conducted and collected by humans on electronic devices.

1.1 Quality Control of Surveys

A survey seeks to capture targeted data of a population by collecting a sample that reflects a proportion of the population. A *quality* survey satisfies the following three characteristics (Brassil, 2021):

1) **integrity** – an interview or data collection occurred without fabrication
   (artificially crafted surveys, duplicate surveys, etc.)
2) **reliability** – the sampling happened with correct methodology
   (missing target population, biased sampling method, etc.)

3) **accuracy** – the respondent's true metrics/opinion were given, recorded and processed into the data set (human/machine error while reading responses, "straight-lining," contradicting responses, etc.)

After post-field collection of survey data, an analyst may be interested in verifying these characteristics to ensure the quality of the survey. Violation of any of the three characteristics may lead to survey results that do not reflect the true information regarding the population. Automation of survey quality control via statistical methods has the potential to help relieve analysts from heavy work. For example, D3 Systems created ValkyRie, which is an application that automates common survey quality control tests to catch instances of duplication, straight-lining, non-response, etc. (Brassil, 2021). Their analytic tool focuses on the survey responses themselves to catch violation of integrity and accuracy. However, it is difficult to measure reliability of a survey, as the survey itself does not show *how* the survey was performed. Analyzing the geographic locations of the executed surveys offers a promising way to help verify the reliability of surveys. Modern survey collection is often performed through software on electronic tablets, which records the time and location of the survey as metadata. An analyst can display the survey locations to visualize the surveyor behavior and detect violation of protocols. However, visualization becomes infeasible as the number of devices increase and surveys cover a larger geographic area. Currently, there is lack of geospatial big data analytic tools that help automate this survey quality control.

2. Related Works

There is rather limited existing literature on identifying location-based outliers in survey quality control perhaps due to the limited complexity of a purely geospatial data point. This is especially true on a geospatial data set that is limited to a two-dimensional surface. One location-based outlier that is of interest is an isolated point. Breunig, Kriegel, Ng, and Sander (2000) defined *outlier factor* that captures the degree of isolation compared to a surrounding clustering structure rather than compared to the global distribution. On the other hand, there have been numerous advances in identifying attribute-based outliers. Simple methods include performing univariate statistical analysis on an attribute among identified clusters of data. Mean or median-based analysis is highly influenced by the distribution of the attribute or existence of an extreme outlier. Singh and Lalitha (2017) introduced location quotient (LQ) as an alternative to mean and median-based analysis that may be more robust against extremities. Further, multivariate statistical analysis can help identify subtle outliers that are hard to identify when attributes are examined independently (Ben-Gal, 2005). In this work, we test the effectiveness and generalizability of some of these methods, attempting to interpret the outliers detected as possible violation of the surveys' reliability.

3. Methods

In this section, we briefly discuss survey data exploration and review two distance metrics and two algorithms that use geospatial data to perform outlier detection. The goal is to identify instances of the survey that display geographical anomalies within a relevant group and, thus, help identify violation of the reliability of the survey.

3.1 Survey Data Exploration

The data sets explored are from real-world surveys conducted by D3 Systems in two countries: Cameroon and Philippines. Human agents conducted surveys among general population using electronic devices. Our data sets are the collection of metadata from the electronic devices, some of which recorded the GPS coordinate at the time of each survey. We refer to each instance of the survey as a *point* and the set of points that have been collected on a same electronic device as a *group*. Our algorithms aim to find points within each group that display geographic anomaly such as isolation or concentration. To properly apply the methods described in the subsequent sub-sections, we only consider groups that contain at least 10 points with properly recorded GPS locations. Table 1 highlights descriptive analysis comparing the two country data sets.

Table 1. Descriptive Analysis Comparing the Two Country Data Sets Explored

Country Data Set	Total Group Count	Total Point Count	Mean Point Count Among Groups	Median Point Count Among Groups	Max Point Count Among Groups
Cameroon	46	1721	45.7	41	135
Philippines	59	1388	21.4	21	40

ISBN: 97819384962-2-6

3.2 Distance Metrics for Outlier Detection

Distance metrics measure how different a point is from another point. We consider two different distance metrics, Haversine and Mahalanobis, that are then used within the two outlier detection algorithms that we implemented.

3.2.1 Haversine Distance

Most GPS data provides geographic coordinates in terms of latitude and longitude. The haversine distance is the great-circle distance between two GPS coordinates. Suppose we know the radius of the Earth r and two points on a map are given as tuples of latitude and longitude $a = (\phi_1, \lambda_1)$ *and* $b = (\phi_2, \lambda_2)$, we calculate the haversine distance $d(a, b)$ as follows:

$$d(a,b) = 2r \arcsin\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cdot\cos(\phi_2)\cdot\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)} \quad (1)$$

3.2.2 Mahalanobis Distance

The Mahalanobis distance accounts for the distribution of the points in the group by taking the correlation among the attributes into consideration. Suppose we consider each point as a vector of normalized attributes with mean 0 and standard deviation 1 within the group to which the point belongs. Then, for two points $\vec{a}$ and $\vec{b}$ that belong to a same group *G*:

$$d(\vec{a}, \vec{b}) = \sqrt{(\vec{a} - \vec{b})^{\mathrm{T}}\mathbf{S}^{-1}(\vec{a} - \vec{b})} \quad (2)$$

where **S** is the nonsingular covariant matrix of the points in G. An advantage of Mahalanobis distance is that we can use additional numerical attributes to measure the degree of difference between two points. However, we only consider geospatial attributes (longitude and latitude) to calculate Mahalanobis distance for one-to-one comparison with haversine distance.

3.3 Outlier Detection Algorithms

Geospatial outlier detection algorithms seek to assign a scalar value to each point based on the distance metrics to other points within a group. If the scalar values of points in the group make a distribution centered around a mean, for example, then we can identify the points with values greater or lower than certain threshold away from the mean as outliers. We will use two standard deviations away from the mean as the threshold in this work. Here, we describe two outlier detection algorithms: *mean-distance method* and *outlier factor method*. Each method requires calculation of distances between every pair of points for each group. Instead of calculating distance between points every time required by the algorithm, it is more efficient (timewise) to create a dictionary data structure with the key as a pair of points in the same group and value as the distance between the pair. Thus, given a dataset D, we divide D into disjoint subsets of data with common group ID. Let $\mathrm{Gr}(D)$ be such partition. We refer to each subset as a *group*. For each group $G \in \mathrm{Gr}(D)$, we calculate and store $d(a, b)$ for all $a, b \in G$ where $d: G \times G \to \mathbb{R}^+ \cup \{0\}$ is the chosen metric. We note that any distance metric can be used for the algorithm. However, using a different metric will produce a different distribution of scalar values assigned to the points, and hence identify different outliers.

3.3.1 Mean-Distance Method

The *mean-distance method* calculates the mean of distances of a point to all other points within the group to which the point belongs. One advantage of the mean-distance method is that we can assign a scalar value *group mean* to each group to identify groups whose points are abnormally farther apart or concentrated compared to other groups. See Figure 1 for the pseudocode of Algorithm 1.

3.3.2 Outlier Factor Method

The *outlier factor method* calculates the degree of local isolation of each point from a nearby cluster. This requires introduction of a hyperparameter *k* which we set as 5. The outlier factors of points in a group create a positive distribution centered at 1. Outlier factors greater than 1 indicate points that are more isolated whereas values less than 1 indicate points that are more concentrated compared to other points within a group. See Figure 1 for the pseudocode of Algorithm 2.

4. Computational Experimentation and Results

In this experiment, we identify geospatial outliers among data points of the Cameroon and Philippines data sets discussed in Section 3.1 with different combinations of the distance metrics and the outlier detection algorithms described in Sections 3.2 and 3.3. Table 2 shows the proportions of flagged points by different algorithms when the distance metrics used are fixed. The percentage values measure the proportions out of the total flagged.

We note that most of the flagged points are not flagged by both algorithms for each metric. Only 10-19% of the flagged points are flagged by both algorithms. Initially, this is surprising given both algorithms seek to measure the degree of isolation of each point in a group. Further examination reveals that this is due to the geographical distribution of the points within each group. Many survey collection devices were used in multiple cities/towns resulting in clusters of points for each group. While the mean-distance method considers all points within each group to assign a scalar, outlier factor method only considers nearest clusters nearby each point as determined by the hyperparameter *k*. Therefore, a point considered as an outlier by the mean distance method do not have to be considered an outlier by the outlier factor method and vice versa. Note that the outlier factor method flagged more points (50.30-54.82%) in the Cameroon data set compared to the Philippines data set (24.32-30.29%).

Algorithm 1 Mean-Distance

Suppose for a group $G \in Gr(D)$, we have a data point $a \in G$. Calculate the average distance from the point to other points within the group as the *case mean*.

$$\mathrm{cam}(a) = \frac{\sum_{b \in G \setminus \{a\}} d(a,b)}{|G| - 1}$$

Calculate the *group mean* of G as the average of all case means within the group.

$$\mathrm{grm}(G) = \frac{\sum_{a \in G} \mathrm{cam}(a)}{|G|}$$

Calculate the *group standard deviation* as the standard deviation of the set of case means within the group.

$$\mathrm{grsd}(G) = \sqrt{\frac{1}{|G|} \sum_{a \in G} (\mathrm{cam}(a) - \mathrm{grm}(G))^2}$$

Then, a is an outlier in G if $|\mathrm{cam}(a) - \mathrm{grm}(G)| > 2 \cdot \mathrm{grsd}(G)$.

Algorithm 2 Outlier Factor

Let $N_k(a)$ be the set of k nearest neighbors of a in G according to a metric $d : G \times G \to \mathbb{R}^+ \cup \{0\}$. Let $d_k(a) = \max\{d(a,c) | c \in N_k(a)\}$. That is, the distance of a from kth nearest neighbor. Define *reachability distance*.

$$\mathrm{rd}_k(a,b) = \max\{d_k(b), d(a,b)\}$$

Define *local reachability density*.

$$\mathrm{lrd}_k(a) = \frac{|N_k(b)|}{\sum_{b \in N_k(a)} \mathrm{rd}_k(a,b)}$$

For each data point, calculate *local outlier factor*.

$$\mathrm{lof}_k(a) = \frac{\sum_{b \in N_k(a)} \mathrm{lrd}_k(b)}{|N_k(a)| \cdot \mathrm{lrd}_k(a)}$$

We calculate the mean m and standard deviation s of the local outlier factors with in G.
Then, a is an outlier in G if $|\mathrm{lof}(a) - m| > 2 \cdot s$.

Figure 1. Descriptions and Pseudocodes of the Two Outlier Detection Algorithms

Table 2. Counts of Flagged Points Between Two Algorithms for each Distance Metric

Country Data Set / Distance Metric Used	Total Flagged	Flagged by Both Algorithms	Flagged by Mean-Distance Only	Flagged by Outlier Factor Only	Flagged by None
Cameroon /					
Haversine	166	23 (13.86%)	52 (31.33%)	91 (54.82%)	1525
Mahalanobis	169	17 (10.06%)	67 (39.64%)	85 (50.30%)	1522
Philippines /					
Haversine	175	33 (18.86%)	89 (50.86%)	53 (30.29%)	1085
Mahalanobis	185	34 (18.38%)	106 (57.30%)	45 (24.32%)	1075

As displayed in Table 3, we now fix the outlier detection algorithm and vary the distance metrics. There is greater overlap of flagged points between different metrics than the overlap between different algorithms. Figure 2 depicts the plots of the flagged points identified in Tables 2 and 3 on geospatial maps of Cameroon and Philippines, respectively.

ISBN: 97819384962-2-6

Table 3. Comparison of Flagged Points Between Two Distance Metrics for each Algorithm

Country Data Set / Outlier Detection Algorithm Used	Total Flagged	Flagged by Both Metrics	Flagged by Haversine Only	Flagged by Mahalanobis Only	Flagged by None
Cameroon /					
Mean-Based	183	46 (25.14%)	29 (15.85%)	108 (59.02%)	1508
Outlier Factor	179	43 (24.02%)	71 (39.66%)	65 (36.31%)	1512
Philippines /					
Mean-Distance	179	71 (39.66%)	51 (28.49%)	57 (31.84%)	1081
Outlier Factor	136	47 (34.56%)	39 (28.68%)	50 (36.76%)	1124

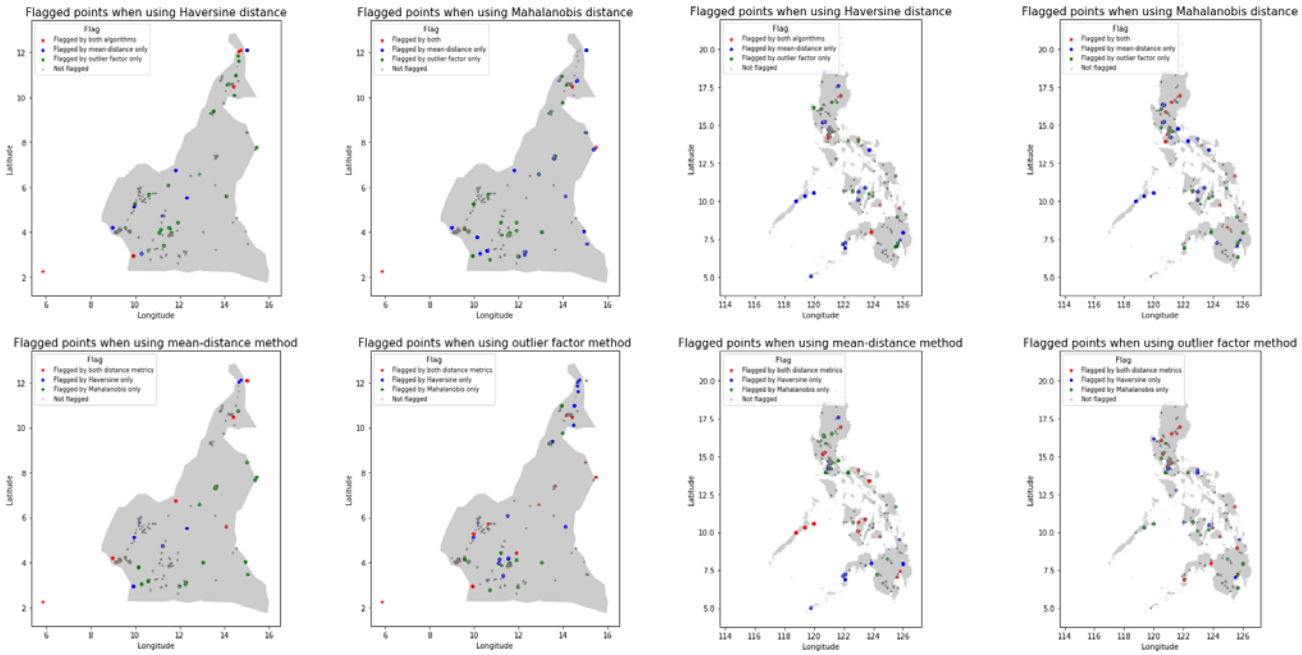

Figure 2. Flagged Points in the Cameroon/Philippines Data Sets with Various Metric/Algorithm Combinations

Next, we analyze the computational time associated with implementing each method. Assuming amortized constant time for adding an item to a dictionary, the expected time complexity of creating a distance dictionary is at least $O(mn^2)$, where m is the number of groups and n is the maximum count of points for a group. The calculation of Mahalanobis distance is far more costly than the calculation of haversine distance as it requires calculation of a covariance matrix per attribute per group in addition to matrix-vector multiplication.

Table 4 displays the comparison of the computational times to create distance dictionaries for the two distance metrics used. Since we are using dictionaries to store distances instead of calculating them when required, our choice of distance metric does not affect the computational times of the outlier detection algorithms themselves. Table 5 shows the comparison of computation times of the algorithms applied to the two data sets.

Table 4. Comparison of Computation Times for Creating Distance Dictionary (Average of 100 Iterations)

	Haversine	Mahalanobis	Ratio
Cameroon (46 groups, 1721 points)	1.32 s	8.38 s	6.32
Philippines (59 groups 1399 points)	0.504 s	3.61 s	7.16

ISBN: 97819384962-2-6

Table 5. Comparison of Computation Times for Outlier Detection Algorithms

	Mean-Distance	Outlier Factor	Ratio
Cameroon (46 groups, 1721 points)	36.4 ms	169.0 ms	4.64
Philippines (59 groups 1399 points)	16.7 ms	49.8 ms	2.98

While the outlier factor method is more computationally expensive than the mean-distance method, we note that most of the total computational time will be used on creating the distance dictionaries rather than performing the algorithms themselves. Focusing on hastening calculation of distance metrics will better optimize the running time.

4.1 Verification and Validation

Verification is the process of ensuring that the outlier detection algorithms are implemented correctly, which was achieved through testing the algorithms for correctness, stability, and convergence. Algorithm code reviews and inspections were performed in addition to some baseline testing to verify that our implementation of the algorithms indeed detect relevant geographic outliers. For example, we created an artificial data set of a random distribution of points with clear instances of outliers that were visibly away from nearby clusters. Both algorithms reliably detected geographical outliers when using the Haversine distance metric. However, the presence of an extreme outlier had significant impact on the overall correlation between the coordinates such that both algorithms sometimes failed to identify desired outlier points when using the Mahalanobis metric.

Validation is the process of ensuring that the outlier detection algorithms adequately capture the phenomenon of interest (i.e., actual survey outliers) by comparing with experiments and/or observations. Further work is needed to fully validate our implemented algorithms, which will entail comparing the results of our implemented algorithms to ground truth data provided by D3 Systems where their survey analysts identified and confirmed actual outlier points.

5. Conclusion and Future Work

In this work, we explored and implemented two different algorithms and two different distance metrics to identify statistical anomalies among geospatial data points from two real-world surveys. By viewing distance metrics as not simply the physical distance between two points but a measure of *difference* between the points, one can use more than locational data for geospatial outlier detection algorithms. If the devices used to conduct interviews/surveys can collect additional attributes such as interview duration or distance/time since previous interview/survey, the Mahalanobis distance may better measure how an interview differs from other interviews and help identify instances of policy violation when used in outlier detection algorithms.

We noticed that using the outlier factor method, which considers the local distribution of points in a dataset, identifies a very different set of outliers when compared to using mean-distance, which does not. Since there is only a small overlap (10-19%) of flagged points between the two algorithms explored in this work, it may be helpful for survey analysts to focus on the fewer points that are flagged by both methods rather than all the "outlier" points that are flagged by each algorithm. Alternatively, we suggest a visual inspection of the points within each group to decide on the geospatial outlier detection algorithm to be applied. If the locations of the points span multiple geographically distant jurisdictions, the outlier factor method will better identify relevant geographical outliers within each jurisdiction. However, we currently have no evidence to assert that the flagged points show instances of policy violation by the surveyors/interviewers.

Future work includes the creation of additional artificial survey data sets for continued verification of whether the algorithms catch instances of violation of reliability and to help determine which algorithm may better catch certain violation of reliability such as straight-walking or standing still. Moreover, future work will include thorough algorithm validation, as we were limited by the lack of actual survey outlier ground truth data. Additional future work entails building a geospatial big data analytics application with a user interface to visualize and interact with the data while allowing the user to select the algorithm and distance metric to use. An ability to plot flagged data points on a zoomed-in satellite image could help survey analysts understand why the points were flagged as an outlier and decide whether they warrant further investigation. While the flags can help reduce dependence on visual analysis by reducing the number of points examined, this tool will help automate survey quality control for analysts to ensure the reliability of surveys while drawing important conclusions in a timely fashion.

ISBN: 97819384962-2-6

6. References

Ben-Gal, I. (2005). Outlier detection. In *Data mining and knowledge discovery handbook* (pp. 131-146). Springer, Boston, MA.

Brassil, C. (2021, August 27). *D3 Systems*. D3: Designs, Data, Decisions. https://www.d3systems.com/our-process/

Brassil, C. (2021b, August 27). *GPS Training III – Basic Analysis*. D3: Designs, Data, Decisions. https://www.d3systems.com

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (pp. 93-104).

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (1999, September). Optics-of: Identifying local outliers. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 262-270). Springer, Berlin, Heidelberg.

Singh, A. K., & Lalitha, S. (2018). A novel spatial outlier detection technique. *Communications in Statistics-Theory and Methods*, 47(1), 247-257.