Geographically Masking Addresses to Study COVID-19 Clusters

Walid Houfaf-Khoufaf
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, geomasking, spatial anonymisation, address, k-anonymity

Posted Date: February 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-128679/v2

ARTICLE TEMPLATE

Geographically Masking Addresses to Study COVID-19 Clusters

Walid Houfaf-Khoufaf[a] and Guillaume Touya[b]

[a]Univ Gustave Eiffel, ENSG, IGN, 6-8 Avenue Blaise Pascal, F-77420, Champs-sur-Marne, France; [b]LASTIG, Univ Gustave Eiffel, ENSG, IGN, 73 avenue de Paris, F-94160 Saint-Mande, France

ARTICLE HISTORY
Compiled February 3, 2021

ABSTRACT
The spatial analysis of health data usually raises geoprivacy issues. But with the virulence of COVID-19, scientists and crisis managers do need to analyse the spatio-temporal distribution and spreading of the disease with spatially precise data. In particular, it is useful to locate each case on a map to identify clusters of cases in space and time. To allow such analyses with breach of geoprivacy, geomasking techniques are necessary. This paper experiments the geomasking techniques from the literature to solve this problem: masking the real address of positive cases while preserving the local 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 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.

KEYWORDS
COVID-19; geomasking; spatial anonymisation; address; k-anonymity;

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, Delmelle, & Eberth, 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, Hohl, & Delmelle, 2020). Geo-visual analysis can also be used to understand the epidemic (Delmelle, Dony, Casas, Jia, & Tang, 2014), or to simulate the spatial infection (Chen, Li, Gao, Kang, & Shi, 2020).

As we are participating in a local initiative to track and break the clusters at a very local scale (a street or a neighbourhood), following strategies that proved efficient against cholera (Piarroux, 2019). We want to provide both spatial analysis and geovisualisation tools to help the epidemiologists that lead this initiative. But tracking and visualising the addresses of people tested for COVID-19 raises severe privacy issues.

CONTACT A. N. Author. Email: guillaume.touya@ign.fr

Privacy problems are inherent to health spatial data (Sherman & Fetters, 2007), and even more prominent when sciences tries to be reproducible (Ajayakumar, Curtis, & Curtis, 2019). And a recent study shows that visualisation with a high level of detail (e.g. with point symbols for addresses) are perceived as more risky for geoprivacy than heat maps for instance (Kim, Kwan, Levenstein, & Richardson, 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. But, 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 can be 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. In order 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 on Paris and a less dense surrounding region.

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 for location privacy is now well acknowledged in many countries. In Europe, the General Data Protection Regulation (GDPR) protects personal data (Georgiadou, de By, & Kounadi, 2019), and location information can cause location-based spam, attacks to 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. And the exact address can be re-engineered from rougly masked spatial data, either by light field surveys (Curtis, Mills, & Leitner, 2006), or by automated methods (Cassa, Wieland, & Mandl, 2008). False identifications is also a risk of attack to location data (Seidl, Jankowski, & Clarke, 2018).

There are different ways to prevent these types of attacks and preserve the privacy of the people included in a geospatial health dataset (Katsomallos, Tzompanaki, & Kotzinos, 2019), and the one that is the most interesting in our COVID-19 use case is geomasking or spatial anonymisation. In the past twenty years, scholars proposed different types of geomasking techniques. There are global transformations of the dataset such as an affine transformation (Armstrong, Rushton, & Zimmerman, 1999), or location swapping (Zhang, Freundschuh, Lenzer, & Zandbergen, 2017), which is not really 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, or a perturbation donut where some specific areas are excluded (Lu,

Yorke, & Zhan, 2012). There are also geographical aggregation techniques, based on a Voronoï diagram (Seidl, Paulus, Jankowski, & Regenfelder, 2015), a predefined grid (Seidl, Jankowski, & Tsou, 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, Liu, Zhou, & Lan, 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 to this use is still possible with a good accuracy (Scheider, Wang, Mol, Schmitz, & Karssenberg, 2020).

These geomasking techniques are really sensible 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 of a good geomasking, rather than on new techniques. Geomasking is always a tradeoff between protection and accuracy of the masked data (Gao, Rao, Kang, Huang, & App, 2019; Kwan, Casas, & Schmitz, 2004). An interesting recent study measured how much spatial distributions and patterns can be modified by different geomasking techniques (Broen, Trangucci, & Zelner, 2021). Geo-indistinguishability is an interesting notion to measure this tradeoff (Andrés, Bordenabe, Chatzikokolakis, & Palamidessi, 2013).

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 sixth Section of the paper.

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 is already partly anonymized as it only contains addresses for each person being tested positive to one of the tests for COVID-19, and the date of the test. Age, gender, or Body Mass Index are not stored in the dataset, and as a consequence, do not require any anonymization. But this data is only released aggregated at the region level (the French "départements"), even for epidemiologists. And the aggregated data is clearly not the good scale to identify clusters at the scale of a street, or of a neighbourhood. When designing a geomasking technique for sensible geospatial health data, the geographical information scientists 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 administrations[1] to disaggregate the official data[2]. In this dataset, we have everyday, 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, and rather favours 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: for a given date n, and a given incidence rate r_0, for each each case of date $n-1$, we pick a number of incidental cases with a gaussian probability centred on r_0. The incidental cases are randomly selected in the address points in a radius of 300 m. The datasets generated with this method were considered realistic by epidemiologists having access to the real SI-DEP data, with cluster appearing over space and time. 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 (Figure 1), with a timespan of three months. These three months correspond to the complete timespan available at the time of the experiment, and is enough to see spatio-temporal 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 neighbourhood. 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. From this point of the paper, we only focus on the spatial dimension of these data, as we decided not to anonymize the date, as it is already fuzzy because people were sometimes tested at the beginning of the disease, and sometimes at the end.

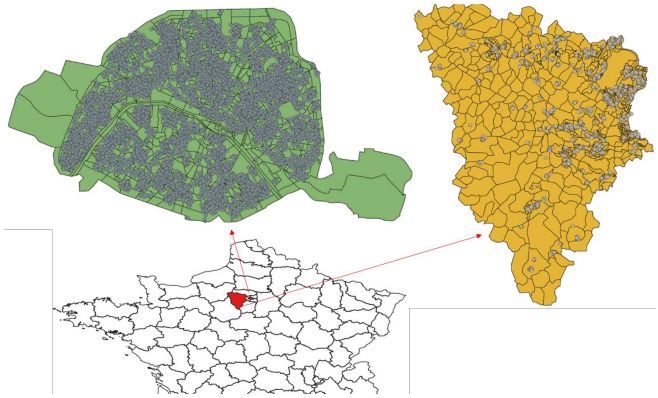

Figure 1. The generated synthetic COVID-19 data, for Paris on the left, and Yvelines on the right, both for three months. The gray point symbols are the cases, displayed on top of census administrative cells. The extract of France below shows the location of both regions and their respective size.

3.2. Evaluation of the Anonymisation

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 its elements, it cannot be distinguished from at

[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

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 a definition of spatial k-anonymity similar to the one proposed in Wang and Kwan (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 2).

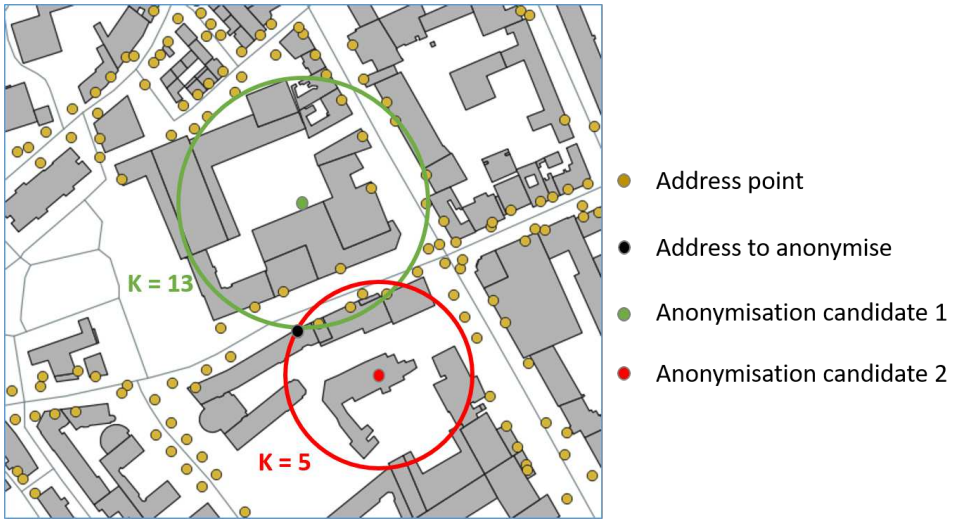

Figure 2. Principles of k-anonymity computation: anonymisation candidate 1 (in green) has 13 address points in the radius from the candidate to the address to anonymise (in black), so has k-anonymity value of 13. The other candidate position (in red) has a k-anonymity value of 5.

As the main use case of this data is the observation and analysis of spatio-temporal very local to large scale 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, Kriegel, Sander, & Xu, 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 1 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.

$$IoU(A, B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{1}$$

The preservation of the general spatial distribution can also be measured by different spatial statistics such as Moran's I, as proposed previously (Broen et al., 2021). However, our first experiments showed a strong correlation between cluster preservation and spatial distribution preservation, so spatial distribution preservation measures are omitted in the paper for the sake of simplicity.

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 adapted to our use case. The methods experimented here were chosen because of their successful use on similar data in the past. Other methods form the literature could have been similarly tested, but we limited the experiments to these ones for 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 (Armstrong et al., 1999). 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 a varying size, depending on the population density. The census cells in our two test areas are visible in Figure 1 below the cases symbols, and we can see that these cells are smaller and more regular in Paris than in the Yvelines region. This method gives very good results in terms of k-anonymity, with a mean of 95 in Paris, and 413 in the Yvelines region, with very few points having a k-anonymity of 5 or less (2% in Paris, and 0.8% in the Yvelines). However, cluster preservation is really bad 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 3 shows the blocks computed in Paris. The results obtained with block cells are logical with a smaller k-anonymity, but a better cluster preservation. In this case, the mean k-anonymity is 42 in Paris (with 6.6% of address points with k-anonymity equal or below 5), and 31 in the Yvelines (with 16% of address points with k-anonymity equal or below 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 generalisation of urban areas (Peng & Touya, 2017). Figure 4 shows how the aggregates are created: when dilated buildings are close to each other, their dilated polygon 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 a better cluster preservation. In Paris for instance, the mean k-anonymity is 17.3, and the median k-anonymity is 14. 13% of the geomasked address points have a k-anonymity lower or

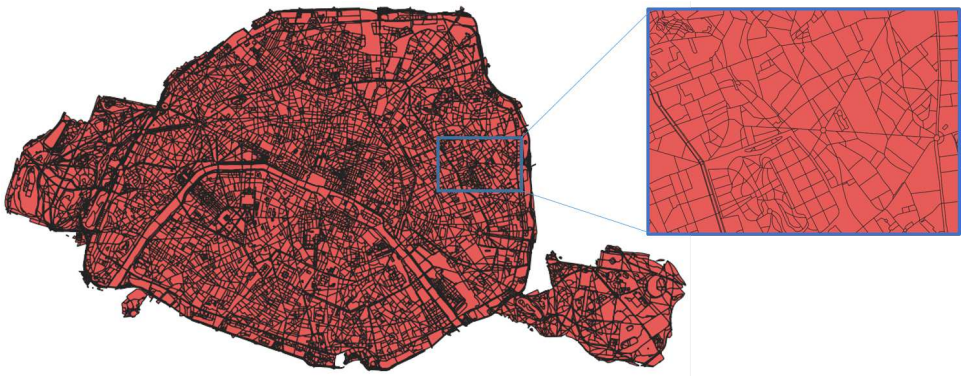

Figure 3. The blocks computed using the road network graph. There are much more cells than with the census.

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.

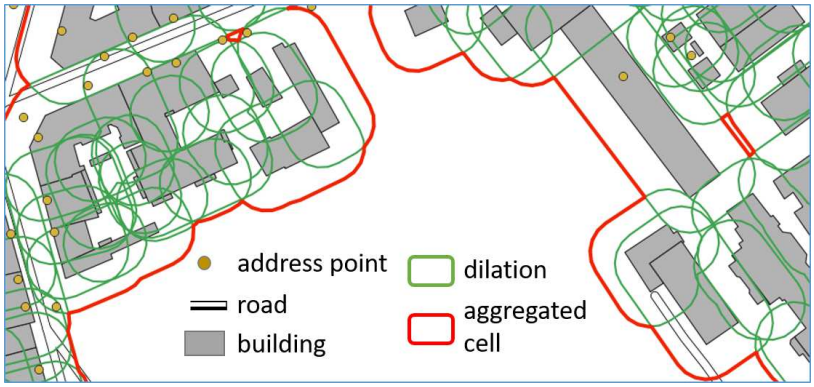

Figure 4. Principles of buildings aggregates based on the dilation of building polygons.

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). According to the literature previously presented, the best perturbation method seems to be the bimodal gaussian that 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) pick a random angle $\alpha \in [-\pi, \pi]$;
(2) randomly pick G_1 or G_2;
(3) randomly pick a distance value d in the chosen gaussian distribution;
(4) displace the point with orientation α 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 is backtracked and the point is displaced with the opposite α angle and the same distance. If k-anonymity is still below 5, the point is pushed 5 m farther in the same direction.

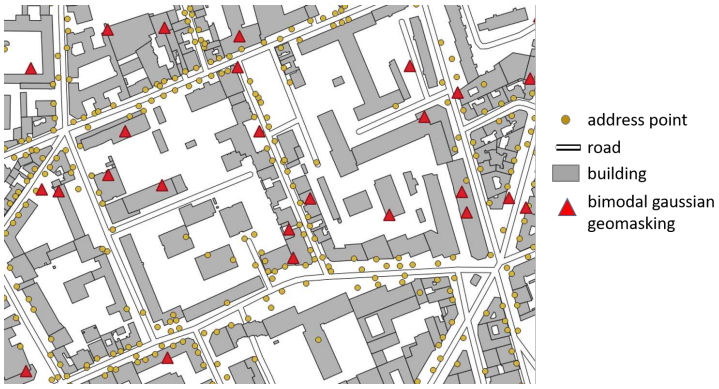

Figure 5. Results obtained with the enhanced bimodal gaussian perturbation in Paris.

The optimal values for d_1, d_2, and both gaussian distributions were empirically defined with as sensitivity analysis. d_1 was set to 30 m, and d_2 was set to 60 m. Figure 5 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 6 shows the results on 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 really with 87 clusters out of 97 in Paris preserved with IoU > 0.75, and 95 clusters have an IoU > 0.5.

4.3. Cluster Oriented Method

As a good geomasking technique for our case would guarantee both a high k-anonymity and a 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

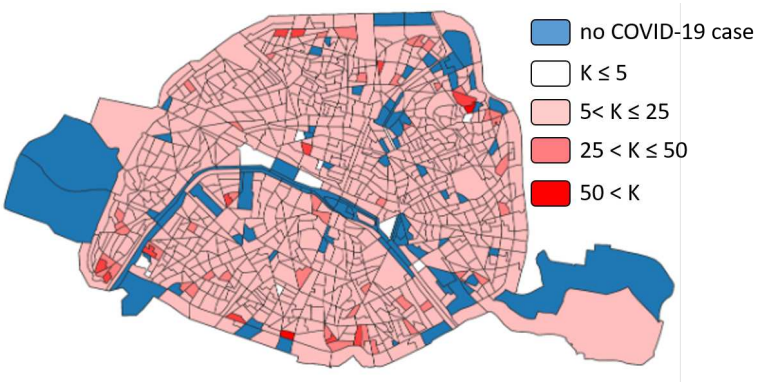

Figure 6. Mean k-anonymity in the census cells of Paris after an enhanced bimodal gaussian perturbation. The blue cells did not contain any (fake) COVID-19 case to geomask.

sure all 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 7).

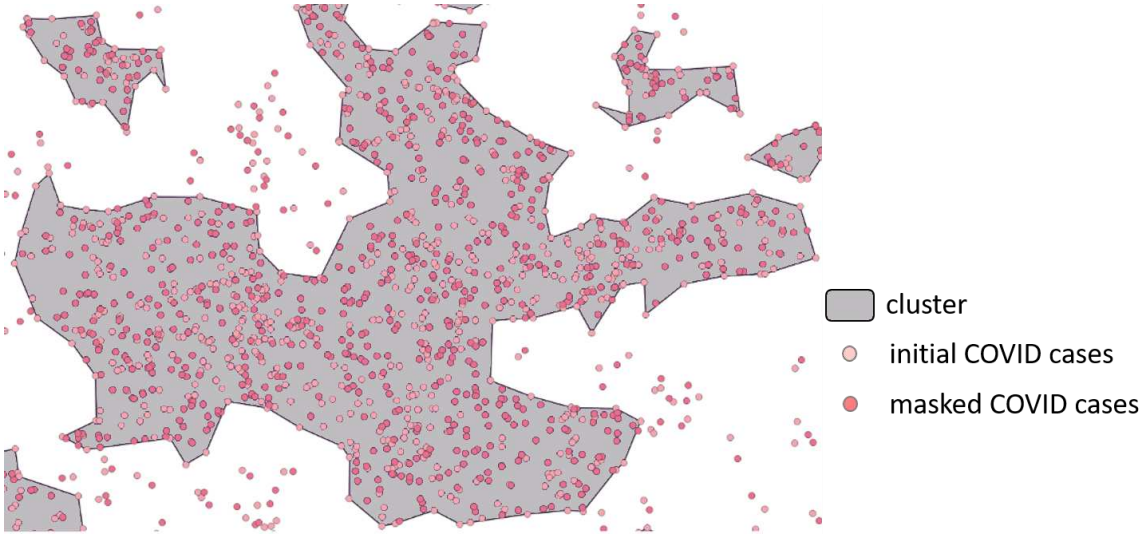

Figure 7. Principles of simulated crowding, new points are generated randomly in the clusters to guarantee cluster preservation.

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 outliers 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 the ones that are geomasked by generating the same amount of new points, randomly, inside the polygon extent of the cluster. Generating random points in a polygon, even with holes, is a pretty straightforward spatial analysis function, accessible in all GIS softwares or libraries. But if points are generated randomly in the cluster, they can lie very close to the initial points,

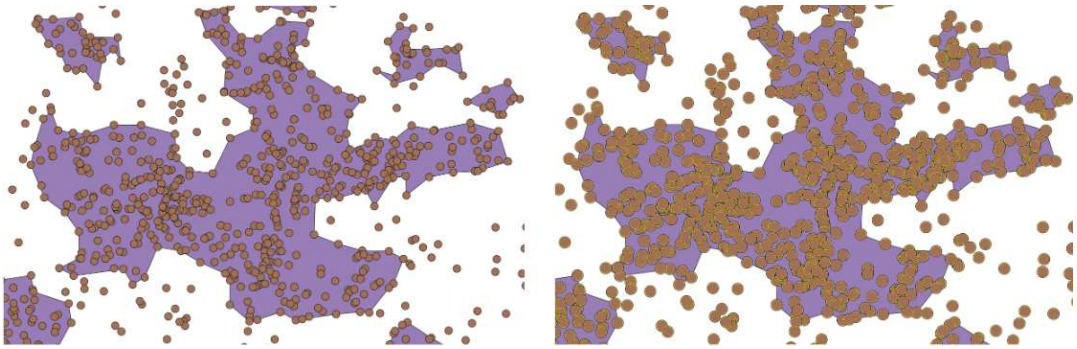

Figure 8. The extent of a cluster and the area covered by the buffer around the initial points (20m buffer on the left, 30m buffer on the right).

leading to a poor k-anonymity. To avoid that the new generated points appear too close to initial, 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 poins 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 better k-anonymity is, but the smaller the area of the pierced polygon is. Figure 8 shows an example cluster with buffers of 20m and 30m. In both cases, there is enough room in the pierced polygon to get a realistic spatial distribution of the randomly generated points. However, if we increase the buffer size to 40m (Figure 9 to improve the resulting k-anonymity, 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 30m buffer area as a maximum in our COVID-19 use case.

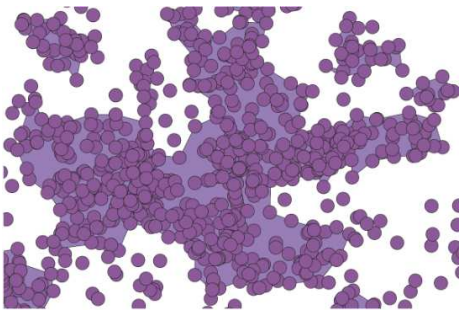

Figure 9. When the buffer size is increased to 40m, the remaining area to generate masked points is too small and fragmented.

In order to compute k-anonymity as proposed in this paper, we need a matching between a masked point and a real address point. But, for the points contained in clusters, the simulated crowding method generates new points independently from the initial ones, so there is no obvious matching to compute k-anonymity. A first basic method to match points would be to randomly match points, but we dismissed this method because it gave very high k-anonymity values, which did not represent the real masking quality. The algorithm below shows how we propose to compute a more pessimistic matching for k-anonymity, matching masked points to the closest unmatched initial address point. Points outside clusters are masked using the bimodal

gaussian perturbation, so the matching is direct.

Algorithm 1 Pseudo-code of the algorithm to match address points and masked points in a cluster.

```
1:  procedure MATCH(list_initial, list_masked)
2:      initialize map_matches
3:      for p in list_masked do
4:          dist_min = ∞
5:          nearest = null
6:          for q in list_initial do
7:              dist = distance(p, q)
8:              if dist < dist_min then
9:                  dist_min = dist
10:                 nearest = q
11:         list_initial.remove(nearest)
12:         map_matches.put(p, nearest)
13:     return map_matches
```

As foreseen, the cluster preservation is really good with the cluster oriented simulated crowding: 100% of the clusters in Paris have an IoU above 0.75, which is by far our best results. Regarding the k-anonymity, the simulated crowding with a 20m buffer has a median value of 22, with 11% of the points having a k-anonymity value below or equal to 5. The simulated crowding with a 30m buffer has a median value of 29, with 7% of the points having a k-anonymity value below or equal to 5. To demonstrate the importance of piercing the cluster polygons with the small buffer areas around each initial point, we computed the k-anonymity of a purely random simulated crowding: the median k-anonymity is 17 and almost 20% of the points have a k-anonymity value below or equal to 5. Figure 10 shows the results of this method on a small extract in Paris. It appears that the generated masked points often lie inside the blocks, while address points are generally located along roads.

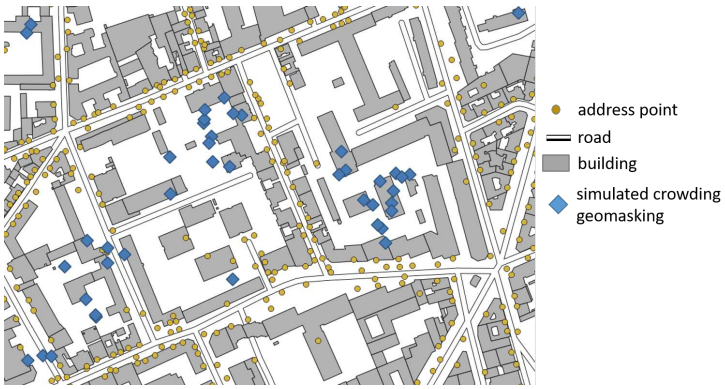

Figure 10. Results of the simulated crowding method with a 30m buffer. The empty spaces where points are grouped often correspond to the interior of blocks, while address points are generally located along roads.

5. Discussion

5.1. Comparison of the Geomasking Methods

From our experiments to mask COVID-19 data in Paris (see Table 1) and the Yvelines region, there is no clear evidence of one method being better than the others. But depending on what we want to do with the geomasked data, some methods may appear better than the others. These different perspectives are discussed in this section.

Table 1. Synthesis of geomasking results in Paris.

Geomasking method	median k-anonym.	% of k-anonym. ≤ 5	% $IoU > 0.75$
Census cell aggregation	**74**	3	19
Block aggregation	27	7	28
Building group aggregation	14	13	48
Bimodal gaussian	13	9	87
Enhanced bimodal gaussian	13	**0.1**	77
Simulated crowding (20m)	22	11	**100**
Simulated crowding (30m)	29	7	**100**

If we want to maximize k-anonymity, regardless of cluster preservation, the aggregation on census cells is the safest method. Unfortunately, this method also minimzes cluster preservation, so it should be kept only for uses where very local clusters are not very important and the census cell scale is sufficient.

If we want to maximize k-anonymity consistency, regardless of cluster preservation, the enhanced bimodal gaussian perturbation is the best method, as only 0.1% of the geomasked points remain with a low 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.

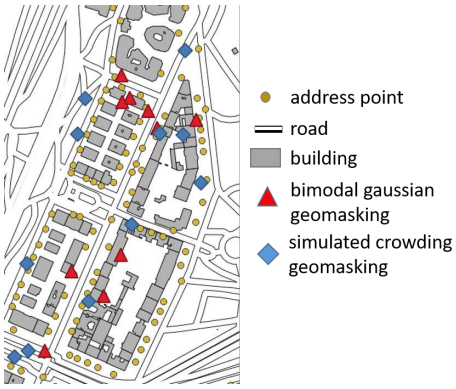

Figure 11. Difference between the enhanced bimodal gaussian perturbation, and the simulated crowding at the border of Paris. The perturbation pushes the points inside the city, while simulated crowding uses empty spaces at the border

If we want to maximize cluster preservation, the best method is the cluster oriented simulated 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, even with the best parameters (30m buffer excluded around initial points).

If we want to analyze rural areas, the block 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 simulated crowding also gives good results in rural areas because the large clusters let a lot of free space to generate masked points, far from the initial points.

If we want the best overall technique, the enhanced bimodal gaussian perturbation seems to be the more balanced one. It provides a good k-anonymity, with very few points badly masked in particular, and preserves local clusters really well. And with its density analysis step, it adapts well to rural and urban areas. However, we believe that the cluster oriented simulated crowding has the potential for this best technique spot, if we manage to improve its main defect, the lack of consistency, with 7% of points insufficiently masked. Figure 12 shows the results of both techniques on the same extract in Paris. The ability of simulated crowding to concentrate masked points inside the blocks explains why the median k-anonymity is so high compared to the perturbation technique. And for points outside clusters, the same perturbation is used so the results are similar.

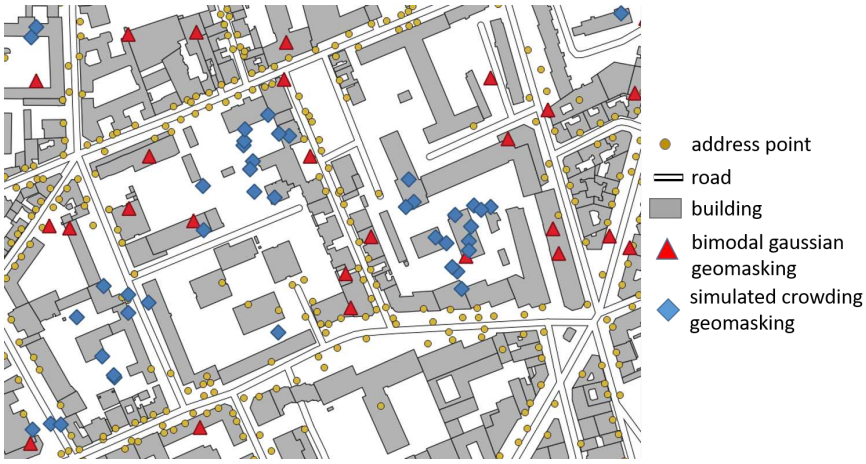

Figure 12. Results of the enhanced bimodal gaussian perturbation, and the cluster oriented simulated crowding on the same extract of the Paris area.

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 a strong anonymisation 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. And 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 discarded this approach.

Voronoï masking is that proved very effective in the recent publications on geomasking (Broen et al., 2021; Seidl et al., 2015). The principles of Voronoï masking is to generate a Voronoï 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 Voronoï 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 Elimination (AAE) (Kounadi & Leitner, 2016), and Adaptive Areal Masking (AAM) (Charleux & Schofield, 2020). Both methods are based on the principle of building anonymization polygons in which there is enough population to mask points. As a consequence, these polygons are larger in areas with low density. The main advantage of this method is its consistency in areas with heterogeneous density. Similarly to Voronoï masking, the drawback seems to be the processing of areas with very dense clusters, where the anonymization polygons might be too small. But, once again similarly to Voronoï 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 to 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 clusters spatial distribution. Overall, the enhanced bimodal gaussian perturbation resulted in the best comprimise between anonymity and cluster preservation. Then, we proposed a specific cluster oriented simulated crowding method, which fully preserves clusters while providing a satisfying anonymisation. 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 to test Voronoï 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 crowding method. One idea would be to exclude other areas where the number of surrounding addresses is low, for instance at the border of the cluster. One way could be to erode the polygon a little. 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 interesting to measure how much these methods may create false clusters. This problem is one of the reasons why the buffer

of the simulated crowding was not pushed beyond 30m, but it would be important to measure how much all the methods tend to create false clusters by displacing. 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.

Acknowledgement(s)

An unnumbered section, e.g. `\section*{Acknowledgements}`, may be used for thanks, etc. if required and included *in the non-anonymous version* before any Notes or References.

Disclosure statement

An unnumbered section, e.g. `\section*{Disclosure statement}`, may be used to declare any potential conflict of interest and included *in the non-anonymous version* before any Notes or References, after any Acknowledgements and before any Funding information.

Funding

An unnumbered section, e.g. `\section*{Funding}`, may be used for grant details, etc. if required and included *in the non-anonymous version* before any Notes or References.

Notes on contributor(s)

An unnumbered section, e.g. `\section*{Notes on contributors}`, may be included *in the non-anonymous version* if required. A photograph may be added if requested.

Nomenclature/Notation

An unnumbered section, e.g. `\section*{Nomenclature}` (or `\section*{Notation}`), may be included if required, before any Notes or References.

Notes

An unnumbered 'Notes' section may be included before the References (if using the `endnotes` package, use the command `\theendnotes` where the notes are to appear, instead of creating a `\section*`).

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 anal-

ysis while maintaining patient confidentiality. *International Journal of Health Geographics*, *18*(1), 30.

Allshouse, W. B., Fitch, M. K., Hampton, K. H., Gesink, D. C., Doherty, I. A., Leone, P. A., ... Miller, W. C. (2010, October). Geomasking sensitive health data and privacy protection: an evaluation using an E911 database. *Geocarto international*, *25*(6), 443–452.

Andrés, M. E., Bordenabe, N. E., Chatzikokolakis, K., & Palamidessi, C. (2013). Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 acm sigsac 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, March). Geographically masking health data to preserve confidentiality. *Statistics in Medicine*, *18*(5), 497–525. (Publisher: John Wiley & Sons, Ltd)

Boffet, A. (2000). Creating urban information for cartographic generalisation. In *International Symposium on Spatial Data Handling (SDH).* Beijing, China. (event-place: Beijing, China)

Broen, K., Trangucci, R., & Zelner, J. (2021, January). 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.

Cassa, C. A., Wieland, S. C., & Mandl, K. D. (2008, August). Re-identification of home addresses from spatial locations anonymized by Gaussian skew. *International Journal of Health Geographics*, *7*(1), 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*, 1–8. Retrieved 2020-11-19, from `https://www.nature.com/articles/s41586-020-2923-3` (Publisher: Nature Publishing Group)

Charleux, L., & Schofield, K. (2020, November). True spatial k-anonymity: adaptive areal elimination vs. adaptive areal masking. *Cartography and Geographic Information Science*, *47*(6), 537–549. Retrieved 2021-02-03, from `https://doi.org/10.1080/15230406.2020.1794975`

Chen, S., Li, Q., Gao, S., Kang, Y., & Shi, X. (2020, December). State-specific projection of COVID-19 infection in the United States and evaluation of three major control measures. *Scientific Reports*, *10*(1), 22429. (Number: 1 Publisher: Nature Publishing Group)

Clarke, K. C. (2016, February). A multiscale masking method for point geographic data. *International Journal of Geographical Information Science*, *30*(2), 300–315.

Curtis, A. J., Mills, J. W., & Leitner, M. (2006, October). Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina. *International Journal of Health Geographics*, *5*(1), 44.

Danis, K., Epaulard, O., Bénet, T., Gaymard, A., Campoy, S., Botelho-Nevers, E., ... Team, I. (2020, 04). Cluster of Coronavirus Disease 2019 (COVID-19) in the French Alps, February 2020. *Clinical Infectious Diseases*, *71*(15), 825-832. Retrieved from `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. Retrieved from `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.

Duckham, M., & Kulik, L. (2006). Location privacy and location-aware computing. In *Dynamic and Mobile GIS* (CRC Press ed., pp. 63–80).

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 Spa-*

tial Information Science, *19*, 510. Retrieved from http://josis.org/index.php/josis/article/view/510

Georgiadou, Y., de By, R. A., & Kounadi, O. (2019). Location privacy in the wake of the gdpr. *ISPRS International Journal of Geo-Information*, *8*(3). Retrieved from https://www.mdpi.com/2220-9964/8/3/157

Hampton, K. H., Fitch, M. K., Allshouse, W. B., Doherty, I. A., Gesink, D. C., Leone, P. A., ... Miller, W. C. (2010, November). Mapping Health Data: Improved Privacy Protection With Donut Method Geomasking. *American Journal of Epidemiology*, *172*(9), 1062–1069. Retrieved 2020-11-03, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2984253/

Jaccard, P. (1901, 01). Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, *37*, 241-72.

Katsomallos, M., Tzompanaki, K., & Kotzinos, D. (2019). Privacy, Space and Time: a Survey on Privacy-Preserving Continuous Data Publishing. *Journal of Spatial Information Science*, *19*, 57–103. Retrieved from http://josis.org/index.php/josis/article/view/493

Kim, J., Kwan, M.-P., Levenstein, M. C., & Richardson, D. B. (2021, January). How do people perceive the disclosure risk of maps? Examining the perceived disclosure risk of maps and its implications for geoprivacy protection. *Cartography and Geographic Information Science*, *48*(1), 2–20.

Kirby, R. S., Delmelle, E., & Eberth, J. M. (2017). Advances in spatial epidemiology and geographic information systems. *Annals of Epidemiology*, *27*(1), 1 - 9. (GIS and Spatial Methods in Epidemiology Symposium)

Kounadi, O., & Leitner, M. (2016, May). Adaptive areal elimination (AAE): A transparent way of disclosing protected spatial datasets. *Computers, Environment and Urban Systems*, *57*, 59–67.

Kwan, M.-P., Casas, I., & Schmitz, B. (2004, June). Protection of Geoprivacy and Accuracy of Spatial Information: How Effective Are Geographical Masks? *Cartographica: The International Journal for Geographic Information and Geovisualization*, *39*(2), 15–28. (Publisher: University of Toronto Press)

Lu, Y., Yorke, C., & Zhan, F. B. (2012, January). Considering Risk Locations When Defining Perturbation Zones for Geomasking. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *47*(3), 168–178. (Publisher: University of Toronto Press)

Peng, D., & Touya, G. (2017, October). Continuously Generalizing Buildings to Built-up Areas by Aggregating and Growing. In *Proceedings of 3rd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics (UrbanGIS'17).* Redondo Beach, CA, USA: ACM. Retrieved from http://dx.doi.org/10.1145/3152178.3152188

Piarroux, R. (2019). *Choléra. haïti 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. (Publisher: Taylor & Francis _eprint: 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, April). Privacy and spatial pattern preservation in masked GPS trajectory data. *International Journal of Geographical Information Science*, *30*(4), 785–800.

Seidl, D. E., Paulus, G., Jankowski, P., & Regenfelder, M. (2015, September). Spatial obfuscation methods for privacy protection of household-level data. *Applied Geography*, *63*, 253–263.

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.

Sweeney, L. (2002, October). k-Anonymity: A Model for Protecting Privacy. *International*

Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, *10*(05), 557–570. (Publisher: World Scientific Publishing Co.)

Wang, J., & Kwan, M.-P. (2020, March). Daily activity locations k-anonymity for the evaluation of disclosure risk of individual GPS datasets. *International Journal of Health Geographics*, *19*(1), 7.

Wang, Z., Liu, L., Zhou, H., & Lan, M. (2019, December). How Is the Confidentiality of Crime Locations Affected by Parameters in Kernel Density Estimation? *ISPRS International Journal of Geo-Information*, *8*(12), 544. (Number: 12 Publisher: Multidisciplinary Digital Publishing Institute)

Yong, S. E. F., Anderson, D. E., Wei, W. E., Pang, J., Chia, W. N., Tan, C. W., ... Lee, V. J. M. (2020). Connecting clusters of COVID-19: an epidemiological and serological investigation. *The Lancet Infectious Diseases*, *20*(7), 809 – 815.

Zandbergen, P. A. (2014, April). Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data. *Advances in Medicine*, *2014*, 567049. (Publisher: Hindawi Publishing Corporation)

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

7. Appendices

Any appendices should be placed after the list of references, beginning with the command `\appendix` followed by the command `\section` for each appendix title, e.g.

```
\appendix
\section{This is the title of the first appendix}
\section{This is the title of the second appendix}
```

produces:

Appendix A. This is the title of the first appendix

Appendix B. This is the title of the second appendix

Subsections, equations, figures, tables, etc. within appendices will then be automatically numbered as appropriate. Some theorem-like environments may need to have their counters reset manually (e.g. if they are not numbered within sections in the main text). You can achieve this by using `\numberwithin{remark}{section}` (for example) just after the `\appendix` command.

Note that if the `endfloat` package is used on a document containing any appendices, the `\processdelayedfloats` command must be included immediately before the `\appendix` command in order to ensure that the floats belonging to the main body of the text are numbered as such.

Appendix A. Troubleshooting

Authors may occasionally encounter problems with the preparation of a manuscript using LaTeX. The appropriate action to take will depend on the nature of the problem:

(i) If the problem is with LaTeX itself, rather than with the actual macros, please consult an appropriate LaTeX 2_ε manual for initial advice. If the solution cannot

be found, or if you suspect that the problem does lie with the macros, then please contact Taylor & Francis for assistance (`latex.helpdesk@tandf.co.uk`), clearly stating the title of the journal to which you are submitting.

(ii) Problems with page make-up (e.g. occasional overlong lines of text; figures or tables appearing out of order): please do not try to fix these using 'hard' page make-up commands – the typesetter will deal with such problems. (You may, if you wish, draw attention to particular problems when submitting the final version of your manuscript.)

(iii) If a required font is not available on your system, allow TEX to substitute the font and specify which font is required in a covering letter accompanying your files.

Appendix B. Obtaining the template and class file

B.1. Via the Taylor & Francis website

This article template and the `interact` class file may be obtained via the 'Instructions for Authors' pages of selected Taylor & Francis journals.

Please note that the class file calls up the open-source LATEX packages booktabs.sty, epsfig.sty and rotating.sty, which will, for convenience, unpack with the downloaded template and class file. The template optionally calls for natbib.sty and subfig.sty, which are also supplied for convenience.

B.2. Via e-mail

This article template, the `interact` class file and the associated open-source LATEX packages are also available via e-mail. Requests should be addressed to `latex.helpdesk@tandf.co.uk`, clearly stating for which journal you require the template and class file.

Figures

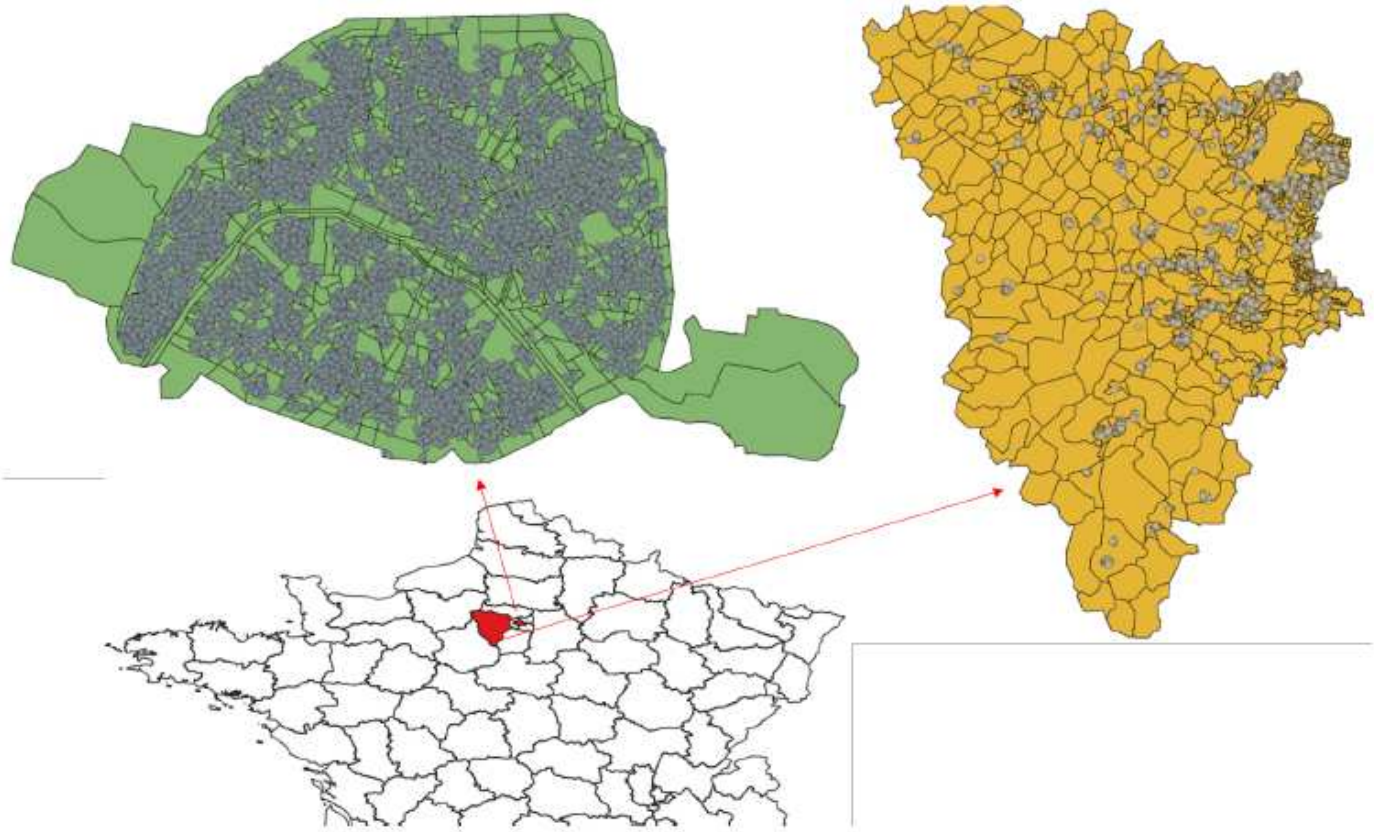

Figure 1

The generated synthetic COVID-19 data, for Paris on the left, and Yvelines on the right, both for three months. The gray point symbols are the cases, displayed on top of census administrative cells. The extract of France below shows the location of both regions and their respective size.

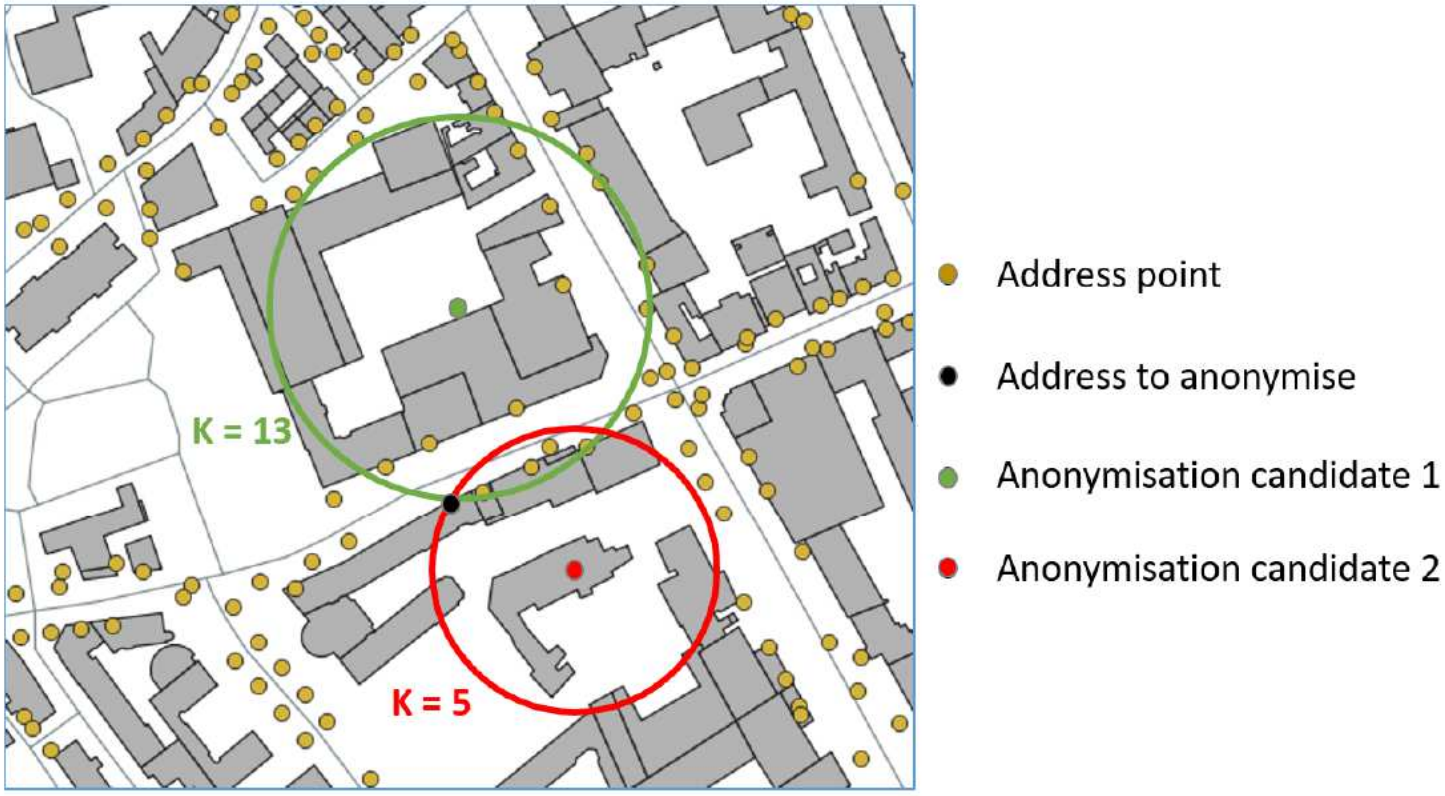

Figure 2

Principles of k-anonymity computation: anonymisation candidate 1 (in green) has 13 address points in the radius from the candidate to the address to anonymise (in black), so has k-anonymity value of 13. The other candidate position (in red) has a k-anonymity value of 5.

Figure 3

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

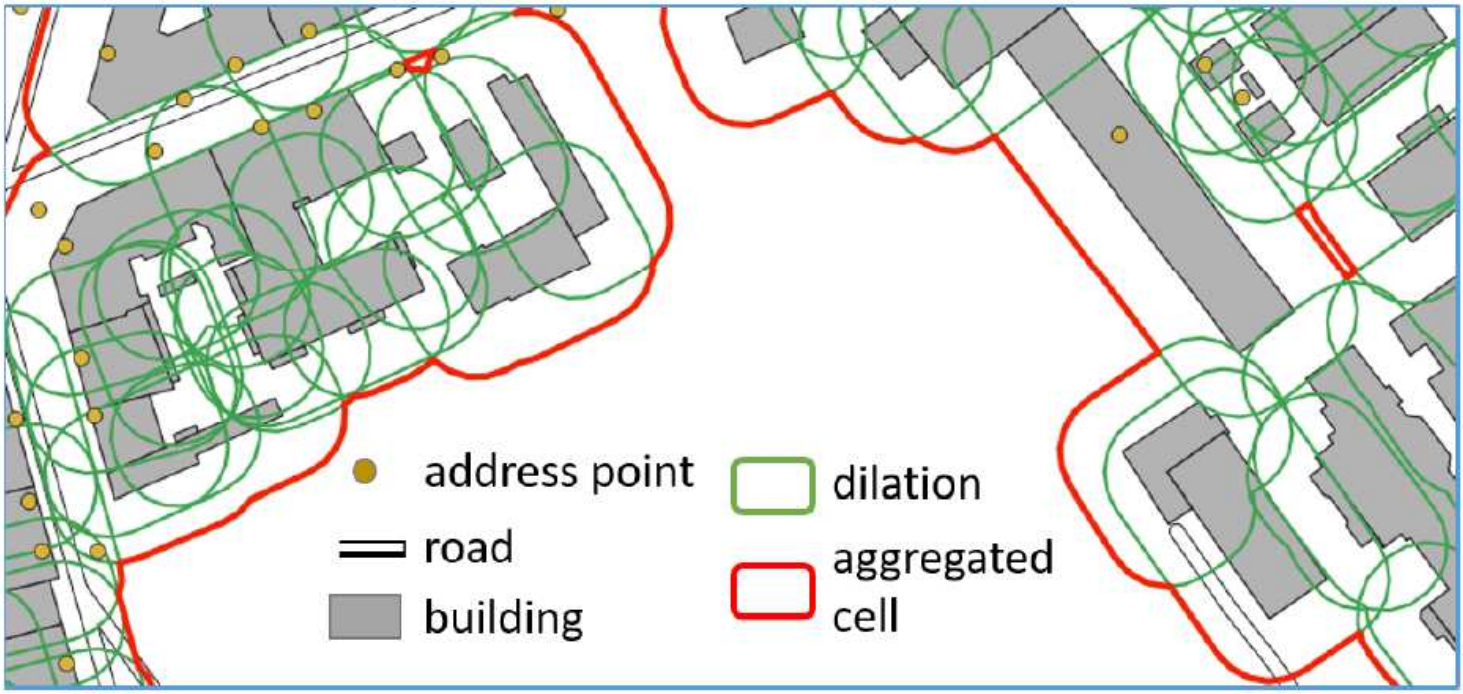

Figure 4

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

Figure 5

Results obtained with the enhanced bimodal gaussian perturbation in Paris.

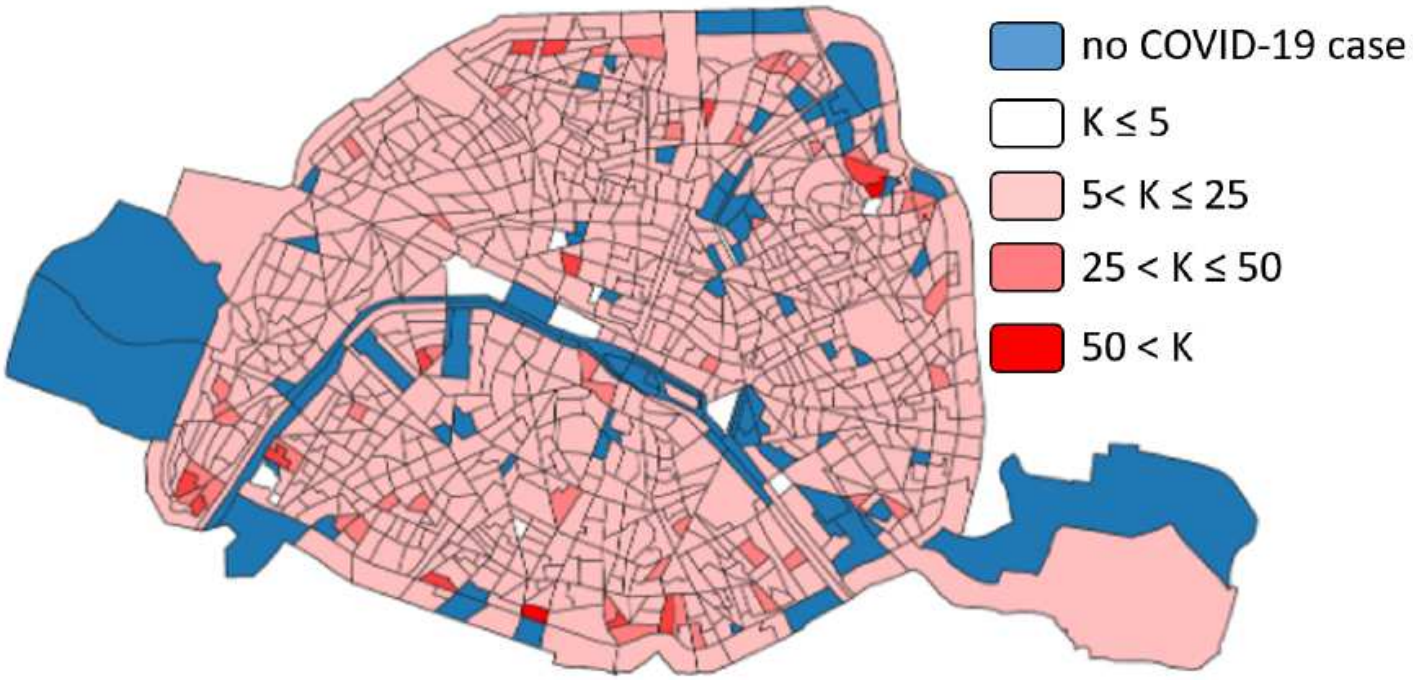

Figure 6

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.

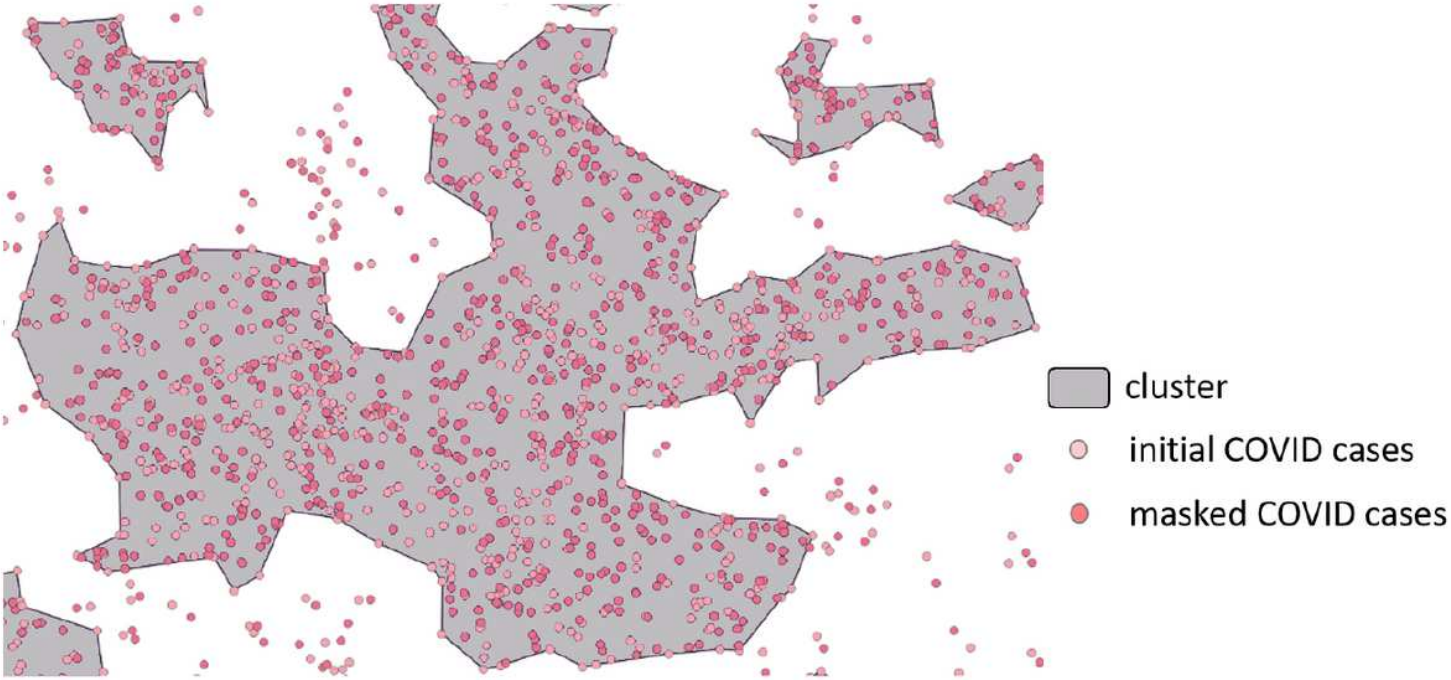

Figure 7

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

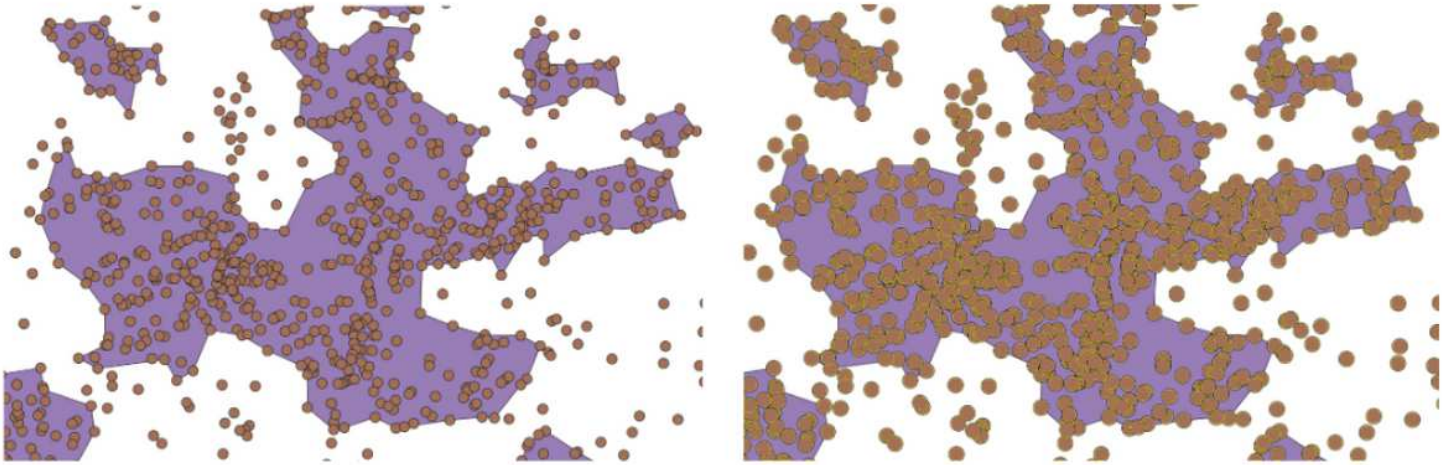

Figure 8

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).

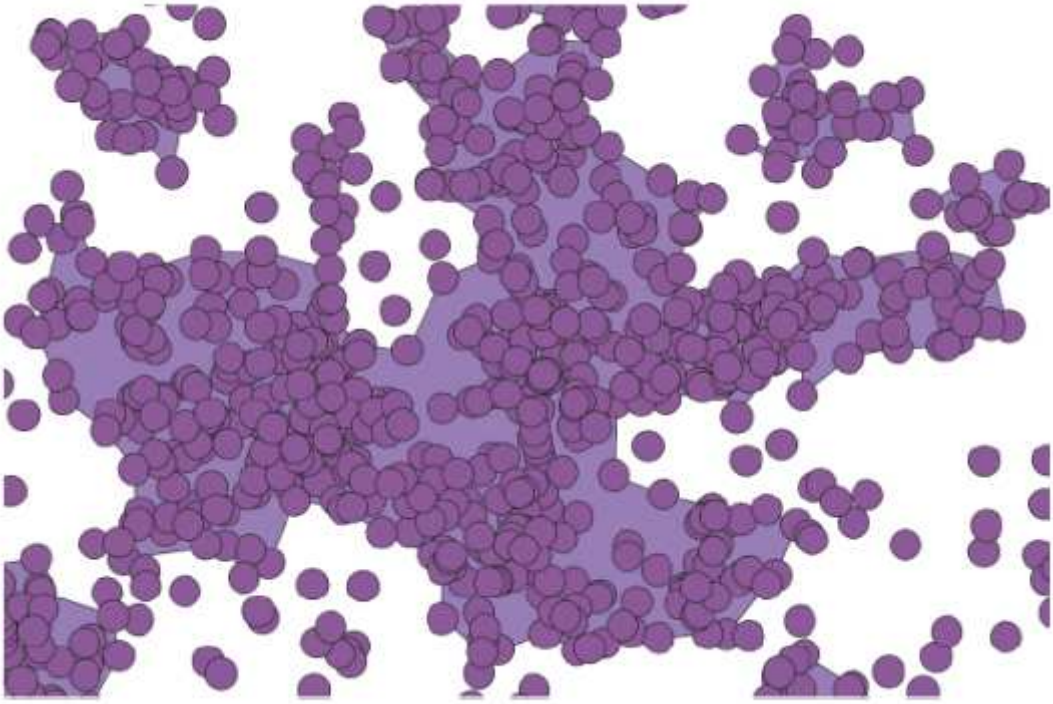

Figure 9

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

Figure 10

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.

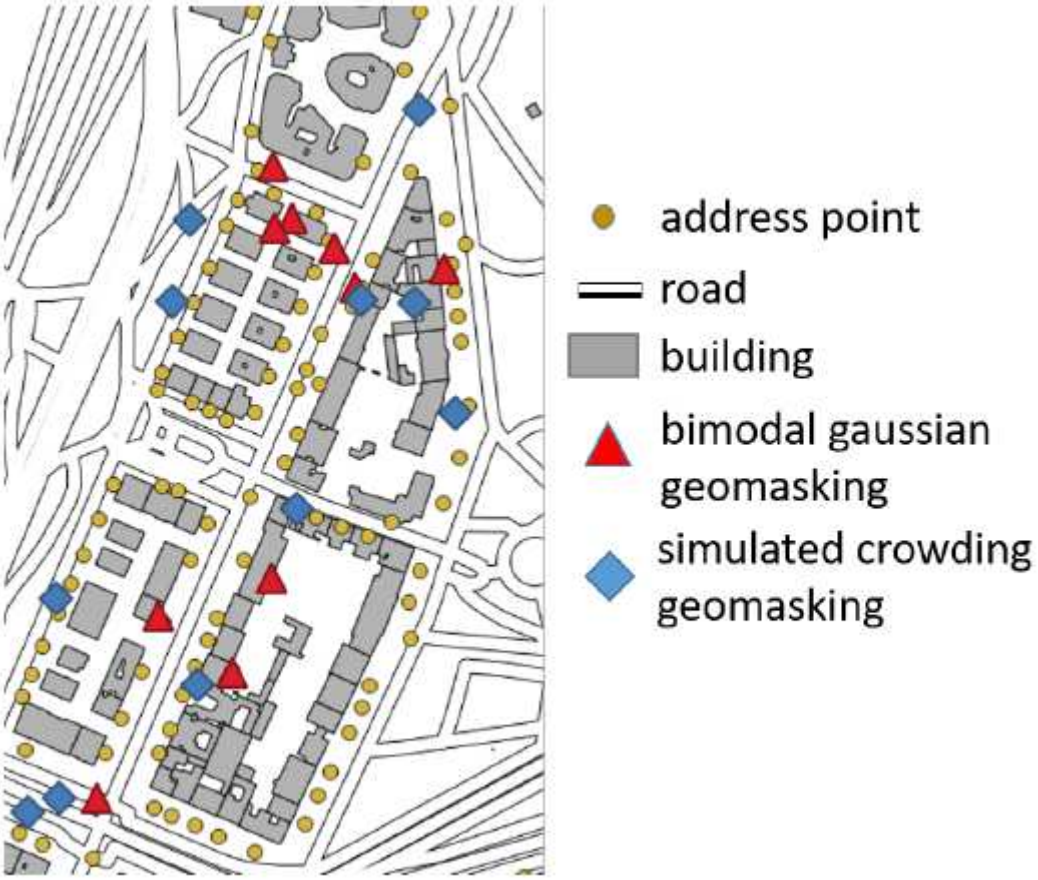

Figure 11

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

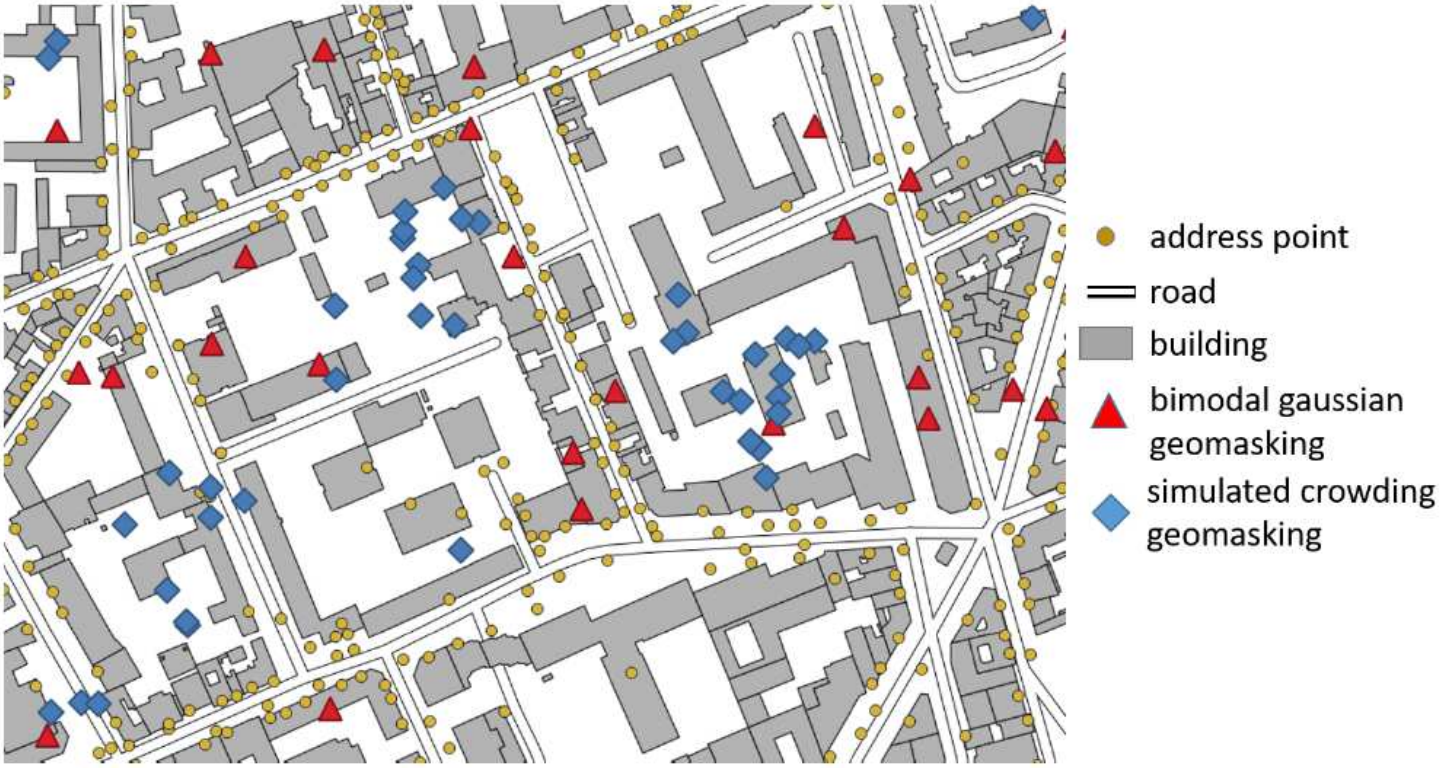

Figure 12

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