Geographically Masking Addresses to Study COVID-19 Clusters

Walid Houfah-Khoufah

Universite Gustave Eiffel, ENSG, IGN

Guillaume Touya (✉ guillaume.touya@ign.fr)

Institut national de l'information géographique et forestière <https://orcid.org/0000-0001-6113-6903>

Research

Keywords: COVID-19, spatial anonymisation, address, k-anonymity

Posted Date: December 16th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-128679/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: The spatio-temporal analysis of cases is a good way to study an epidemic, and the recent COVID-19 pandemic unfortunately generated a huge amount of data. But analysing this raw data, with for instance the address of the people who contracted COVID-19, raises some privacy issues, and geomasking is necessary to

preserve both people privacy and the spatial accuracy required for analysis. This paper proposes different geomasking techniques adapted to this COVID-19 data.

Methods: Different techniques are adapted from the literature, and tested on a synthetic dataset mimicking the COVID-19 spatio-temporal spreading in Paris and a more rural nearby region. These techniques are assessed in terms of k-anonymity and cluster preservation.

Results: Three adapted geomasking techniques are proposed: aggregation, bimodal gaussian perturbation, and simulated crowding. All three can be useful in different use cases, but the bimodal gaussian perturbation is the overall best technique, and the simulated crowding is the most promising one, provided some improvements are introduced to avoid points with a low k-anonymity.

Conclusions: It is possible to use geomasking techniques on addresses of people who caught COVID-19, while preserving the important spatial patterns.

Full Text

This preprint is available for [download as a PDF](#).

Figures

Figure 1

The generated synthetic COVID-19 data, for Paris on the left, and Yvelines on the right, both for three months.

Figure 2

A geomasked address point has k -anonymity if at least $k-1$ address points are closer to this address than the initial position of the address point (Figure 2).

Figure 3

The census cells used to aggregate the address points in Paris (on the left), and the Yvelines (on the right). The points show the centroid of each cell, where the address are displaced when aggregated to the cell.

Figure 4

The blocks computed using the road network graph. There are much more cells than with the census.

Figure 5

Principles of buildings aggregates based on the dilation of building polygons.

Figure 6

Results obtained with the enhanced bimodal gaussian perturbation in Paris.

Figure 7

Mean k-anonymity in the census cells of Paris after an enhanced bimodal gaussian perturbation. The blue cells did not contain any (fake) COVID-19 case to geomask.

Figure 8

Principles of simulated crowding, new points are generated randomly in the clusters to guarantee cluster preservation.

Figure 9

The extent of a cluster and the area covered by the buffer around the initial points (20m buffer on the left, 30m buffer on the right).

Figure 10

When the buffer size is increased to 40m, the remaining area to generate masked points is too small and fragmented.

Figure 11

Results of the simulated crowding method with a 30m buffer. The empty spaces where points are grouped often correspond to the interior of blocks, while address points are generally located along roads.

Figure 12

Difference between the enhanced bimodal gaussian perturbation, and the simulated crowding at the border of Paris. The perturbation pushes the points inside the city, while simulated crowding uses empty spaces at the border

Figure 13

Results of the enhanced bimodal gaussian perturbation, and the cluster oriented simulated crowding on the same extract of the Paris area.