Comparing Label Local Differential Privacy with Blurred Geospatial Data

Student

André von Aarburg

Advisor
Prof. Stefan F. Keller

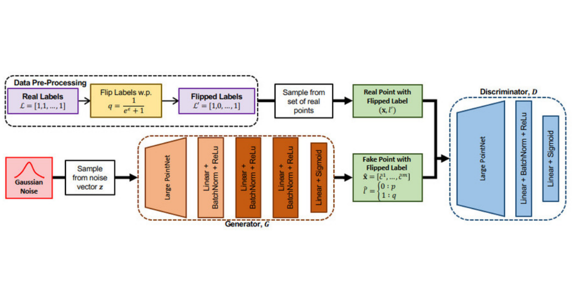
Subject Area
Data Science

Problem: This project addresses growing privacy concerns in geospatial data by developing and evaluating privacy-preserving synthetic data generators. With the widespread use of location-based services, the sharing of raw geospatial data can lead to privacy breaches and expose sensitive information about individuals' locations.

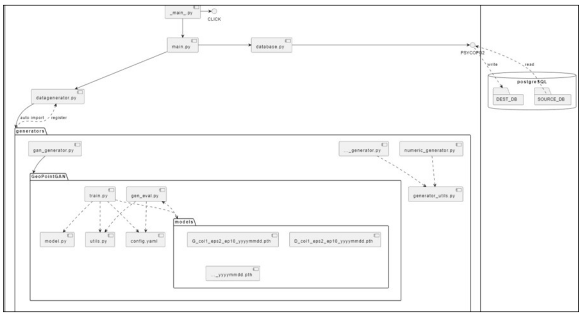
Approach: To address these challenges, two synthetic data generators are introduced: the GAN generator and the blurring generator. The GAN generator is based on the GeoPointGAN model of Klemmer et al. (2022). By using differential privacy mechanisms, this generator ensures privacy while producing synthetic data that closely matches the original data distribution. On the other hand, the blurring generator uses blurring techniques to perturb the geographic coordinates of the data points. This approach provides different levels of privacy based on a blur radius parameter. Both generators offer trade-offs between privacy and data utility. To evaluate the quality of the generated synthetic data, the project uses a set of evaluation metrics, including Earth-Mover-Distance, Chamfer-Distance, Mean Absolute Error of Range Queries, and Sørensen Dice Coefficient of Hotspot Analysis. These metrics compare the performance of the generators in preserving data distribution and supporting geospatial analysis tasks.

Result: To make the tools available, they are implemented in the existing pgsynthdata tool. pgsynthdata is a command-line tool for creating synthetic data for PostgreSQL databases. The two implemented generator extensions can be customized by the user via YAML configuration files. To execute pgsynthdata, the usual commands are used, in addition, the corresponding evaluation methods can be displayed for the evaluation of the results. The generators are tested and evaluated on a database of well coordinates. In comparison, the GAN generator shows more stable behavior in terms of preserving statistical values against changing privacy settings. The results demonstrate the importance of finding a balance between privacy and data utility, and provide guidance for future advances in privacy and geospatial computing.

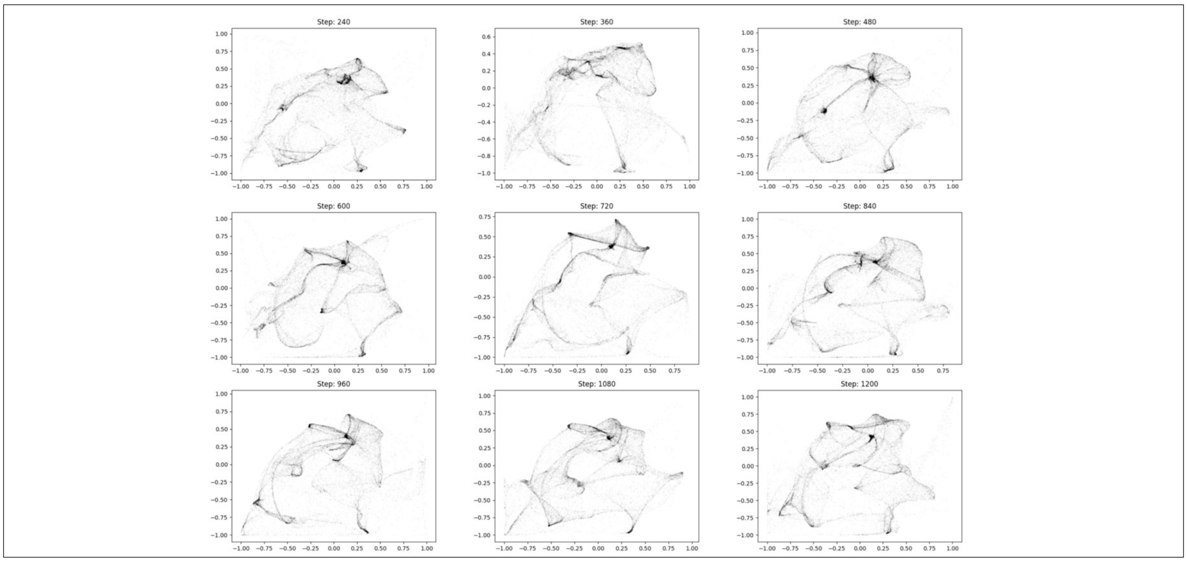
Data flow and model sequence of GeoPointGAN synthetic spatial data generator
From Klemmer et al. (2022)


Data Pre-Processing
Real Labels
$\mathcal{L} = [1,1,\dots,1]$
Flip Labels w.p.
$q = \frac{1}{e^\epsilon + 1}$
Flipped Labels
$\mathcal{L}' = [1,0,\dots,1]$
Sample from set of real points
Real Point with Flipped Label
$(\mathbf{x}, l')$
Gaussian Noise
Sample from noise vector z
Large PointNet
Linear + BatchNorm + ReLu
Linear + BatchNorm + ReLu
Linear + BatchNorm + ReLu
Linear + Sigmoid
Generator, G
Fake Point with Flipped Label
$\hat{\mathbf{x}} = [\hat{c}^1,\dots,\hat{c}^m]$
$\hat{l}' = \begin{cases} 0 : p \\ 1 : q \end{cases}$
Discriminator, D
Large PointNet
Linear + BatchNorm + ReLu
Linear + Sigmoid

UML Component Diagram of the GAN synthetic spatial data generator
Adapted from Erhart & Elmer (2021)



Training output of the GAN generator from synthetic well coordinates (epochs two to ten, with a privacy budget of 0.5)
Own presentment


Step: 240
Step: 360
Step: 480
Step: 600
Step: 720
Step: 840
Step: 960
Step: 1080
Step: 1200