Geographically masking addresses to study COVID-19 clusters

Walid Houfaf-Khoufaf\textsuperscript{a}, Guillaume Touya\textsuperscript{b} and Arnaud Le Guilcher\textsuperscript{b}

\textsuperscript{a}ENSG, IGN, Université Gustave Eiffel, France; \textsuperscript{b}LASTIG, ENSG, IGN, Université Gustave Eiffel, France

\section*{ABSTRACT}

The spatial analysis of health data usually raises geoprivacy issues. Due to the virulence of COVID-19, scientists and crisis managers do need to analyze the distribution and spread of the disease with spatially precise data. In particular, it is useful to locate each case on a map to identify clusters of cases. To allow such analyses without breach of geoprivacy, geomasking techniques are necessary. This paper experiments with the geomasking techniques from the literature to solve this problem: masking the real address of positive cases while preserving the local spatial cluster patterns. In particular, two different approaches based on aggregation and perturbation are adapted to the geomasking of addresses in areas with different densities of population. A new simulated cluster crowding method is also proposed to preserve clusters as much as possible. The results show that geomasking techniques can spatially anonymize addresses while preserving clusters, and the best geomasking method depends on the use of the anonymized data.

\section*{1. Introduction}

With the unprecedented impact of the COVID-19 pandemic, clusters of cases have become a crucial issue. Finding these clusters as quickly as possible is important to understand the virus transmission (Yong et al., 2020), and for health policies that try to stop the pandemic (Danis et al., 2020). Since the times of the John Snow map, cartography and spatial analysis have been useful tools to find epidemic clusters, or for epidemiology in general (Kirby et al., 2017), and it is also the case for COVID-19, as large scale clusters can be detected by spatio-temporal analyses of the cases (Desjardins et al., 2020). Geo-visual analysis can also be used to understand the epidemic (Delmelle et al., 2014), or to simulate the spatial infection (Chen et al., 2020).

We are participating in a local initiative to track and break the clusters at a very local scale (a street or a neighborhood), following strategies that proved efficient against cholera (Piarroux, 2019). We want to provide both spatial analysis and geovisualization tools to help the epidemiologists that lead this initiative. However, tracking and visualizing the addresses of people tested for COVID-19 raises severe privacy issues. Privacy problems are inherent to health spatial data (Sherman & Fetters, 2007), and even more prominent when sciences try to be reproducible (Ajayakumar et al., 2019). A recent study shows that visualization with a high level of detail (e.g. with point symbols for addresses) is perceived as riskier for geoprivacy than heat maps for instance, (Kim et al., 2021).

Geomasking the addresses is a way to use this data to find COVID-19 clusters without uncovering the real addresses of people registered in this dataset. Geomasking is not new and several interesting techniques have been proposed in the literature. Do the geomasking techniques from the literature preserve address privacy while preserving the spatial properties of COVID-19 clusters? Is it possible to design a new technique more adapted to our use? And can these geomasking techniques handle areas with high density, such as the city of Paris where the first wave of the epidemic was extremely severe, and rural areas where there are isolated dwellings? In very dense areas, the difficulty is cluster preservation as they can be spatially small; in rural areas, the difficulty is the low density of households. To answer these questions, this paper reports experiments to compare and adapt existing geomasking techniques that proved successful for other kinds of health data, with simulated COVID-19 data from Paris and a less dense surrounding region.

\section*{2. Related work}

Geospatial health data are mainly data related to people, so privacy issues are inherent to such data (Sherman & Fetters, 2007). The privacy issue is important when
location is encoded with addresses, but can be even more prominent when phone tracks are used to understand the places responsible for a cluster (Chang et al., 2020).

The debates on the applications tracking the users to avoid SARS-Cov-2 spreading show that the right to location privacy is now well acknowledged in many countries. In Europe, the General Data Protection Regulation (GDPR) protects personal data (Georgiadou et al., 2019), and location information can cause location-based spam, attacks on personal safety, or intrusive inferences (Duckham & Kulik, 2006). When location and specifically the address is attached to health data, it can reveal even more important personal information. The exact address can be re-engineered from roughly masked spatial data, either by light field surveys (Curtis et al., 2006), or by automated methods (Cassa et al., 2008). False identifications are also a risk of attack on location data (Seidl et al., 2018) and can be avoided by assigning unrealistic masked locations (Swanlund et al., 2020).

There are different ways to prevent these types of attacks and preserve the privacy of the people included in a geospatial health dataset (Katsomallas et al., 2019), and the one that is the most interesting in our COVID-19 use case is geomasking or spatial anonymization. In the past twenty years, scholars have proposed different types of geomasking techniques. There are global transformations of the dataset such as an affine transformation (Armstrong et al., 1999), or location swapping (Zhang et al., 2017), which is not relevant for our use case because there is no additional information attached to the address. There are also different techniques that apply a perturbation in the location of each point: a random perturbation (Armstrong et al., 1999); the donut method where the point is randomly displaced in a donut around its initial position (Hampton et al., 2010); a Gaussian perturbation (Zandbergen, 2014) that can apply to both previous methods; a perturbation donut where some specific areas are excluded (Lu et al., 2012); or a perturbation where points are snapped to a road junction close to the initial location (Swanlund et al., 2020). There are also geographical aggregation techniques, based on a Voronoï diagram (Seidl et al., 2015), a predefined grid (Seidl et al., 2016), or the military grid reference system (Clarke, 2016). Point data can also be geomasked by generating a heat map, with a kernel density estimation for instance, (Z. Wang et al., 2019). These techniques generally mask location very well, at the expense of the accuracy of the aggregated data. Finally, simulated crowding is a technique that focuses on the future use of the geomasked data, to guarantee this use is still possible with good accuracy (Scheider et al., 2020).

These geomasking techniques are sensitive to population density heterogeneity. For instance, with the donut method, the distance values that define the donut cannot be optimal for both urban and rural areas (Allshouse et al., 2010). Some geomasking techniques were specifically proposed to mitigate this density heterogeneity, e.g. adaptive areal elimination (Kounadi & Leitner, 2016), and adaptive areal masking (Charleux & Schofield, 2020).

Other past research on geomasking focused more on the evaluation and comparison of geomasking methods, rather than on new techniques. Geomasking is always a tradeoff between protection and accuracy of the masked data (Gao et al., 2019; Kwan et al., 2004). A recent study measured how much spatial distributions and patterns can be modified by different geomasking techniques (Broen et al., 2021). Geo-indistinguishability is an interesting notion to measure this tradeoff (Andrés et al., 2013). Many metrics have been used in the literature to measure how much information can be lost during geomasking. Beyond a basic analysis of point displacement, we can measure how much spatial distribution has been preserved by the geomasking technique. Several measures of spatial distribution have been proposed: a nearest neighbor analysis, kernel density estimation (Charleux & Schofield, 2020; Kounadi & Leitner, 2016), Hotspot’s Divergence (Kounadi & Leitner, 2016), or Moran’s I (Broen et al., 2021) that measures spatial autocorrelation. Regarding the preservation of clusters, the cluster specificity (Cassa et al., 2006; Hampton et al., 2010), or the Ripley’s K function (Swanlund et al., 2020) measure how much the clusters after geomasking are similar to the clusters before geomasking. Regarding geographic consistency, we can measure the preservation of the access to roads for addresses (Charleux & Schofield, 2020), or the landcover agreement (Swanlund et al., 2020; Zhang et al., 2017), i.e. if the geomasked location is in the same type of landcover as the initial location.

Privacy issues are not restricted to geospatial health, and as a consequence, geomasking techniques were developed for other types of geospatial data: for instance, Twitter (Gao et al., 2019), GPS tracks (Scheider et al., 2020; J. Wang & Kwan, 2020), or crime data (Z. Wang et al., 2019). These methods can be useful also for health data, and the simulated crowding is adapted here in Section 4.3.

3. Description of the use case

3.1. Dataset

The French health authorities collect data related to COVID-19 in a central dataset called SI-DEP. This dataset only contains addresses for each person testing
positive to one of the tests for COVID-19, and the date of the test. Personally identifying information is not stored in the dataset, and as a consequence, does not require any anonymization. As the health authorities do not know how to geomask this dataset yet, it is only released aggregated at the region level (the French “départements”), even for epidemiologists. The aggregated data is clearly not a good scale to identify clusters at the scale of a street, or of a neighborhood. When designing a geomasking technique for geospatial health data, the geographical information scientist can be caught in a conundrum as they cannot access the data they want to mask before knowing how to mask them. This is why we needed to demonstrate the efficiency of geomasking on synthetic data.

First, as the text of the complete address contained in the SI-DEP dataset can easily be geocoded to geographic coordinates, we decided to anonymize the addresses as 2D points. To generate the synthetic dataset, we used the open address dataset proposed by the French administrations¹ to force the synthetic data to look like a disaggregation of the official data.² In this dataset, we have every day, for each region, the number of new positive tests, and the number of people tested. A simple way to disaggregate this data is, for every day from the beginning of the epidemic, to randomly select \( n \) address points from the address dataset, where \( n \) is the real number of positive tests in the given region. Obviously, this method does not guarantee clusters with the expected characteristics, and rather favors an even spatial distribution over space, without connection between cases. This is why we adopted a more sophisticated approach, based on the incidence rate of the disease, which is available for each day and region, in the official data: at a given date \( n \), and a given incidence rate \( r_0 \), for each case of date \( n - 4 \), we pick several incidental cases with a Gaussian probability centered on \( r_0 \). For instance, for day with a \( r_0 \), there is a high probability that each case generates two incidental cases, but some generate only one or even none. The incidental cases are randomly selected in the address points in a radius of 300 m (this value was empirically chosen as it gave a realistic distribution according to people having access to the real data). The datasets generated with this method were considered realistic by epidemiologists having access to the real SI-DEP data, with clusters appearing over space and time. It was just a visual check that the generated clusters had the expected dimensions in space and time, but it was enough to pursue the geomasking experiments with the synthetic data. We generated two datasets for the regions of Paris (very dense, 8921 points) and Yvelines (rural on its western part, 3001 points) using this method, with a timespan of three months. These three months correspond to the complete timespan available at the time of the experiment, and are enough to see spatio-temporal COVID-19 clusters appear and disappear. The Paris area was chosen because of the large number of cases, and the density that makes very local clusters, at the scale of a street or a neighborhood. Then, the Yvelines area was chosen in contrast, because of heterogeneous densities across the area, and a significant number of cases compared to other rural areas in France. We only focus on the spatial dimension of these data, we decided not to anonymize the date of the case, as it is already fuzzy because people are sometimes tested at the beginning of the disease, and sometimes at the end. And we make the assumption that a 2D geomasking technique that preserves spatial clusters, all preserves to some extent the spatio-temporal clusters our users are interested in.

### 3.2. Evaluation of the anonymization

Usually, in geomasking research, and more generally in data anonymization research, the quality of masking is evaluated with k-anonymity (Sweeney, 2002). A masked dataset has k-anonymity, if for each of its elements, it cannot be distinguished from at least k-1 other elements of the masked dataset. This concept was designed to evaluate the anonymization of field-structured data, but can be extended to spatial data. In this research, we use two different definitions of spatial k-anonymity, depending on the method we study. For some methods, i.e. aggregation-oriented methods, and the simulated cluster crowding, the usual definition of k-anonymity is applicable: k-anonymity for each masked case in an area is equal to the number of real addresses in this area. However, k-anonymity is not self-sufficient to describe how privacy is protected in this case. If two different areas both contain 100 real addresses, but one contains 10 COVID cases to mask, while the other contains 50 cases, they have the same k-anonymity equal to 100. But privacy is less protected in the second cluster because the probability of one of the real addresses is hosting a COVID case is higher. This is why we introduce a privacy ratio \( R_p \) and a Privacy Guarantee \( G_p \) computed as proposed in Equations (1) and (2).

\[
R_p = \frac{n_mask}{n} 
\]

(1)

\[
G_p = \frac{1}{R_p} 
\]

(2)
where \( n_{mask} \) is the number of COVID-19 cases to mask in the area where aggregation or simulated crowding is performed, and \( n \) is the total number of real addresses in the same area. In the example given above, \( R_p \) is 0.1 in the first area, and 0.5 in the second area.

If we put aside the distinction between cases and addresses, and thus the fact that more than one case can be associated with a single address, \( R_p \) is the probability that a given address has a positive case, from the point of view of someone who knows only the anonymized dataset. As one address can have more than one case, \( R_p \) is in fact an upper boundary of this probability. In our analysis, we use both the value of \( k \), which enables an homogeneous comparison with other methods, and the value of \( G_p \), which acts more like a privacy guarantee, and is more comparable to values of \( k \) computed for displacements methods.

For the other methods, i.e. displacement methods, we use a definition of spatial k-anonymity similar to the one proposed in J. Wang & Kwan (2020): 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 1).

As the main use case of this data is the observation and analysis of spatio-temporal very local to large size clusters, the geomasking techniques have to be evaluated for their ability to preserve these clusters while masking the real addresses. We used the classical DBSCAN method (Ester et al., 1996) to find spatial clusters in our generated datasets, and then we used it once again to compute clusters in the geomasked datasets. Then, the ability to preserve clusters is computed by comparing both sets of clusters. We used the intersection over union (IoU) to measure the similarity between two clusters of address points (Jaccard, 1901). Equation (3) gives the value of IoU for two clusters A and B, and IoU is 1 if A and B contain exactly the same address points. The intersection is computed by keeping track of the unique identifier of the geomasked address: if an address and the masked version are both in the clusters being compared, they belong to the intersection. We also computed other measures to compare clusters, but only IoU is reported as it is sufficient to distinguish the ability of the presented masking methods to preserve clusters.

\[
\text{IoU}(A, B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}
\] (3)

The preservation of the general spatial distribution can also be measured by different spatial statistics such as the ones listed in the literature review. However, our first experiments showed a strong correlation between cluster preservation and spatial distribution preservation, so spatial distribution preservation measures were limited in the paper to a nearest neighbor analysis. For all the address points in one of our two test areas, we compute for each point its nearest neighbor among the other points. Two indices are used, the mean distance to the nearest neighbor, and the nearest neighbor index (NNI) computed with Equation (4).

\[
\text{NNI} = \frac{\text{Mean}_{distance}}{0.5 \sqrt{\frac{a}{n}}}
\] (4)

where \( \text{Mean}_{distance} \) is the mean of the distances to the nearest neighbor, \( a \) is the total area under study (here the French departments), and \( n \) is the number of points.

4. Experiments with geomasking methods

In this section, experiments are reported with three types of geomasking methods. The first two are adaptations from the literature, and the last one is a new proposition that appeared to be adapted to our use case. The methods experimented with here were chosen because of their successful use on similar data in the past. Other successful methods from the literature could have been similarly tested, but we limited the experiments to these because of time constraints. The others are discussed in Section 5.2.

4.1. Aggregation oriented methods

In aggregation-oriented geomasking methods, cells are defined at the appropriate size to mask details, and all elements contained in a cell are aggregated to this one...
cell. Rather than using regular grids to define the cells for aggregation (Armstrong et al., 1999), we used three different ways to define geographical cells of different sizes: census cells, blocks, and building aggregates. In each case, the address points are aggregated to the centroid of the cell.

The census cells are a partition of the French territory that all approximately contain 2,000 inhabitants. As a consequence, the cells have varying sizes, depending on the population density. The census cells are smaller and more regular in Paris than in the Yvelines region. For this method, the usual definition of k-anonymity is relevant; as all data are aggregated in a cell, the value of k is the size of the cell (number of inhabitants or number of addresses). This method gives very good results in terms of k-anonymity, with a mean of 149 in Paris, and 879 in the Yvelines region, with very few points having a k-anonymity of 5 or less (0.002% in Paris, and 0% in the Yvelines). Regarding Privacy Guarantee, the mean is 24, the median is 14, and 17% of the geomasked address points have a value lower or equal to 5. However, cluster preservation is not good in Paris, with only 18 clusters out of 97 with an IoU value above 0.75. 51 of the 97 clusters even have an IoU value below 0.5. In the Yvelines region, the cluster preservation is better with 56 clusters out of 86 with IoU above 0.75.

Blocks are obtained by computing the faces of the planar graph formed by the road network. These blocks create cells that are smaller than census cells but remain large enough to mask the address points. Figure 2 shows the blocks computed in Paris. The results obtained with block cells are logical with a smaller k-anonymity, but better cluster preservation. In this case, the mean k-anonymity is 30 in Paris (with 3% of address points with k-anonymity equal to or below 5), and 39 in the Yvelines (with 5.4% of address points with k-anonymity equal to or below 5). Regarding Privacy Guarantee, the mean is 22, the median is 16, and 20% of the geomasked address points have a value lower or equal to 5. Regarding cluster preservation, there are only 27 out of 97 with an IoU value above 0.75, and 55 with an IoU value above 0.5.

As cluster preservation was still not satisfying with the blocks, we developed a method to create smaller cells, with building aggregates. Our proposition is to dilate the building polygons to create building aggregates, using a standard morphological dilation, or buffer operation. The principle is similar to the method to derive a built-up area from building polygons (Boffet, 2000), or to the method for the continuous cartographic generalization of urban areas (Peng & Touya, 2017).

Figure 3 shows how the aggregates are created: when dilated buildings are close to each other, their dilated polygons intersect, so we just merge the dilated polygons that intersect each other to create the cells. To make sure the cells are large enough, we used a 10 m morphological dilation. As foreseen, these smaller cells give a lower k-anonymity and better cluster preservation. In Paris for instance, the mean k-anonymity is 29, and the median k-anonymity is 26.4% of the geomasked address points have a k-anonymity lower or equal to 5. Regarding Privacy Guarantee, the mean is 24, the median is 17, and 4% of the geomasked address points have a value lower or equal to 5. Regarding cluster preservation, there are 43 out of 97 with an IoU value above 0.75, and 71 with an IoU value above 0.5.

Finally, we tested the aggregation option of the Adaptive Areal Elimination (AAE) method (Kounadi & Leitner, 2016). It is similar to the aggregation to blocks, but it first measures the number of addresses inside each block. If a block contains fewer addresses than the expected k-anonymity, the block is aggregated to a neighboring block, and the process repeats itself until all blocks contain enough addresses to guarantee an acceptable level of k-anonymity. In our implementation of the method, we used a minimum k-anonymity value of 20, and the method gives higher levels of k-anonymity. For instance, in the Yvelines, where blocks usually contain fewer addresses, the mean k-anonymity is 636, with obviously 0 cases below 5. Even the Privacy Guarantee remains high with a mean of 213 and a median of 62. In response, the cluster preservation is worse than block aggregation, with
only 20% of clusters with an IoU value above 0.75. All the results in the Yvelines region are summarized in Table 1.

4.2. Perturbation oriented method

Perturbation-oriented methods apply a small random perturbation, i.e. displacement here, of all the elements to mask, independently (Armstrong et al., 1999). As the output of this method is a collection of points, the usual definition of k-anonymity is not relevant, and we need to use the definition of J. Wang and Kwan (2020) to evaluate k-anonymity properly. According to the literature, the bimodal Gaussian perturbation (Cassa et al., 2006) is particularly interesting because it merges the benefits of the donut and Gaussian perturbations (Zandbergen, 2014). This is why we used the bimodal Gaussian perturbation as a baseline. As in the donut perturbation, there are two distances $d_1$ and $d_2$ that represent the minimum and maximum displacements allowed, but in the case of the bimodal Gaussian perturbation, $d_1$ and $d_2$ are the centers of two Gaussian distributions of distances $G_1$ and $G_2$. The algorithm works as follows:

1. picks a random angle $\alpha \in [-\pi, \pi]$;
2. randomly picks $G_1$ or $G_2$;
3. randomly picks a distance value $d$ in the chosen Gaussian distribution;
4. displaces the point with orientation $\alpha$ and distance $d$.

However, this method has two main drawbacks: (1) it is hard to find an optimal couple $(d_1, d_2)$ when population and address density vary a lot; (2) the points can be displaced in empty areas, due to the border of the study zone, or to large uninhabited areas (e.g. large rivers, parks, forests, cemeteries, etc.). This is why we propose two enhancements of the bimodal Gaussian method.

To solve problem (1), we introduce a factor $l \in [0, 2]$ that is multiplied to $d$, to reduce the displacement distance, when the density of addresses around the processed point is high, and to increase the displacement distance when the density around the processed point is low. This density is computed by counting the number of address points in a radius around the point. The radius was empirically set to 500 m, as we can see significant differences of density with this ratio, in urban, suburban, or rural areas.

To solve problem (2), we introduce an iterative perturbation process. K-anonymity is computed after a perturbation, and if it is below 5, the perturbation

Table 1. Synthesis of geomasking results with the aggregation oriented methods on the Yvelines area.

| Geomasking method              | Mean k-anonym. (Gp) | Median k-anonym. (Gp) | % of k-anonym. (Gp) < 5 | % IoU > 0,75 |
|-------------------------------|---------------------|-----------------------|-------------------------|-------------|
| Census cell aggregation       | 149 (24)            | 139 (14)              | 0.02 (17)               | 19          |
| Block aggregation             | 39 (22)             | 28 (16)               | 5.4 (20)                | 28          |
| Building group aggregation    | 40 (24)             | 29 (17)               | 4.2 (11)                | 48          |
| AAE                           | 636 (213)           | 224 (62)              | 0 (0)                   | 20          |

Figure 3. Principles of buildings aggregates based on the dilation of building polygons.
is backtracked and the point is displaced with the opposite $\alpha$ angle and the same distance. If k-anonymity is still below 5, the point is pushed 5 m farther in the same direction.

The optimal values for $d_1$, $d_2$, and both Gaussian distributions were empirically defined with sensitivity analysis. $d_1$ was set to 30 m, and $d_2$ was set to 60 m. Figure 4 shows the results of this enhanced perturbation with these optimal values, on a small extract of the Paris test area. It is clearly difficult to find the original address of the geomasked points, as there is no clear pattern. However, the fact that points are not displaced too far enables the preservation of local clusters.

Figure 5 shows the results for Paris, aggregated on the census cells. Most of the cells have a k-anonymity between 5 and 25, with a median of 13 and a mean of 18.5. This figure shows that the cells that contain points with a low k-anonymity are the ones that contain large uninhabited areas such as the Seine river or large parks. But these points are very rare with only 0.17% of points below 5. In the Yvelines, the introduction of the factor $l$ increases the mean k-anonymity from 12 to 24.

Regarding cluster preservation, it is good with 87 clusters out of 97 in Paris preserved with IoU > 0.75, and 95 clusters have an IoU > 0.5. Regarding the nearest neighbor analysis, it confirms that the method tends to spread the points a little, with a mean distance of 36 m and a nearest neighbor index of 0.67, while the initial points have a mean distance of 32 m and a nearest neighbor index of 0.59. The results of the perturbation-oriented methods are summarized for both regions in Table 2.

### 4.3. Simulated cluster crowding method

As a good geomasking technique for our case would guarantee both a high k-anonymity and high cluster preservation, we decided to propose a new technique, loosely based on the simulated crowding of bike GPS tracks (Scheider et al., 2020), which makes sure all

![Figure 4](image-url)  
**Figure 4.** Results obtained with the enhanced bimodal gaussian perturbation in Paris.

![Figure 5](image-url)  
**Figure 5.** 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.
clusters are fully preserved. The principle of the method is to generate new random points inside the extent of existing clusters until we reach the initial number of points in the cluster (Figure 6).

The first step is to generate the clusters to preserve. In our experiments, we used the DBSCAN algorithm as we also use it to evaluate cluster preservation. With DBSCAN, all of the elements are not necessarily grouped in a cluster, some are left alone as outliers, which corresponds to the actual spatio-temporal distribution of COVID-19. Then, the outlier points and the points contained in a cluster are geomasked differently. The outlier points do not need to preserve any cluster, so we mask them with the bimodal Gaussian perturbation method presented in the previous section.

The points contained in a cluster are those that are geomasked by generating the same amount of new points, randomly, inside the polygon extent of the cluster. The polygon extent of the cluster is computed using the k-nearest neighbor concave hull algorithm (Moreira & Yasmina Santos, 2007). Generating random points in a polygon, even with holes, is a pretty straightforward spatial analysis function, accessible in all GIS software or libraries. But if points are generated randomly in the cluster, they can create unexpected, or unrealistic patterns. This is why we introduce two alternative versions of this simulated cluster crowding geomasking method. We call this initial method SCCv1.

For this method, the output is a set of points, but for cases that are in a cluster, these points are chosen uniformly in the area of the cluster, and in particular their

Table 2. Synthesis of geomasking results with the perturbation oriented methods.

| Geomasking method                      | Mean k-anonym. | Median k-anonym. | % of k-anonym. <5 | % IoU > 0.75 | NNI |
|----------------------------------------|----------------|------------------|-------------------|--------------|-----|
| Bimodal gaussian (Paris)               | 18             | 13               | 9                 | 87           | 0.68|
| Enhanced bimodal gaussian (Paris)      | 22.4           | 13               | 0.1               | 77           | 0.66|
| Bimodal gaussian (Yvelines)            | 12             | 8                | 31                | 91           | 0.17|
| Enhanced bimodal gaussian (Yvelines)   | 24             | 17               | 2.3               | 87           | 0.16|

Figure 6. Principles of simulated crowding, new points are generated randomly in the clusters to guarantee cluster preservation.
spatial distribution in the anonymized dataset is independent from the initial distribution of cases in the cluster. Therefore, from the point of view of an opponent willing to infer sensitive information, the anonymized dataset contains the same information as a dataset giving only the boundaries of the clusters and the number of cases for each cluster. As a consequence, for all privacy preservation aspects, this method acts like an aggregation method inside the clusters, and the usual definition of k-anonymity is once again the more relevant, the value of k being the size of the cluster. The cases lying outside clusters are evaluated with the k-anonymity definition used for perturbation-oriented methods.

To avoid the new generated points appearing too close to initial points, we generate a buffer area around each initial point, and we pierce the cluster polygon with these buffer areas: the polygon delineating the cluster now contains holes around each address to mask. Then, we generate the random points in this pierced polygon to guarantee a minimum distance to the initial points. We conducted experiments to find an optimal size for the buffers or holes: the larger the holes are, the further away the initial points marked points are, but the smaller the area of the pierced polygon is. Figure 7 shows an example cluster with buffers of 20 m and 30 m. In both cases, there is enough room in the pierced polygon to get a spatial distribution of the randomly generated points, which does not reveal the location of the holes. However, if we increase the buffer size to 40 m (Figure 7(c)), the pierced area is often small and fragmented. The resulting spatial distribution will exhibit fake smaller clusters, thus misleading the possible analyses of the geomasked data. This is why we consider the 30 m buffer area as a maximum in our COVID-19 use case. We call this alternative method SCCv2.

Finally, we observe that both first simulated cluster crowding methods tend to generate points away from roads, inside the blocks, which is not realistic, as most address points are located close to the roads (Figure 8). This is why we propose an additional geomasking method, called SCCv3. We dilate the roads with a 15 m buffer and then join all these polygons into one big geometry. The 15 m threshold was determined empirically as 97% of the address points in our test datasets are located within this 15 m radius around roads. Finally, each cluster is intersected with this big geometry, and the fake points are generated inside this intersected geometry instead of inside the whole cluster polygon.

As foreseen, the cluster preservation is good with the simulated cluster crowding: 100% of the clusters in Paris have an IoU above 0.75, which is by far our best result. Regarding the k-anonymity, all three versions of simulated cluster crowding give a median value of 640, with a mean value of 1208. 5.8% of the points have a k-anonymity value below or equal to 5. The Privacy Guarantee is less positive with a mean value of 9, a median value of 7, and 11% of the points have a value below or equal to 5. This difference between k-anonymity and Privacy Guarantee can be explained by the large number of points to anonymize contained in each cluster. Table 3 summarizes the results obtained with SCC.

![Figure 7](image.png)  
*Figure 7. The extent of a cluster and the area covered by the buffer around the initial points: a) 20 m buffer, b) 30 m buffer, c) 40 m buffer where the remaining area to generate masked points is too small and fragmented.*
Figure 9 shows the results of this method on a small extract in Paris. The figure shows that the generated masked points often lie inside the blocks with versions 1 and 2. The figure also shows that SCCv2 does create sub-clusters, which is not the case for SCCv1 and SCCv3. This is confirmed by the nearest neighbor evaluation of the geomasked points. SCCv2 has a mean distance of 25 m, while the initial points and the points masked by SCCv1 and SCCv3 have mean distance values between 32 and 33 m. The nearest neighbor index also shows that SCCv2 generates points that are more clustered than the initial ones. And the NNI for SCCv2 also shows a more clustered distribution than the points generated by SCCv1 and SCCv3 (0.46 vs 0.59 for all other three point clouds).

5. Discussion

5.1. Comparison of the geomasking methods

From our experiments to mask COVID-19 data in Paris (see Table 4) and the Yvelines region, there is no clear evidence of one method being better than the others (Figure 10). Some methods may appear better than the others depending on what we want to do with the geomasked data. These different perspectives are discussed in this section.

Table 3. Synthesis of geomasking results with the simulated cluster crowding. $R_p$ is only computed for the points inside a cluster.

| Geomasking method | Mean k-anonym. ($G_p$) | Median k-anonym. ($G_p$) | % of k-anonym. ($G_p$) < 5 | NNI |
|-------------------|------------------------|--------------------------|----------------------------|-----|
| SCC v1            | 1208 (9)               | 610 (7)                  | 5.8 (11)                   | 0.56|
| SCC v2            | 1208 (9)               | 610 (7)                  | 5.8 (11)                   | 0.60|
| SCC v3            | 1208 (9)               | 610 (7)                  | 5.8 (11)                   | 0.58|

Figure 8. This extract of Paris addresses show that most address points are located within a 15 m radius around roads.

Figure 9. Results of the three versions of the simulated crowding method (with a 30 m buffer for v2).
If we want to maximize k-anonymity, regardless of cluster preservation, the SCC is the safest method, with the AAE as second best. Unfortunately, this method is also quite inconsistent regarding the k-anonymity of the cases outside the clusters, with 5.8% of all cases below 5. Regarding the second best method, AAE, it also minimizes cluster preservation (with the census cell aggregation method), so it should be kept only for uses where very local clusters are not very important and the census cell scale is sufficient. If the Privacy Guarantee is considered instead of k-anonymity, AAE is the best method, as the aggregation cells of SCC contain many points to anonymize by nature.

If we want to maximize k-anonymity consistency, regardless of cluster preservation, AAE is the best method as all points are masked with a k-anonymity and a Privacy Guarantee value above 5. If we take cluster preservation into account, the enhanced bimodal Gaussian perturbation is the best method, as only 0.1% of the geomasked points remain with a low IoU.
k-anonymity. With this method, the k-anonymity is consistent with a large majority of points in a range between 10 and 20 of k-anonymity. Figure 11 shows an example of this consistency at the border of Paris.

If we want to maximize cluster preservation, the best method is the simulated cluster crowding, as it was designed to maximize cluster preservation. Although the median k-anonymity is quite high with this method, its defect is its consistency, as too many points still have a low k-anonymity.

If we want to analyze rural areas, the building group aggregation seems to be the best method because it maintains a very high k-anonymity in areas of low density while preserving clusters correctly. This is due to the size of the clusters in areas of low density with sprawled dwellings. The SCC also gives good results in rural areas because the large clusters allow much free space to generate masked points. The AAE is also a good solution for rural areas: it guarantees a good anonymity as the aggregation units are merged to their neighbors when they do not contain enough addresses for masking, which is common in rural areas.

If we want the best overall technique, the enhanced bimodal Gaussian perturbation seems to be the most balanced one between k-anonymity and cluster preservation. It provides a good k-anonymity, with very few points badly masked in particular, and preserves local clusters well. With its density analysis step, it adapts well to rural and urban areas. However, we believe that the simulated cluster crowding has the potential for this best technique spot, if we manage to improve its main defect, the lack of consistency, with 11% of points insufficiently masked. Figure 12 shows the results of both techniques on the same extract in Paris. The ability of simulated cluster crowding to concentrate masked points inside the blocks Figure 11. 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 12. Results of the enhanced bimodal gaussian perturbation, and the cluster oriented simulated crowding on the same extract of the Paris area.
explains why the median k-anonymity is so high compared to the perturbation technique. For points outside clusters, the same perturbation is used so the results are similar.

5.2. Alternative geomasking methods

Beyond the proposed geomasking techniques, other techniques from the literature could be used to mask addresses while preserving COVID-19 cases clusters. Some of these techniques are discussed in this section. First, it is possible to aggregate the addresses to a regular grid rather than geographic cells (Armstrong et al., 1999; Seidl et al., 2016). The main advantage of this method is strong anonymization if the grid cells are large enough. The main drawback is cluster preservation in dense areas as clusters might be smaller than the cells of the grid. It was shown that grid preservation is not effective to preserve spatial patterns in general (Broen et al., 2021).

Transforming the point information into heat maps is another way to guarantee a good k-anonymity (Z. Wang et al., 2019), and heat maps are perceived as not disclosing privacy by map readers (Kim et al., 2021). Besides k-anonymity, the theoretical advantage of heat maps is the preservation of large clusters that are converted into hot spots in the heat map. The main drawback of this method is the limitation of possible further analysis, compared to a set of spatio-temporal points. This is why our end-users, i.e. the epidemiologists searching for local clusters, discarded this approach.

Voronoi masking proved very effective in recent publications on geomasking (Broen et al., 2021; Seidl et al., 2015). The principles of Voronoi masking is to generate a Voronoi diagram of the points to mask, and then to project each point on the nearest edge of the diagram. The advantages are both a good overall k-anonymity while preserving clusters by construction. The method is also adapted to varying densities because the Voronoi cells are larger in areas of low density. The drawback is the limited perturbation performed in areas with a very high density, which occurs quite often in cities with COVID-19. However, this method has the potential to be at least as good as the ones proposed in this paper.

Finally, Adaptive Areal Masking (AAM) (Charleux & Schofield, 2020) is an improved version of AAE. The main advantage of this method is its consistency in areas with heterogeneous density. Similarly to Voronoi masking, the drawback seems to be the processing of areas with very dense clusters, where the anonymization polygons might be too small. Once again similarly to Voronoi masking, this method has the potential to be at least as good as the ones proposed in this paper.

6. Conclusions and future work

To conclude, we can answer the three research questions raised in the introduction of the paper. First, the aggregation and perturbation methods from the literature were adapted to provide a significant geomasking of the addresses of COVID-19 cases, while preserving more or less the clusters in the spatial distribution. Overall, the enhanced bimodal Gaussian perturbation resulted in the best compromise between anonymity and cluster preservation in the tested methods. Then, we proposed a specific simulated cluster crowding method, which fully preserves clusters while providing satisfying anonymization. Finally, while the aggregation methods appeared to be very sensitive to population density, it is possible to make the other two methods adaptive to population density, and they provided similar results in the very dense test area of Paris, and in the more heterogeneous area of Yvelines. Our goal is now to help the health authorities in France (and maybe elsewhere), to adopt these methods, in order to safely diffuse these important datasets.

To go further, the first step would be to test Voronoi masking and Adaptive Areal Masking on the same datasets, to compare them to the proposed method. Meanwhile, it is still possible to improve the geomasking techniques proposed in this paper. Our priority is to improve the simulated cluster crowding method. One idea would be to propose a better masking method for the points outside the cluster, using the same idea of generating fake points randomly in a polygon. Another idea would be to generate fake points based on the density of initial points rather than randomly in the concave hull of the cluster. We also need to deal with uninhabited areas, which cause low k-anonymity values for all techniques. The use of a mask of uninhabited areas (Lu et al., 2012) to exclude these areas from the potential places to move or generate a point should improve all our proposed techniques. Another idea is to investigate the quality of geomasking regarding false identification, i.e. how much masked addresses are close to other real addresses, as false identifications have been identified as a major risk with geospatial health data (Seidl et al., 2018). In the same vein, it would be interesting to measure how much these methods may create false clusters. This problem is one of the reasons why the buffer of the simulated cluster crowding was not pushed beyond 30 m, but it would be important to measure how much all the methods tend to create false clusters by displacing or masking points. Finally, we would like to test if these techniques, particularly the new cluster-oriented simulated crowding, can be useful for the geomasking of other types of geospatial health data.
Notes

1. https://www.data.gouv.fr/fr/datasets/base-adresse-nationale/
2. https://www.data.gouv.fr/fr/datasets/donnees-relatives-aux-resultats-des-tests-virologiques-covid-19

Acknowledgments

The authors thank Renaud and Martine Piarroux from Hôpitaux de Paris, for their valuable time.

Disclosure statement

No potential conflict of interest was reported by the author(s)

ORCID

Guillaume Touya http://orcid.org/0000-0001-6113-6903

References

Ajayakumar, J., Curtis, A. J., & Curtis, J. (2019, December). Addressing the data guardian and geospatial scientist collaborator dilemma: How to share health records for spatial analysis while maintaining patient confidentiality. International Journal of Health Geographics, 18(1), 30. https://doi.org/10.1186/s12942-019-0194-8

Allshouse, W. B., Fitch, M. K., Hampton, K. H., Gesink, D. C., Doherty, I. A., Leone, P. A., Serre, M. L., & Miller, W. C. (2010). Geomasking sensitive health data and privacy protection: An evaluation using an E911 database. Geocarto International, 25(6), 443–452. https://doi.org/10.1080/10106049.2010.496496

Andrés, M. E., Bordenebre, N. E., Chatzikokolakis, K., & Palamidessi, C. (2013). Geoindistinguishability: Differential privacy for location-based systems. Proceedings of the 2013 acm sigac conference on computer communications security (p. 901–914). New York, NY, USA: Association for Computing Machinery.

Armstrong, M. P., Rushton, G., & Zimmerman, D. L. (1999). Geographically masking health data to preserve confidentiality. Statistics in Medicine, 18(5), 497–525.

Boffet, A. (2000). Creating urban information for cartographic generalisation. International Symposium on Spatial Data Handling (SDH). Beijing, China: International Geographical Union.

Broen, K., Trangucci, R., & Zelner, J. (2021). Measuring the impact of spatial perturbations on the relationship between data privacy and validity of descriptive statistics. International Journal of Health Geographics, 20(1), 3. https://doi.org/10.1186/s12942-020-00256-8

Cassa, C. A., Grannis, S. J., Overhage, J. M., & Mandl, K. D. (2006). A context- sensitive approach to anonymizing spatial surveillance data: Impact on outbreak detection. Journal of the American Medical Informatics Association: JAMIA, 13(2), 160–165. https://doi.org/10.1197/jamia.M1920

Cassa, C. A., Wieland, S. C., & Mandl, K. D. (2008). Re-identification of home addresses from spatial locations anonymized by Gaussian skew. International Journal of Health Geographics, 7(1), 45. https://doi.org/10.1186/1476-072X-7-45

Chang, S., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D., & Leskovec, J. (2020, November). Mobility network models of COVID-19 explain inequities and inform reopening. Nature, 82–87. https://www.nature.com/articles/s41586-020-2923-3

Charleux, L., & Schofield, K. (2020). True spatial k-anonymity: Adaptive areal elimination vs. adaptive areal masking. Cartography and Geographic Information Science, 47(6), 537–549. https://doi.org/10.1080/15230406.2020.1794975

Chen, S., Li, Q., Gao, S., Kang, Y., & Shi, X. (2020). State-specific projection of COVID-19 infection in the United States and evaluation of three major control measures. Scientific Reports, 10(1), 22429. https://doi.org/10.1038/s41598-020-80044-3

Clarke, K. C. (2016). A multiscale masking method for point geographic data. International Journal of Geographical Information Science, 30(2), 300–315. https://doi.org/10.1080/13658816.2015.1085540

Curtis, A. J., Mills, J. W., & Leitner, M. (2006). Spatial confidentiality and GIS: Re-engineering mortality locations from published maps about Hurricane Katrina. International Journal of Health Geographics, 5(1), 44. https://doi.org/10.1186/1476-072X-5-44

Danis, K., Épaulard, O., Bénét, T., Gaymard, A., Campoy, S., Botelho-Nevers, E., Team, I., Ader, F., Mailles, A., Boudalaaz, Z., Tolsma, V., Berra, J., Vaux, S., Forestier, E., Landelle, C., Fougere, E., Thabuis, A., Berthelot, P., Veil, R., Mounayar, A.-L., & Bouscambert-Duchamp, M. (2020). Cluster of Coronavirus disease 2019 (COVID-19) in the French Alps, February 2020. Clinical Infectious Diseases, 71(15), 825–832. https://doi.org/10.1093/cid/ciaa424

Delmelle, E., Dony, C., Casas, I., Jia, M., & Tang, W. (2014). Visualizing the impact of space-time uncertainties on dengue fever patterns. International Journal of Geographical Information Science, 28(5), 1107–1127. https://doi.org/10.1080/13658816.2013.871285

Desjardins, M., Hohl, A., & Delmelle, E. (2020). Rapid surveillance of covid-19 in the united states using a prospective space-time scan statistic: Detecting and evaluating emerging clusters. Applied Geography, 118, 102202. https://doi.org/10.1016/j.apgeog.2020.102202

Duckham, M., & Kulik, L. (2006). Location privacy and location-aware computing. In E. Joao, R. Billen & D. Forrest (Eds.), Dynamic and mobile GIS (pp. 63–80). CRC Press.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the second international conference on knowledge discovery and data mining (p. 226–231). AAAI Press.

Gao, S., Rao, J., Kang, Y., Huang, Q., & App, J. (2019). Exploring the effectiveness of geomasking techniques for protecting the geoprivacy of Twitter users. Journal of Spatial Information Science, 19, 510. http://josis.org/index.php/josis/article/view/510
Peng, D., & Touya, G. (2017). Continuously generalizing buildings to built-up areas by aggregating and growing. Proceedings of 3rd ACM SIGSPATIAL workshop on smart cities and urban analytics (UrbanGIS’17). Redondo Beach, CA. https://doi.org/10.1145/3152178.3152188

Piarroux, R. (2019). Choléra. haiti 2010-2018: Histoire d’un désastre. Cnrs.

Scheider, S., Wang, J., Mol, M., Schmitz, O., & Karssenberg, D. (2020). Obfuscating spatial point tracks with simulated crowding. International Journal of Geographical Information Science, 34(7), 1398–1427. https://doi.org/10.1080/13658816.2020.1712402

Seidl, D. E., Jankowski, P., & Clarke, K. C. (2018). Privacy and false identification risk in geomasking techniques. Geographical Analysis, 50(3), 280–297. eprint. https://onlinelibrary.wiley.com/doi/pdf/10.1111/gean.12144

Seidl, D. E., Jankowski, P., & Tsou, M.-H. (2016). Privacy and spatial pattern preservation in masked GPS trajectory data. International Journal of Geographical Information Science, 30(4), 785–800. https://doi.org/10.1080/13658816.2015.1101767

Seidl, D. E., Paulus, G., Jankowski, P., & Regenfelder, M. (2015). Spatial obfuscation methods for privacy protection of household-level data. Applied Geography, 63, 253–263. https://doi.org/10.1016/j.apgeog.2015.07.001

Sherman, J. E., & Fetters, T. L. (2007). Confidentiality concerns with mapping survey data in reproductive health research. Studies in Family Planning, 38(4), 309–321. https://doi.org/10.1111/j.1728-4465.2007.00143.x

Swanlund, D., Schurman, N., Zandbergen, P., & Brussoni, M. (2020). Street masking: A network-based geographic mask for easily protecting geoprivacy. International Journal of Health Geographics, 19(1), 26. https://doi.org/10.1186/s12942-020-00219-z

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557–570. https://doi.org/10.1142/S0218478802001648

Wang, J., & Kwan, M.-P. (2020). Daily activity locations k-anonymity for the evaluation of disclosure risk of individual GPS datasets. International Journal of Health Geographics, 19(1), 7. https://doi.org/10.1186/s12942-020-00201-9

Wang, Z., Liu, L., Zhou, H., & Lan, M. (2019). How is the confidentiality of crime locations affected by parameters in kernel density estimation? ISPRS International Journal of Geo-Information, 8(12), 544. https://doi.org/10.3390/ijgi8120544

Yong, S. E. F., Anderson, D. E., Wei, W. E., Pang, J., Chia, W. N., Tan, C. W., Teoh, Y. L., Rajendram, P., Toh, M. P., Poh, C., Koh, V. T. (2020). Connecting clusters of COVID-19: An epidemiological and serological investigation. The Lancet Infectious Diseases, 20(7), 809–815. https://doi.org/10.1016/S1473-3099(20)30273-5

Zandbergen, P. A. (2014). Ensuring confidentiality of geoded health data: Assessing geographic masking strategies for individual-level data. Advances in Medicine, 2014, 567049. https://doi.org/10.1155/2014/567049

Zhang, S., Freundschuh, S. M., Lenzer, K., & Zandbergen, P. A. (2017). The location swapping method for geomasking. Cartography and Geographic Information Science, 44(1), 22–34. https://doi.org/10.1080/15230406.2015.1095655