An Evaluation Framework for Assessing the Impact of Location Privacy on Geospatial Analysis

MAYRA A. ZURBARÁN1, AUGUSTO SALAZAR1, MARIA ANTONIA BROVELLI2, AND PEDRO M. WIGHTMAN1, (Senior Member, IEEE)

1Department of Systems Engineering and Computing, Universidad del Norte, Barranquilla 080001, Colombia
2Department of Civil and Environmental Engineering, Politecnico di Milano, 20133 Milano, Italy

Corresponding author: Pedro M. Wightman (pwightman@uninorte.edu.co)

ABSTRACT The need for addressing geoprivacy in location based services has increased the offer of mechanisms that protect location information, however, these algorithms are not always developed to ensure the usability of the data and therefore, their adoption is not wide. In this work, a framework is presented to evaluate the effects of geoprivacy mechanisms on the quality of geodata to provide insights into how the data is affected for geospatial analysis. For this purpose, a toolkit of indices was developed to evaluate different characteristics of the data before and after a geoprivacy mechanism is implemented, providing a criterion to select one of them. The indices measure the changes in the presence of clusters through the quantification of hotspots in hotspot analysis and the difference observed in heatmaps of the concentration of the geodata. Variations in global indices like the Nearest Neighbor Index (NNI) and the orientation of the standard deviational ellipse are also measured. For demonstration, the data of crime arrests in New York was used for the month of January in 2017 and 2018. Five mechanisms were tested with different settings, resulting with the NRand-K algorithm producing fewer alterations to the reference data, preserving its initial characteristics better than the other mechanisms.

INDEX TERMS Exploratory spatial data analysis, geoprivacy, geospatial analysis, location based services, location obfuscation, masking.

I. INTRODUCTION

The need for addressing privacy issues while being able to obtain feedback in a Location-Based Service (LBS) has been a concern for some years. In [1], the authors state that “enabling location privacy is going to become increasingly important in a world where location-aware services are available over larger and larger geographical areas. Deploying systems that support location privacy for users and provide feedback about the level of anonymity users have may prove critical to the widespread adoption of location-aware services.”

Referring to health data disclosure, in [2] is said: “the consideration of both accuracy and efficiency argues for maintaining health data in a spatially disaggregated form so that it can be aggregated, flexibly, to any spatial unit of interest or any set of small areas that share a common level of environmental exposure.” However, in order to protect the privacy for public disclosure, the location data may be altered considerably, decreasing the value of the information it carried for performing geospatial analysis and it may not reflect reality anymore. Location Privacy Protection Mechanisms (LPPMs) should aim to take into account to what measure the data is affected, and therefore if it remains useful for research and analysis. Companies whose core business depends on the quality of location data for their analyzes will not adopt a location privacy protection mechanism that damages their business model.

In this light, measuring the impact of LPPMs in geospatial analysis is now a critical step, in order to understand how these analyzes will be affected if there is a massive adoption of privacy protection techniques, either due to public awareness or due to legislation and enforced geoprivacy standards; or probably both.

This work is dedicated to define a general criterion to measure the impact that location obfuscation techniques have in geospatial analyses. To achieve this, the performance of tools belonging to Exploratory Spatial Data Analysis (ESDA),
like hotspot analysis (that helps in identifying clusters), heatmaps, and the calculation of different global and local spatial indices are explored as tools to quantify alterations of the data after applying privacy protection mechanisms. The selected LPPMs to test include noise-based algorithms. These algorithms disturb the original location but still provide distorted geographical coordinates as an obfuscated result as seen in Table 1. In particular, algorithms to include in this evaluation are: Donut Masking from [3], Pinwheel from [4], NRand-K from [5], VoKA: Voronoi K-Aggregation from [6], and a mechanism that induces noise by clustering based on VoKA from [7]; the framework will contain different analysis and metrics that will be computed with the datasets in their original form and obfuscated with the different mechanisms. The evaluations will aid in identifying what kind of perturbations are more suitable to preserve geographic properties and in general the framework will be replicable for any kind of LPPM based on noise or masking.

The dataset to use in this evaluation framework is of crime arrests in the city of New York, this historic data was made available in [8]. Enabling the creation of subsets to test different data distributions.

This paper is organised as follows: a subsection detailing the Study Design, a section of Related Work to describe the mechanisms to include in the evaluation and other evaluation techniques, the Methodology used for evaluating the LPPMs independent of the specifics of each mechanism, followed by a section dedicated to the performed Experiments with each of the selected LPPM. At last, a Discussion section comparing the results is dedicated to evaluate the performance of each mechanism in terms of alterations on geospatial analysis.
A. STUDY DESIGN
This work aims to produce a replicable methodology to evaluate degradation of spatial information in terms of longitude and latitude when geoprivacy methods have been applied. For this, the study plan is to recognize the most effective mechanisms to provide geoprivacy and develop steps using the tools of spatial analysis to provide an evaluation framework that measures how these mechanisms perform in terms of distorting location information for geospatial analysis purposes.

II. RELATED WORK
The study of mechanisms to preserve privacy location in location-based services is wide and have been going on for a couple of decades in the research community, according to [2] “The best approach depends on the purpose of the data user as well as the degree of risk of disclosure that the data custodian wishes to tolerate.” It is no surprise that this topic is so explored since the need for mechanisms that work and enable a civilized treatment of sensitive data according to guidelines will persist as long as data is relevant. In [5] is presented a comprehensive study of LPPMs categorizing them according to their purpose, other survey studies are presented in [9], [10] and [11]. In general they convey in different types of mechanisms according to the distortion method applied in them. Below are the ones identified in [5] and the mechanism general definition.

From the categories of LPPMs in Table 1, the ones that enable some kind of geospatial analysis fall under noise-based or masking, since they provide geographic coordinates as output and do not multiply them, in [2] is stressed the limitation of aggregation often implied in k-anonymity for the purposes of research with geodata: “it is clear that geographical analyses must investigate relationships in areas that are defined far more flexibly than is possible using spatially aggregated data for current political and administrative entities.” Aggregation impacts highly the detection of clusters, for these reasons the mechanisms based on noise will be the focus of this work. In some cases some type of aggregation may be used by masking mechanisms to provide obfuscation but the nature of the mechanism is still inducing noise based on location allocation yet providing an obfuscated or masked coordinate for each original location.

In [2], they focus on health data as a sensitive type of information and the importance of being able to perform spatial analysis while keeping confidentiality. They perform tests on different mechanisms specifically for health data and how they preserve some properties after masking, concluding that clustering is affected by high levels of perturbation but with lower levels the clusters found in the original data persist.

In [12] is proposed a spatial masking based on spatial smoothing and the authors evaluate the effects of masking in statistical analyzes like regression using the Mean Squared Error (MSE) parameter when fitting a Generalized Linear Model (GLM) to the masked data. They conclude in providing parameters for smoothing of the data using loglinear models that preserve the utility of the data while preserving privacy. This approach is interesting in terms of statistical impact of the perturbation, however, spatial analysis is still yet to be evaluated.

In [13] is presented a study to evaluate dissimilarities among crime geodata that has been masked. This study focuses on masking algorithms, the algorithms they evaluate consist of random noise and donut masking which creates a cloaking area around the original location and only adds noise that falls further than these cloaking areas, having a maximum radius of noise to add. However, more complex algorithms have emerged and there are more advanced geospatial analysis tools to test their usability, like the hotspot analysis tool now available in QGIS as a plugin. In [13] global indices and local indices are presented to evaluate the behavior of clusters and the distribution of the points globally.

Algorithms that are noise-based or masking are defined as “the means of deliberately degrading the quality of information about an individual’s location in order to protect that individual’s location privacy.” according to [14], some of these are presented in [15], where density is had under consideration in the methods. in [2] is described displacement, change of scale, rotation, among other mechanisms. In [16] there is also proposed radius enlargement which is similar to donut masking, shifting the center for the perturbation, reduction of the maximum radius, and their combination.

From these mechanisms, the selected ones were chosen due to their efficiency in obfuscating location, the possibility of still being able to work with geographic coordinates, and ease for implementation. In addition, clustering based on VoKA, VoKA, NRand-K are already included in QGIS as plugins which enforces their adoption.

A. DONUT MASKING
The donut method for masking ensures that an address is not randomly assigned on or too near its original location. “In donut method geomasking, each geocoded address is relocated in a random direction by at least a minimum distance, but less than a maximum distance” as stated in [3].
FIGURE 2. Pinwheel distribution of points.

Even though donut masking was first introduced with a variable perturbation depending on a desired density around the location, in the implementation of donut masking for this article, polar coordinates will be used to generate the random perturbation with a fixed minimum minDis and maximum distance maxDis (Fig. 1). In [5], polar coordinate generation creates a more even distribution within the area of masking “Polar-generated points are distributed more evenly along the radius. The fact that there seems to be more points closer to the centre is due to the symbol used to represent points, giving the impression of closeness, when in fact is a more uniform distribution.” (p. 41) This will enable comparison based on maximum noise added as a fixed value with other mechanisms.

B. PINWHEEL

Pinwheel is a noise-based algorithm that adds noise based on the Equation (1):

\[ r_\alpha = \frac{\alpha \mod \phi}{\phi} \ast r_{\text{max}} \]  

where \( \phi \) is a parameter that will define the degrees wide of the pinwheel like distribution of the resulting points. A small value of \( \phi \) will result more and thinner wings, a high value will result in less wider wings; therefore this determines the periodicity of the distribution of the output obfuscated location as shown in Fig. 2. For a random \( \alpha \) between \([0^\circ - 360^\circ] \), the range of \( \phi \) should be in \([0^\circ - 359^\circ] \). Examples of different values of \( \phi \) are shown in Fig. 2 Some distributions with high asymmetry of noise proved to be more efficient against filtering attacks, this is possible by adjusting values of \( \phi \) according to [4]; therefore offering more geoprivacy with the same maximum radius of noise than a normal distribution. In order to decrease the probability that an obfuscated location falls too far from its original location, an extra step was taken for the pinwheel implementation in this paper. From Equation (1), this calculated radius will become the maximum radius for a second calculation of the final resulting radius with the same \( \phi \) parameter as in Equation (1), where \( \text{random()} \) is a value between \([0 - 1] \):

\[ r_{\text{final}} = \text{random()} \ast r_\alpha \]  

This final radius from Equation (2) will impact the results in offering less noise, however, this serves to preserve variability and more information of the original location which can be a Point of Interest (PoI) with many locations reported at the same coordinates.

C. VoKA: VORONOI K-AGGREGATION

VoKA is designed to provide geoprivacy at the time of releasing a dataset and therefore its computation is not on-the-fly as the data is being collected, rather after the dataset is formed and ready to be released to the public or a third-party. In [6] four main types of geolocated sources are identified according to density considering that a single source may have many entries or locations reported on it, these types are: isolated living areas, low density living areas, isolated points of interest, and popular areas. This mechanism uses location allocation in order to obfuscate, geolocated sources with low number of entries or locations reported in them will be aggregated with other nearby sources, this process is done by defining a desired level of K-anonymity which is the number of locations each source should represent at least, though this number is not always guaranteed based on the dataset. The location allocation is done by means of getting the neighbors by using the Delauny triangulation.

From Fig. 3, KD-Aggregation is the process used to sort the geolocated sources by the number of entries; from smaller to largest, then by latitude and finally, by longitude. The sources that already comply with K are not visited further for aggregation unless they are neighbors of a source that does not comply.

D. NRand-K

NRand-K is an offline mechanism designed to provide geoprivacy at the time of releasing a dataset. The parameters to add noise depend on the density of the cell the locations are in. It works by dividing the space into a grid of a size depending on the dataset volume and area, and determining a \( K \) which will decide if a cell contains enough locations to be considered geoprivate as in Equation(3). This assumption is based in that cells act as a k-anonymous area for the locations contained in them. If the cell does not contain at least \( K \) locations a maximum radius of noise will be added else, if it contains at least \( K \) locations, a minimum radius of noise is added as shown in Fig. 4. This comes from the notion that crowded cells (more K-anonymous) need less protection against geoprivacy attacks than low density cells.
E. GEOPRIVACY WITH SPATIAL CLUSTERING

In [7], the authors describe mechanisms that follow the principle of aggregation of near locations as is done in VoKA but using traditional clustering mechanisms such as K-Means and DBSCAN. The mechanisms first aggregate in a grid-like manner by truncating decimal digits of the latitude and longitude, where each digit represents a scale in meters for a grid; when more digits are truncated, the grid cells are bigger, and more locations will be aggregated. After this first aggregation, a clustering method is applied, however these are susceptible to induce point loss which is undesirable. For the experiments conducted in this paper, the method using K-Means is selected.

III. EVALUATION FRAMEWORK

The evaluation framework will mainly focus on the effects of masking on clustering by means of analysing the results of hotspot analysis, global indices, and heatmap divergence. Clustering detection is of special importance in spatial analysis and crime data, therefore the framework will be oriented towards the detection of alterations in this aspect.

For this, there are taken into consideration the global Nearest Neighbor Index (NNI) and Z-score, which will initially tell if the data is clustered or not. To calculate the NNI, the QGIS implementation for its calculation is used.

For evaluating the hotspot analysis, the count of alterations in hot spots, not significant and cold spots of the masked data versus the reference datasets will be measured by using the Hotspot divergence index (Hdi) presented in Equation (4):

$$Hdi = \frac{\sum_{i=1}^{7} \text{RefValue}_i - \text{MaskedValue}_i}{\text{Total number of polygons}}$$

Seven corresponds to all the significance values evaluated in the hotspot analysis, this measure will provide an average on how different a mechanism performs in this analysis from the reference dataset output. For a mechanism with no alterations the result would be 0. Each hotspot polygon is compared to the corresponding polygon in the reference dataset.

To analyse how the composition of heatmaps is affected, the normalized heatmap outputs of each mechanism is subtracted from the reference data heatmap using the RGB channels and images of the same size. The Heatmap divergence index (Hedi) is calculated by the average change per pixel using matplotlib. Equation (5) shows the calculation of this index where n is the total number of pixels in each image.

$$Hedi = \frac{\sum_{i=1}^{n} \text{Pixel Error}_i}{n}$$

To understand globally how the data was affected by the masking, the mean divergence index (Mdi in Equation (6)) from [13] is used. This tells how much the mean coordinates of the data shifted after the masking. In addition, the orientation of the standard deviational ellipse is measured by using the orientation divergence index (Odi in Equation (7)) proposed as well in [13].

$$Mdi = \frac{d(\text{original mean, masked mean})}{d(\text{original mean, the farthest point away})}$$

$$Odi = \frac{\sum_{i=1}^{n} \text{Orientation Error}_i}{n}$$
Other global indicators are the maximum and average noise induced by each masking mechanism, which describe how much distortion was added in the geoprivacy process. These are indicators of why the other indices are altered more by one mechanism or other depending not only on the maximum radius for noise set, but the actual average of noise induced in the data by using the mean squared error (MSE) in Equation (8) where \( n \) is the total count of crimes in the datasets.

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} d(\text{original location}_i, \text{masked location}_i) \tag{8}
\]

Aggregation is of great importance in geospatial analysis and the implications of altering the data in the geoprivacy implementation is definitive to how this data will serve to the analysis. For this reason an index is presented to measure how aggregation is affected by applying geoprivacy mechanisms and verifying how the aggregation count varied. The index for this was first introduced in [5] and is described as the difference between the point count per each polygon used to aggregate the original data (original count) and the point count after obfuscation (obfuscated count) is used to calculate the Point Count divergence index (PCdi). This index is calculated using all polygons included in the analysis and is implemented in this framework as in Equation (9), where \( n \) is the total number of polygons in the aggregation:

\[
PCDi = \frac{\sum_{i=1}^{n} (\text{original count}_i - \text{obfuscated count}_i)}{n} \tag{9}
\]

The framework is best represented by a table that describes the explained indices and summarizes the performance of each evaluated mechanism with respect to the reference data. The framework aids in estimating which mechanism alters less the output data after privacy requirements, however, it can happen that some mechanisms perform best in a global scale and poorly while preserving clusters. For this reason, it is up to the analyst to have a criterion on what should be the priority

A. TOOLS

The tools used to conduct experiments and design the indices that compose the evaluation framework are PostgreSQL with the PostGIS extension, QGIS with the Hotspot Analysis and the Standard Deviational Ellipse plugins, python with matplotlib, and Excel.

The Hotspot Analysys functionality is an Exploratory Spatial Data Analysis (ESDA) tool. ESDA is a set of different statistical techniques to describe and visualize distributions and the identification of local patterns of spatial autocorrelation which is most desirable and probable when doing spatial data analysis as stated by Tobler’s law.\(^1\) The Hotspot Analysis calculates the Getis-Ord local statistic \( G^* \) from [17] to identify areas of atypical point density occurrence, either high (hot spots) or low (cold spots).

To be a statistically significant hot spot, a feature will have a high value and be surrounded by other features with high values as well. In this case, it was implemented first by counting the points inside the polygons described by the New York census blocks, serving these counts as weights for the census blocks polygons to compute hotspots using the queen’s case lookup to compare with the scores of statistical significance between neighboring polygons.

The Standard Deviational Ellipse was used with the Crime-Stat configuration and tells if the orientation of the data was or not affected by the masking technique applied when comparing. This is useful to understand how much the overall data is affected besides the clusters.

The Nearest Neighbors Analysis tool of QGIS is used to calculate the values for the Nearest Neighbor Index (NNI), the Z-score of the datasets, and the mean coordinates.

IV. METHODOLOGY

The amount of noise or perturbation to add is still an open discussion and depends highly on the data, a level of noise of 1km may work in the outskirts of a city where the density is low but may result too much in the city center for example. This depends also on the level of aggregation desired for further analysis, too much noise may place a location in a different cell than the one it originally belonged to. In [18] is used maximum noise thresholds of 30, 50, and 250m; in [19] experiments were made with 1032m; and in [20] is demonstrated how inferences can still be made with an induced noise of 1000m, but are less likely than when 500m is used. This means that there is no established number and it rather depends on the purpose of protection, utility of the data, and further analysis.

For this work, maximum noises of 200m and 500m are used for our case study. This was set having into consideration the average area of 20.25km\(^2\) of a census block in New York which is the selected level of aggregation for the spatial analysis.

After selecting the parameters to evaluate each mechanism in a sensible way that allows for comparison, i.e. mechanisms with similar levels of distortion induced, the indices described in the framework are calculated for each obfuscated dataset.

V. EXPERIMENTS

This section will apply the framework described to evaluate if the selected mechanisms preserve the geographic properties of the data enabling geospatial analysis and to what degree.

First the initial indices are calculated for the original data, these are, the reference datasets for Jan 2017 and Jan 2018. These will serve to calculate the different indices that compose the evaluation framework.

---

\(^1\)Tobler’s first law of geography (Tobler, 1970) states that everything is related to everything else, but near things are more related than distant things.
A. DATA

The historic crime data obtained from [8] was subset into two datasets in order to have more than one reference dataset for testing the framework, the datasets contain the crimes registered in January 2017 (25770) and January 2018 (23882). These two will be our original data to be masked with the different algorithms. The locations of the crimes cover an area of 785.64km$^2$.

During the exploratory data analysis it was observed that some of the locations in the data are overlapped and very close
to the census block borders, hence any masking method used will highly affect these overlapped points and the resulting data clusters.

The count of hot spots and cold spots for the original datasets is presented in Table 2 along with the other indices. These counts for Jan 2017 and 2018 will serve as the reference count of hotspots for the masked data. If there are any alterations, these mean the hotspot analysis was affected by the masking technique applied. The measure for this will be as shown in Equation 4.

The difference observed in the total blocks item of Table 2 is due to the presence or not of crime arrests in some census blocks; the total hotspot count is therefore also affected. For 2017 all census blocks contained data whereas for 2018 there were 59 without crime appearances reported and therefore not appearing in the resulting hotspot analysis output. The Nearest Neighbor Index, the mean coordinates, the Orientation of the Standard Deviational Ellipse and the eccentricity of it are very similar for both datasets, this indicates that the reference datasets we are using pose some similarities that help maintain consistency when comparing with the masked data. They do differ however in the appearance of hotspots and the distribution. In the following subsection, the parameters used for the different mechanisms are explained.

### B. PARAMETERS FOR EACH MECHANISM

The parameters for each mechanism where chosen in general for two scenarios, one for a maximum noise of 200m and one for a maximum noise of 500m; however, each mechanism has its specificities and these are the ones that need to be adjusted in order to get the desired maximum noise to get comparable masked data in terms of noise added.
For NRand-K, the 3rd quantile of the census blocks that contain some arrests is taken for the K value which resulted in 2.

For algorithms that cluster neighboring locations like VoKa and K-Means based obfuscation, the parameters were selected based in the number of output clusters in order to make them comparable to the other mechanisms.

The mechanisms where implemented as described in Subsections II-A, II-B, II-C, II-D, II-E and were executed with the parameters listed in Table 3. The output of the evaluated indices is described by Table 4, in this Table are highlighted the best values for the indices with respect to the reference datasets.

Table 4 and Fig. 9 of a radar chart describe the indices that compose the framework as well as noise metrics employed since the mechanisms evaluated were tested on different noise settings. The radar chart aims to visualize what the table represents, in order to calculate the values for the chart, the indices that represent a negative feature were taken as the inverse and scaled to a range of 1 to 100. Both, the chart and table show that there is not a single mechanism that performs best for all indices, rather mechanisms that have an average performance overall indices and some with best performance on single indices.

Doing a further analysis by looking at the values, it shows that NRand-K preserves the NNI and Z-score better compared
FIGURE 8. Heatmap divergence calculation for 01/2018 for the different mechanisms.

TABLE 2. Original datasets hotspot analysis count per census block and indices.

| Significance            | Arrest data 01/2017 | Arrest data 01/2018 |
|-------------------------|---------------------|---------------------|
| Coldspot 99% confidence | 0                   | 0                   |
| Coldspot 95% confidence | 0                   | 0                   |
| Coldspot 90% confidence | 0                   | 0                   |
| Not Significant         | 36127               | 34727               |
| Hotspot 90% confidence  | 291                 | 288                 |
| Hotspot 95% confidence  | 483                 | 443                 |
| Hotspot 99% confidence  | 1291                | 1214                |
| Total blocks            | 38794               | 38672               |

TABLE 3. Location privacy preserving mechanisms.

| Mechanism          | Min Noise | Max Noise | K  | φ  | Search Distance |
|--------------------|-----------|-----------|----|----|-----------------|
| Donut Masking 1    | 50m       | 200m      | –  | –  | –               |
| Donut Masking 2    | 100m      | 500m      | –  | –  | –               |
| NRand-K 1          | 50m       | 200m      | 2  | –  | –               |
| NRand-K 2          | 100m      | 500m      | 2  | –  | –               |
| Pinwheel 1         | –         | 200m      | –  | 30 | –               |
| Pinwheel 2         | –         | 200m      | –  | 100| –               |
| Pinwheel 3         | –         | 500m      | –  | 30 | –               |
| Pinwheel 4         | –         | 500m      | –  | 100| –               |
| VoKA               | –         | –         | 20 | –  | 1km             |
| K-Means            | –         | –         | 400| –  | –               |

The mechanisms that use clustering or aggregation such as VoKa and K-means result in a NNI near zero because all points are clustered after their application, affecting the NNI in this way. These mechanisms also have a higher MSE and maximum noise induced, thereby providing more privacy but having a stronger impact in some of the indices. The Odi is close to zero for all mechanisms, however, variations of Pinwheel in both datasets produce the lowest value. A similar behaviour occurs with the Mdi, all values are close to zero, however, NRand-K 1 and Pinwheel 1 show the smallest values. The values obtained for the PCdi represent the average count of differences in the obfuscated datasets compared to the ones of reference. For this index, Pinwheel 2 showed...
TABLE 4. Evaluation framework indices for masking mechanisms.

| Mechanism   | Hdi     | Hdi   | MSB (m) | Max Noise (m) | Avg Clusters | NNI  | Z-score | Odi    | Mdi    | PCdi  |
|-------------|---------|-------|---------|---------------|--------------|------|---------|--------|--------|-------|
| 01/2017     |         |       |         |               |              |      |         |        |        |       |
| Donut 1     | 0.078   | 0.00240 | 125     | 200           | –            | 0.38 | -190.1  | 0.00001 | 0.000013 | 0.80  |
| Donut 2     | 0.122   | 0.00296 | 300     | 500           | –            | 0.44 | -170.7  | 0.00001 | 0.000050 | 0.90  |
| NRand-K 1   | 0.051   | 0.00108 | 75      | 200           | –            | 0.29 | -217.6  | 0.00004 | 0.000007 | 0.68  |
| NRand-K 2   | 0.069   | 0.00192 | 161     | 500           | –            | 0.35 | -200.86 | 0.00011 | 0.000013 | 0.80  |
| Pinwheel 1  | 0.056   | 0.00068 | 50      | 198           | –            | 0.32 | -208.6  | 0.00001 | 0.000006 | 0.62  |
| Pinwheel 2  | 0.052   | 0.00096 | 46      | 198           | –            | 0.31 | -210.6  | 0.0    | 0.000091 | 0.60  |
| Pinwheel 3  | 0.078   | 0.00115 | 125     | 494           | –            | 0.39 | -188.8  | 0.00005 | 0.000033 | 0.72  |
| Pinwheel 4  | 0.073   | 0.00119 | 116     | 494           | –            | 0.38 | -191.0  | 0.00008 | 0.000245 | 0.70  |
| VoKA        | 0.073   | 0.00070 | 241     | 1000          | 39           | 0.01 | -305.0  | 0.00037 | 0.010054 | 1.13  |
| K-Means     | 0.070   | 0.00823 | 410     | 1200          | 52           | 0.0  | -307.0  | 0.00092 | 0.000700 | 0.93  |
| 01/2018     |         |       |         |               |              |      |         |        |        |       |
| Donut 1     | 0.086   | 0.00305 | 125     | 200           | –            | 0.38 | -184.1  | 0.00001 | 0.000028 | 0.76  |
| Donut 2     | 0.127   | 0.00415 | 299     | 500           | –            | 0.44 | -164.9  | 0.00011 | 0.000045 | 0.85  |
| NRand-K 1   | 0.061   | 0.00132 | 75      | 200           | –            | 0.29 | -210.0  | 0.00004 | 0.000016 | 0.76  |
| NRand-K 2   | 0.078   | 0.00222 | 165     | 500           | –            | 0.34 | -194.3  | 0.0    | 0.000058 | 0.83  |
| Pinwheel 1  | 0.060   | 0.00282 | 50      | 198           | –            | 0.32 | -200.7  | 0.00002 | 0.000018 | 0.58  |
| Pinwheel 2  | 0.060   | 0.00279 | 46      | 198           | –            | 0.31 | -203.4  | 0.00001 | 0.000100 | 0.56  |
| Pinwheel 3  | 0.083   | 0.00316 | 124     | 497           | –            | 0.38 | -182.5  | 0.00008 | 0.000006 | 0.68  |
| Pinwheel 4  | 0.081   | 0.00321 | 116     | 493           | –            | 0.37 | -185.2  | 0.00005 | 0.000248 | 0.66  |
| VoKA        | 0.075   | 0.00647 | 234     | 1000          | 37           | 0.01 | -293.3  | 0.0    | 0.009746 | 1.05  |
| K-Means     | 0.075   | 0.00934 | 393     | 1200          | 50           | 0.0  | -295.1  | 0.00077 | 0.000993 | 0.94  |

the best value, therefore affecting the less this count, this is explained due to the induced noise which was also smallest in both datasets. The values for the Hdi are smaller for algorithms that do not affect the hotspot count which is also related to the PCdi, hence NRand-K 1 and Pinwheel in the variations of 200m work best. Fig. 6 shows the maps for the hotspot analysis where subfigure (d) is more similar to (a) than the maps of other mechanisms for the reference dataset of arrests in New York in 01/2017.

On a general note, it can be said that for these experiments NRand-K 1 performed best overall and should be the mechanism with lesser impact in the geospatial data. However, this may not always be the case and some indices will have been prioritized to reach a decision.

VI. DISCUSSION

From this analysis, NRand-K with 200m of maximum noise setting is the mechanism that performs best overall followed by Pinwheel in the variations of 200m. The variations of 500m should be evaluated considering that they provide more privacy and therefore may alter more the indices, with this setting of maximum noise, NRand-K 2 and Pinwheel 4 are comparable with the results of VoKa and K-Means. Even if for these experiments, NRand-K yields as the best choice, for another analysis the framework may show with less clarity which mechanism affects less the indices; with different mechanisms obtaining the best scores for different indices. In this case, the analyzer should prioritize which indices are more important to the scope that is being applied; it could be that indices relating to global features are maintained by some mechanisms while local clusters are strongly affected.
or vice versa. It may be that the most suitable mechanism did not score best on none of the indices but rather maintained a average behaviour through them. It is important to realize that there are no absolutes when it comes to evaluating methods for distorting locations, there are however, features that can help indicate how the data is affected and the framework presented in this paper is a means to providing a decision tool in that sense. It is worth noting that in this demonstration, the noise inducing mechanisms in general increment the number of hotspots and therefore clusters, while the aggregation based mechanisms generate different clusters to that of the reference data but not as spread, this is evidence as well as the NNI and Z-Score. A similar behaviour is observed in the difference of heatmaps; with the aggregation based mechanism resulting in different shapes and the noise inducing mechanisms enlarging the heatmap area, decreasing the highest concentration heat zones.

VI. CONCLUSION

This paper presents an evaluation framework for the impact of obfuscation and masking techniques on geospatial data. The framework is composed by a toolkit of indices that estimate different characteristics of the data before and after a geoprivacy mechanism has been applied. The condensation of the framework is presented as a table that enables visualizing the overall behaviour of the evaluated mechanisms to enable the analyzer to make a decision based on the results of the indices.

In the experiments performed, the NRand-K mechanism with 200m of maximum noise resulted in less alterations (better index maps) with respect to the reference data in the evaluated features proposed, it was followed by the Pinwheel algorithm also on a maximum noise setting of 200m. On this particular set of experiments, the framework was able to point toward a single best mechanism instead of different indices marking best for different mechanisms, for this reason it is worth noting that even though exhaustive experiments were carried throughout the paper these are not in itself the main contribution, but a means to prove the way the framework operates and how it can be used with any point based geoprivacy mechanism. In other words, the main conclusion of this work is that the NRand-K mechanism is best when compared with others but the evaluation framework introduced that allows to make a decision, not only with the scenarios covered in this paper but with any other geoprivacy mechanism that outputs point data. For further work, it would be pertinent to include this framework as an automated plugin in an open desktop GIS software for a wider adoption and ease of use.

REFERENCES

[1] A. R. Beresford and F. Stajano, “Mix zones: User privacy in location-aware services,” in Proc. IEEE Annu. Conf. Pervasive Comput. Commun. Workshops. Washington, DC, USA: IEEE Computer Society, Mar. 2004, pp. 127–131.

[2] M. P. Armstrong, G. Rushton, and D. L. Zimmerman, “Geographically masking health data to preserve confidentiality,” Statist. Med., vol. 18, no. 5, pp. 497–525, Mar. 1999.

[3] K. H. Hampton, M. K. Fitch, W. B. Allshouse, I. A. Doherty, D. C. Gesink, P. A. Leone, M. L. Serre, and W. C. Miller, “Mapping health data: Improved privacy protection with donut method geomasking,” Amer. J. Epidemiol., vol. 172, no. 9, pp. 1062–1069, Nov. 2010.

[4] P. Wightman, M. Zurbaran, and A. Santander, “High variability geographrc-K aggregation mechanism for privacy in location-based information systems,” in Proc. 47th Int. Carnahan Conf. Secur. Technol. (ICCST), Oct. 2013, pp. 1–6.

[5] M. Zurbarán and P. Wightman, Privacy in Location Based Services, P. M. Wightman, Ed. LAP LAMBERT Academic Publishing, 2018. [Online]. Available: https://www.lap-publishing.com/

[6] M. Zurbaran and P. Wightman, “VoKa: Voronoi K-aggregation mechanism for privacy in location-based information systems,” in Proc. Int. Carnahan Conf. Secur. Technol. (ICCST), Oct. 2017, pp. 1–6.

[7] J. P. D. Ordonez, A. de Jesus Pugliesi Viloria, and P. W. Rojas, “Comparison of spatial clustering techniques for location privacy,” in Proc. IEEE Latin-Am. Conf. Commun. (LATINCOM), Nov. 2019, pp. 1–6.

[8] NYC Open Data. Accessed: Sep. 2, 2020. [Online]. Available: https://opendata.cityofnewyork.us/

[9] C. D. Cottrill, “Location privacy: Who protects?” J. Urban Regional Inf. Syst. Assoc., vol. 23, no. 2, pp. 49–59, 2011.

[10] J. Krumm, A Survey of Computational Location Privacy, vol. 13, no. 6. Springer, Aug. 2009. [Online]. Available: https://link.springer.com/article/10.1007/s00779-008-0212-5#citeas

[11] S. Gambis, M. O. Killijian, and M. N. del Prado Cortez, “Show me how you move and i will tell you who you are,” Trans. Data Privacy, vol. 4, no. 2, pp. 103–126, 2011.

[12] Y. Zhou, F. Dominici, and T. A. Louis, “A smoothing approach for masking spatial data,” Ann. Appl. Statist., vol. 4, no. 3, pp. 1451–1475, Sep. 2010.

[13] K. N. Truong, Ed. Berlin, Germany: Springer, 2007, pp. 47–60.

[14] M. Duckham and L. Kulik, “A formal model of obfuscation and negotiation for location privacy,” in Pervasive Computing (Lecture Notes in Computer Science), vol. 3468, Berlin, Germany: Springer, 2005, pp. 152–170.

[15] A. Murad, B. Hilton, T. Horan, and J. Tangenberg, “Protecting patient geoprivacy via a triangular displacement geo-masking method,” in Proc. 1st ACM SIGSPATIAL Int. Workshop Privacy Geographic Inf. Collection Anal. (GeoPrivacy). New York, NY, USA: Association Computing Machinery, 2014, pp. 1–9.

[16] C. A. Ardagna, M. Cremonini, E. Damiani, S. D. C. Di Vimercati, and P. Samarati, “Location privacy protection through obfuscation-based techniques,” in Proc. IEEE Annu. Conf. Data Anal. Secur. Privacy, Berlin, Germany: Springer, 2007, pp. 47–60.

[17] J. K. Ord and A. Getis, “Local spatial autocorrelation statistics: Distributional issues and an application,” Geograph. Anal., vol. 27, no. 4, pp. 286–306, Sep. 2010.

[18] D. E. Seidl, P. Jankowski, and M.-H. Tsou, “Privacy and spatial pattern preservation in masked GPS trajectory data,” Int. J. Geograph. Inf. Sci., vol. 30, no. 4, pp. 785–800, Apr. 2016.

[19] M.-P. Kwan, I. Casas, and B. Schmitz, “Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks?” Cartographica, Int. J. Geographic Inf. Geovisualization, vol. 39, no. 2, pp. 15–28, Jun. 2004.

[20] J. Krumm, “Inference attacks on location tracks,” in Pervasive Computing (Lecture Notes in Computer Science), A. LaMarca, M. Langheinrich, and K. N. Truong, Eds. Berlin, Germany: Springer, Jan. 2007, pp. 127–143.

MAYRA A. ZURBARÁN was born in Barranquilla, Colombia, in 1990. She received the B.S. degree in systems engineering and the Ph.D. degree in systems engineering and computing from the Universidad del Norte, Barranquilla, in 2011 and 2018, respectively.

From 2011 to 2012, she was a Developer with IT Research and Consulting. From 2013 to 2015, she was a Research Assistant with the Systems Engineering Department, Universidad del Norte. From 2015 to 2016, she was a Ph.D. Intern with the Environmental and Civil Engineering Department, Politecnico di Milano, Milan, Italy, where she was also a Research Fellow, from 2018 to 2020. Since 2020, she has been a Research Fellow with the Systems Engineering Department, Universidad del Norte. She is the author of two books and 27 articles. Her research interests include location privacy, geographic information systems, and geospatial data analysis.
AUGUSTO SALAZAR was born in Bogotá, Colombia, in 1977. He received the B.S. degree in systems engineering from the Universidad del Norte, Barranquilla, Colombia, and the M.S. degree in computer science from National Chiao Tung University, Hsinchu, Taiwan, in 2012. He was with the Industry on embedded systems for a period of nine years on companies, such as Ericsson LMF (FI), Hitron Technologies (TW), and Proscend Communications (TW). Since 2012, he has been an Assistant Professor with the Department of Systems Engineering, Universidad del Norte, where he covers courses on introduction to programming, computer architecture, embedded systems, and mobile applications. His research interests include embedded system application on robotics, game analytics, use of sensors on ubiquitous, and intelligent systems for the detection of anomalies on the city roads.

MARIA ANTONIA BROVELLI was born in Angera, Italy, in 1961. She received the degree (Hons.) in physics and the Ph.D. degree in geodesy and cartography. From 1997 to 2010, she was the Director of the Geomatics Laboratory, Politecnico di Milano (PoliMI), Milan, Italy, where she was a Vice Rector for the Como Campus, from 2011 to 2016. She is currently a Professor of geographic information systems with the Civil and Environmental Department, PoliMI. She is also a member with the School of Doctoral Studies in Data Science, Roma La Sapienza. She is the author of 108 scientific indexed publications. Her research interest includes geomatics. She is a member of ESA Advisory Committee of Earth Observation (ACEO). She is the Chair of the ISPRS WG IV/4 Collaborative Crowdsourced Cloud Mapping (C3M). She is the Co-Chair of the United Nations Open GIS Initiative, the Deputy-Chair of the UN-Global Geospatial Information Management (GGIM) Academic Network, and a Mentor with the PoliMI Chapter of YouthMappers (PoliMappers). She served as a Guest Editor for 11 Special Issues.

PEDRO M. WIGHTMAN (Senior Member, IEEE) was born in Barranquilla, Colombia, in 1980. He received the B.S. degree in systems engineering from the Universidad del Norte, Barranquilla, and the Ph.D. degree in computer science and engineering from the University of South Florida, Tampa, FL, USA. He is currently an Associate Professor with the Systems Engineering Department, Universidad del Norte. He is also the Director of the Networking and Software Engineering Research Group, GReCIS. His research interests include information privacy, mainly location and healthcare data, industrial applications of multiple realities, blockchain, the Internet of Things, and science fiction. He serves as the Chair for the IEEE Colombian Caribbean Section.