RV Institute of Technology and Management®

Department of Computer Science and Engineering
JP Nagar, Kothanur, Bengaluru - 560076

"Privacy Preservation of Cluster Integrity on Web-Scraped Hospital Data"

Presented by:
Hrishik Sai Bojnal - 1RF21CS052
Dharmisht SVK - 1RF21CS035
J Krishna Kaarthik - 1RF21CS055
Dhyaan Kotian - 1RF21CS039

Under the Guidance of:
Dr. Shashidhar V
Assistant Professor
Dept. of CSE
RVITM

RV Institute of Technology and Management®

Contents

1. Introduction

1. Background

- **Current Situation:**
  - The exponential growth of web technologies has revolutionized data collection in diverse fields, with healthcare being a significant beneficiary.
  - Online platforms provide unprecedented access to information about healthcare facilities, such as their locations, ratings, and patient reviews, enabling detailed analyses of healthcare accessibility and quality.
  - However, the rise of such datasets introduces critical challenges, especially concerning the privacy of sensitive geospatial information.
- **Urgency of Privacy Concerns:**
  - Geospatial and demographic data, while invaluable for research, can inadvertently expose sensitive information, leading to privacy breaches.
  - Current anonymization techniques often compromise the utility of data, rendering it less effective for actionable insights.

1.2. Objective

Main Objective:

"To develop an anonymization method that preserves the natural spatial clustering of healthcare facilities while ensuring privacy and maintaining the integrity of geospatial datasets."

Specific Objectives:

- Extract and compile a comprehensive geospatial dataset of healthcare facilities, including location, ratings, and review counts.
- Calculate the regional influence of each facility based on prominence and proximity.
- Implement a cluster-preserving anonymization technique that restricts displacements to within minimal cluster radii.

2. Methodology

Step 1: Dataset Acquisition

- Extract initial data from NIT Jalandhar's published hospital list.
- Use Selenium with Microsoft Edge to scrape additional data (ratings, review counts, coordinates) from an online maps platform.
- Handle discrepancies by considering the first search result, with negligible impact from rare errors.

Step 2: Data Cleaning

- Backfill missing geospatial data based on proximity to other hospitals.
- Fill missing ratings and review counts with median values to maintain dataset integrity.

2. Methodology

Step 3: Data Modeling

•Compute the effective rating $\eta=\alpha \cdot R \cdot \log(N)$ where R is the rating, N is the number of reviews, and α normalizes the influence scale.

•Plot a bubble chart using geospatial coordinates, with bubble size proportional to the effective rating.

Step 4: Data Representation

- Map hospital influences on a 1x1 km grid using their coordinates.
- Calculate influence per grid cell as $i = a\, e^{-\lambda r}$, where r is the distance to the hospital, and a, λ are scaling constants.

RV Institute of Technology and Management®

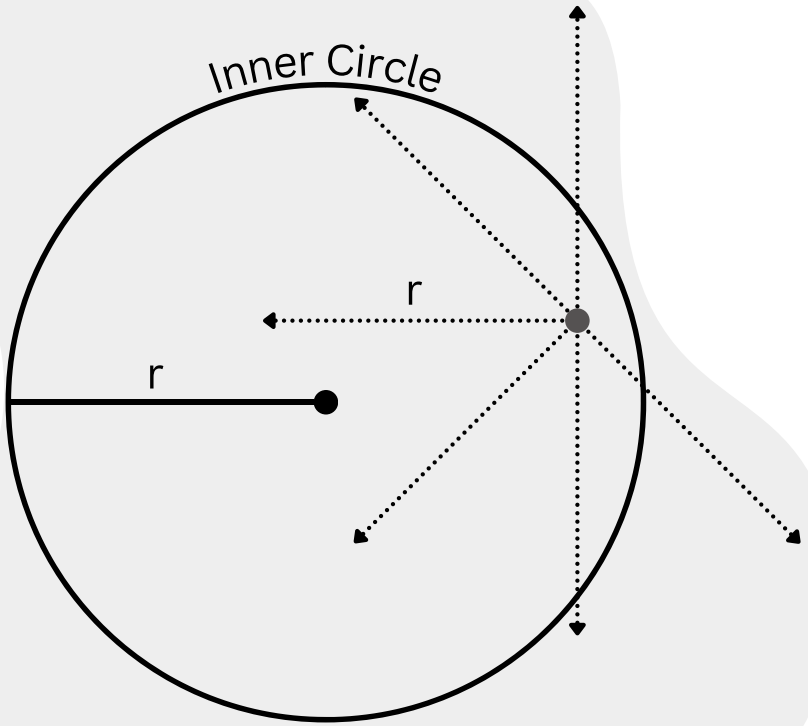


Fig. 1. A typical cluster is shown here, along with its incircle of radius r. The dotted lines, each also measuring r, represent the valid displacements of a sample point within the cluster.

2. Methodology

Step 5: Cluster Identification

•Filter significant influence values (50th percentile and above).

•Define clusters as groups of 8-adjacent cells with non-zero influence values.

•Use the convex hull to identify cluster boundaries and calculate cluster magnitudes by summing influence values.

Step 6: Cluster-Preserving Anonymization

•Compute the centroid and inner circle radius for each cluster using the Quickhull algorithm.

•Displace hospital locations within the cluster, scaling displacement inversely by the square of the hospital's effective rating.

•Ensure displacements stay within cluster boundaries.

2. Methodology

Step 7: Random Anonymization for Comparison

- Add Gaussian noise to coordinates using mean and standard deviation based on cluster inner radii.
- Assess dataset quality deterioration compared to cluster-preserving anonymization.

Step 8: Full Anonymization

- Obfuscate hospital names.
- Add Gaussian noise to ratings and adjust review counts to maintain effective rating consistency.

Step 9: Evaluation of Anonymization

- Compare data signatures (cluster magnitude histograms) of original, cluster-preserving, and randomly anonymized datasets visually and quantitatively using Earth Mover's Distance (EMD).

RV Institute of Technology and Management®

3. Results

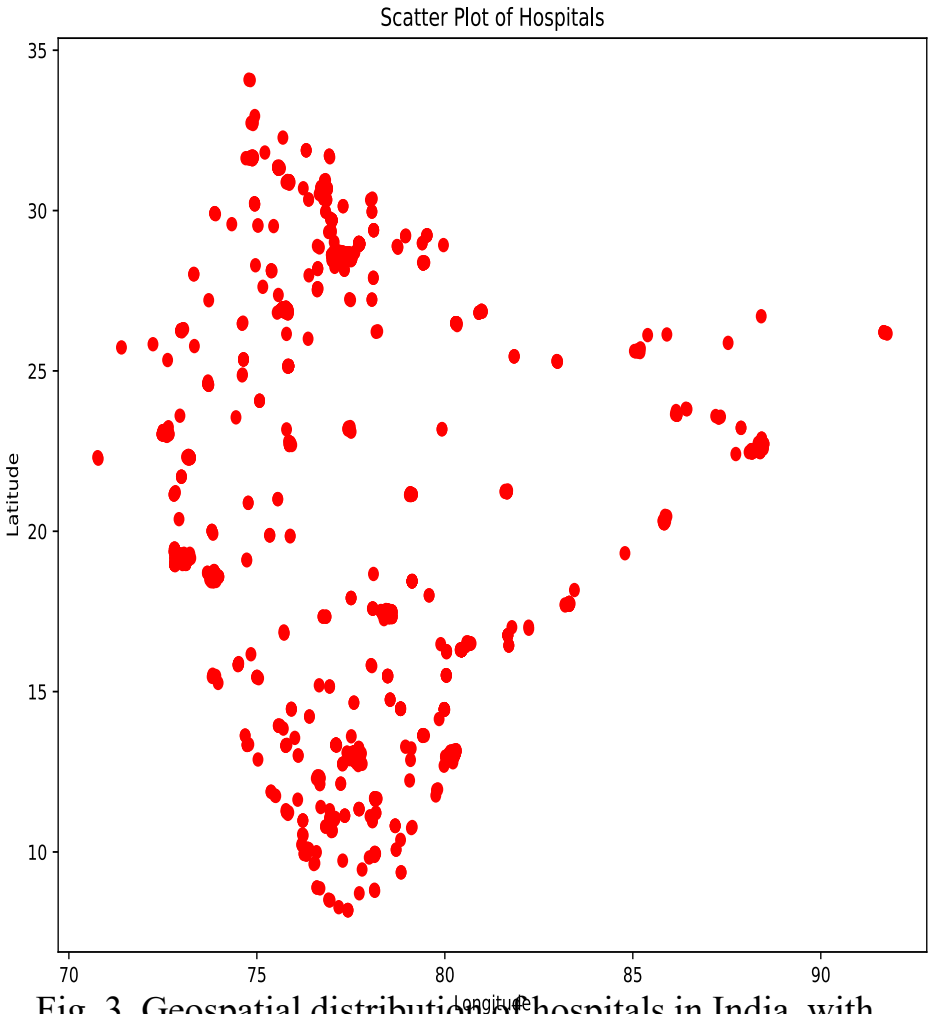


Fig. 3. Geospatial distribution of hospitals in India, with notable clusters in major metropolitan areas and regional population centers.

Scatter Plot of the scraped hospitals:

The scatter plot visually represents the coordinates of scraped hospital data. Due to overlapping, some hospital locations may eclipse others. This plot does not indicate the total number of hospitals in the country but rather shows the distribution across major metropolitan areas like Delhi, Kolkata, Mumbai, Hyderabad, Bengaluru, and Chennai, as well as regional population centers.

The clustering in these major areas suggests higher hospital density and might reflect accessibility and availability of healthcare services. Regional population centers show a similar trend, albeit at a smaller scale.

3. Results

Bubble Plot from the Influences

Each hospital is represented by a bubble in the plot, with the radius indicating the hospital's influence. Larger bubbles correspond to hospitals with a higher effective rating, signifying greater influence in their region. The density and saturation of each bubble's color reflect the proximity of hospital locations within the dataset.

The positioning of these bubbles shows a more pronounced concentration of hospital influence in certain areas and sparser coverage in others. This visual helps to quickly assess the areas with higher healthcare access and the gaps in regions with fewer hospitals.

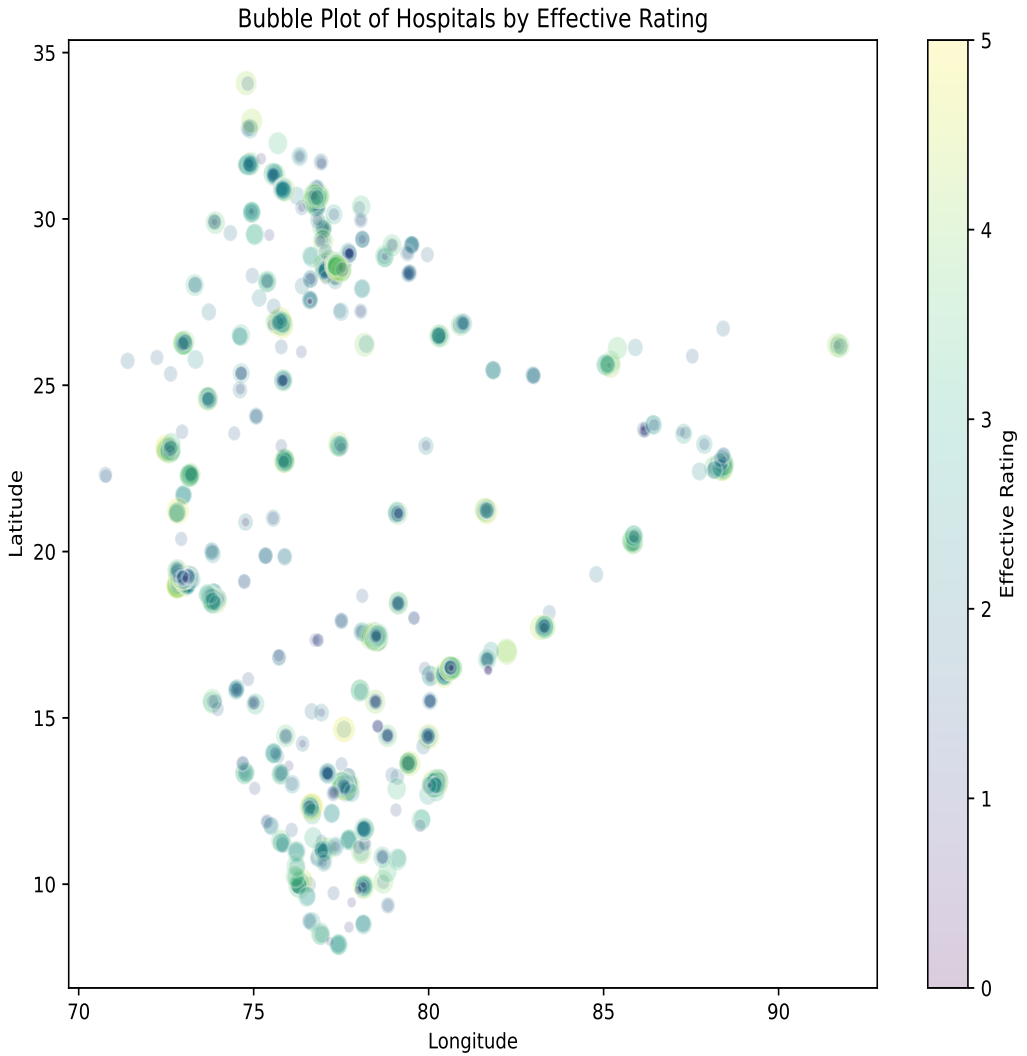


Fig. 4. Bubble plot of hospitals, where bubble size represents effective rating and density illustrates regional healthcare presence.

3. Results

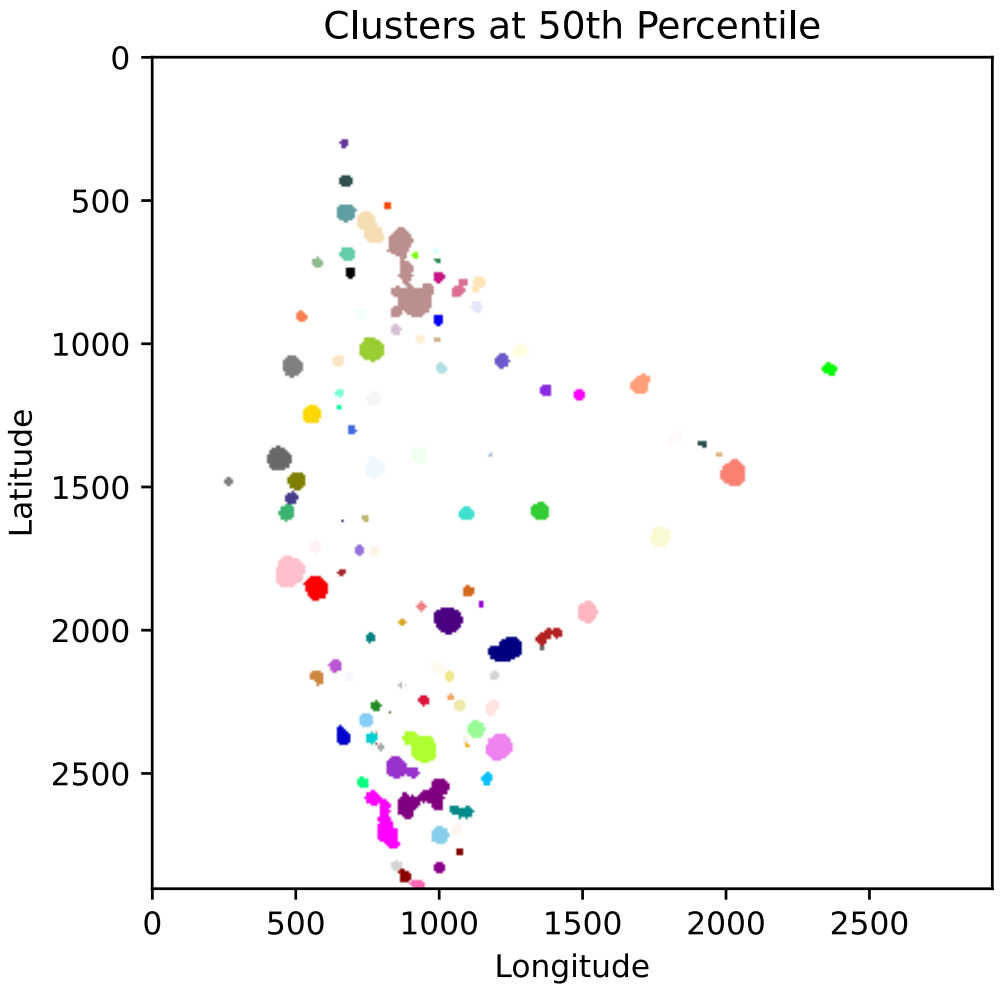


Fig. 5. Clusters representing the top 50% of hospital influences by effective rating

Cluster Map

This map visualizes clusters of hospital influences based on effective ratings. Only the top 50% of non-zero hospital influences are considered, providing a clearer picture of where hospital impacts are concentrated geographically. Each cluster is shown as a distinct region, formed around the concentration of points, indicating areas of high influence.

By considering only the upper half of the values, the map ensures that the clusters reflect denser zones of hospital influences more accurately. This comparison is essential to evaluating the accuracy of the anonymization techniques applied, as it highlights how well the original data structure is preserved.

3. Results

Cluster-Preserving Anonymization

Using the cluster-preserving anonymization method, hospital influence clusters are slightly displaced, but their original structures are maintained. This technique displaces cells within the cluster by the radius r of the incircle of the convex hull. This ensures that the overall layout and density of hospital locations remain intact, preserving the natural grouping found in the original data.

When comparing this graph to the unaltered data in the previous figure, minimal changes are visible, suggesting that the cluster-preserving anonymization is effective in maintaining the distribution of hospital influences without significantly altering their spatial relationships..

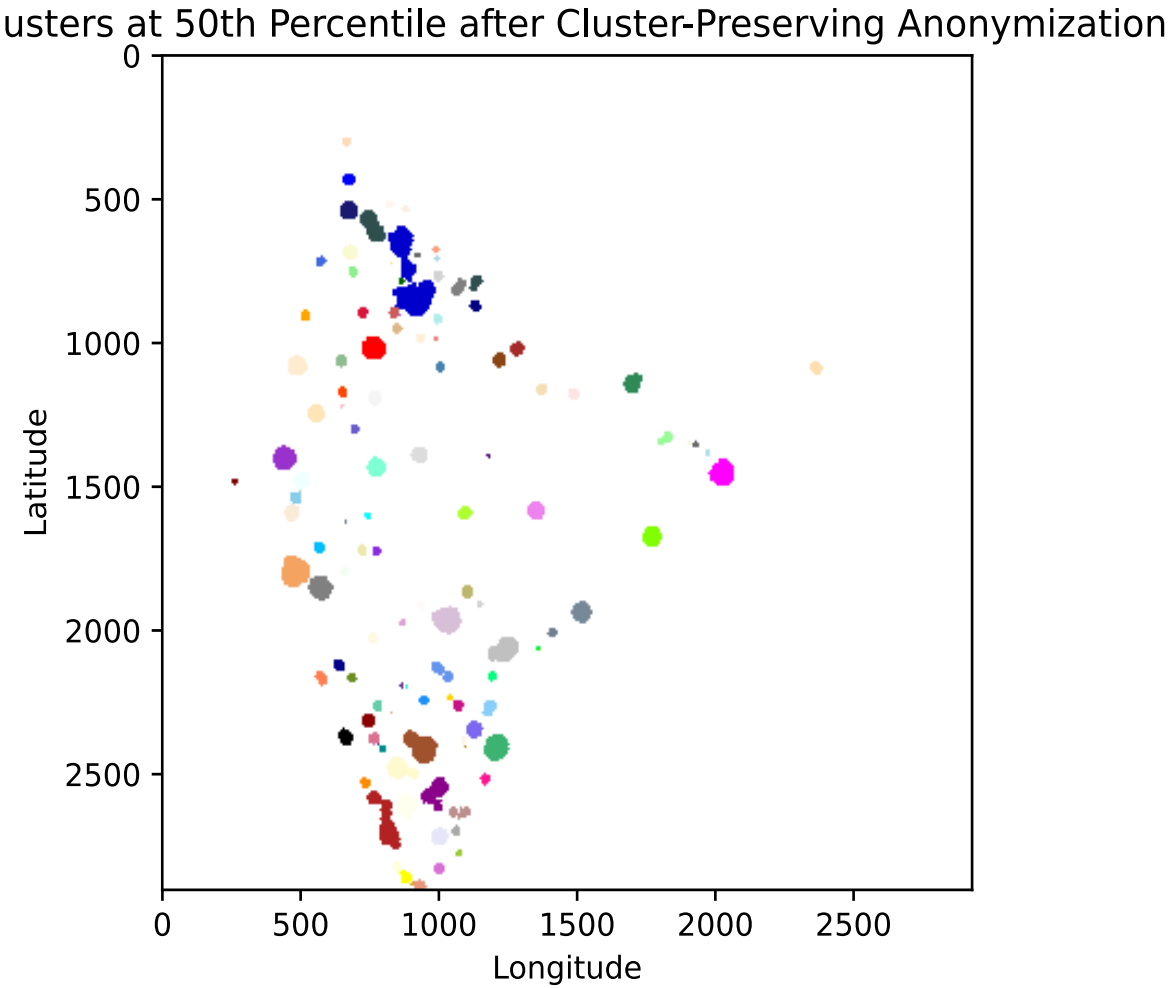


Fig. 6. Hospital-influence clusters after applying clustered anonymization

3. Results

Random Noise Addition

Random noise is added to the original dataset to anonymize the hospital locations, disrupting the preservation of original cluster structures. This method introduces variability to the geospatial coordinates of hospitals within each cluster, making it visually distinct from the original dataset. In particular, the Deccan region shows noticeable differences in scale and structure when compared to the original data, which suggests a loss of spatial integrity due to noise addition.

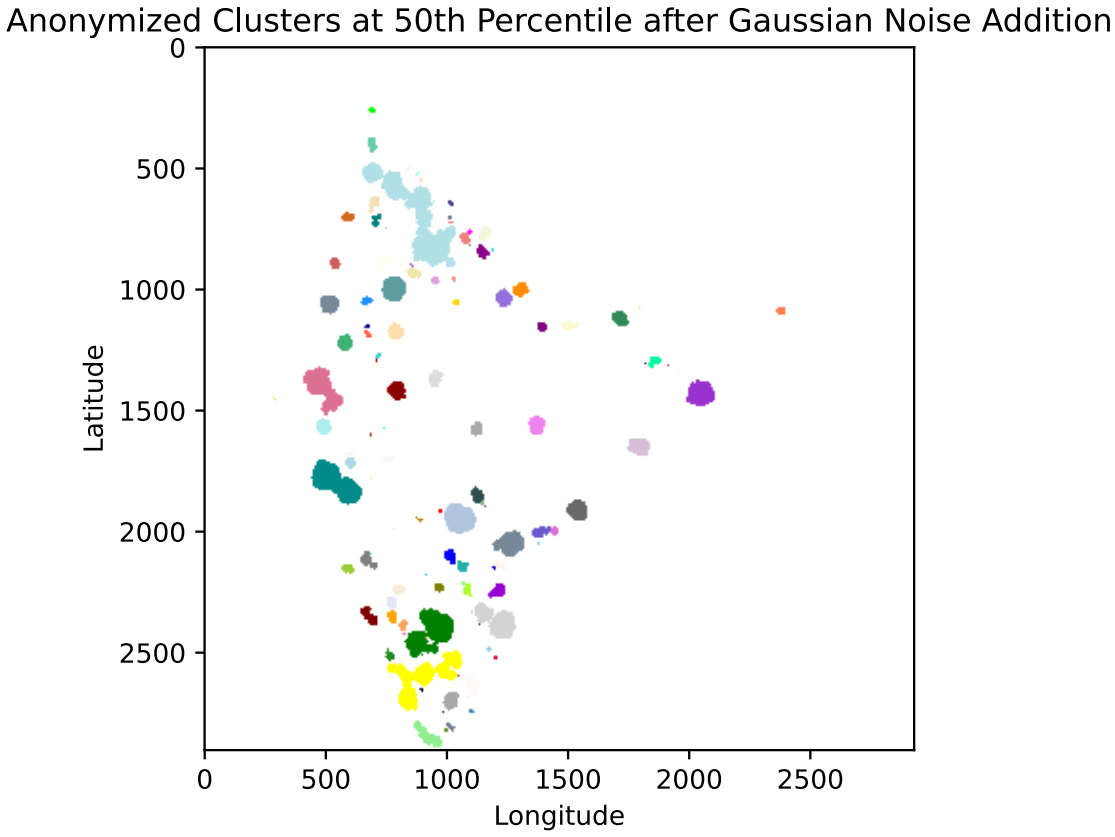
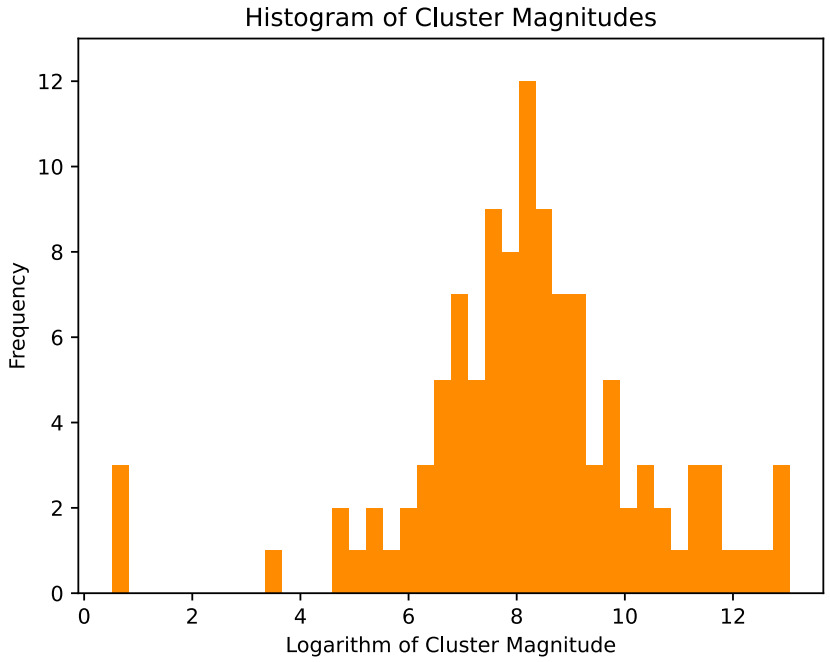


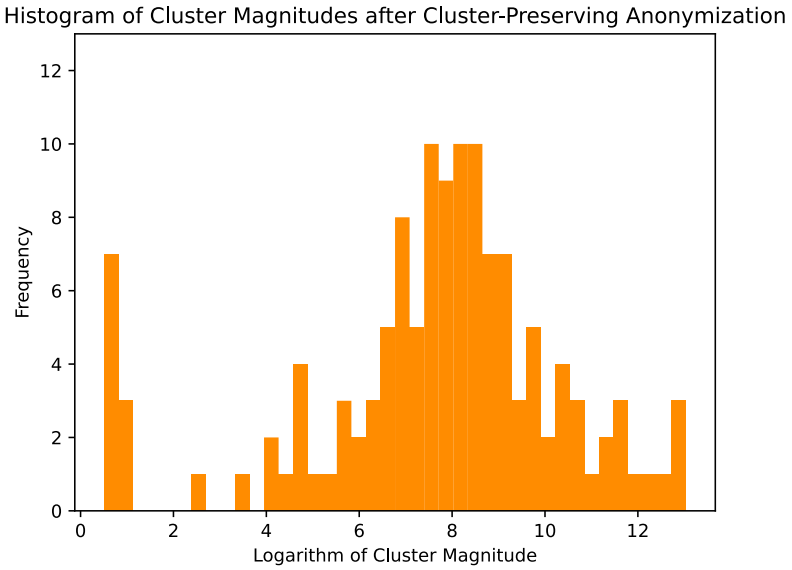
Fig. 8. Hospital-influence clusters after random noise addition,

RV Institute of Technology and Management®

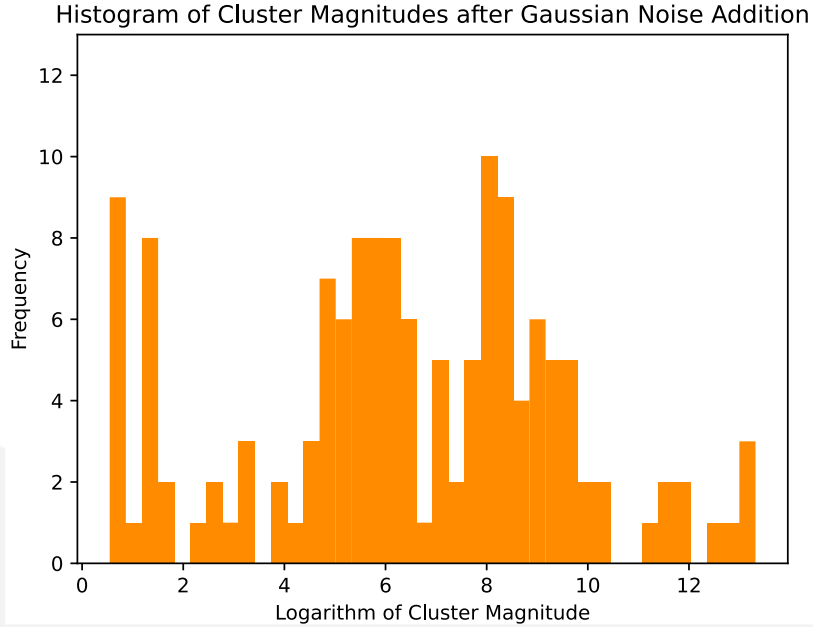
3. Results



Signature of the Original Dataset:
•Displays a central peak indicating the concentration of hospitals with similar effective ratings.
•Tails reflect the presence of hospitals with either low or high effective ratings, highlighting diversity in hospital influence.



Cluster-based Anonymized Signature:
•Retains a similar spread and central peak height as the original dataset, indicating effective preservation of distributional characteristics.
•The Earth Mover's Distance (EMD) of 0.4002 suggests close similarity in distribution to the original dataset.



Naive Anonymized Signature:
•Shows a significantly larger spread compared to the original dataset, indicating a reduction in central concentration.
•The Earth Mover's Distance (EMD) of 1.3015 highlights the loss of spatial integrity compared to cluster-based anonymization.

3. Results

Anonymization Technique	Earth Mover's Distance (EMD)
Original Dataset	-
Cluster-based Anonymization	0.4002
Naive Anonymization	1.3015

4. Conclusion

- In this study, we collected a comprehensive dataset of hospitals in India by scraping online map platforms.
- We gathered key details such as hospital ratings, number of reviews, and exact locations.
- Both random noise addition and our anonymization technique were evaluated to balance data utility and privacy.
- We found that clustering-based anonymization is superior, preserving natural patterns and density.
- Compared to naive anonymization, clustering maintains the layout and frequency of hospital locations.
- It is evident that this method retains the dataset's spatial structure while ensuring privacy.

RV Institute of Technology
and Management®

Thank You!