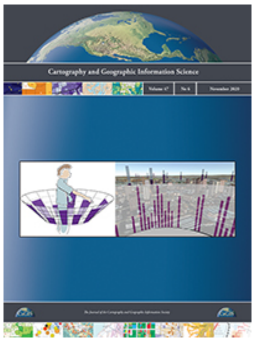


Cartography and Geographic Information Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tcag20

True spatial k-anonymity: adaptive areal elimination vs. adaptive areal masking

Laure Charleux & Katherine Schofield

CARTOGRAPHY AND GEOGRAPHIC INFORMATION SCIENCE
2020, VOL. 47, NO. 6, 537–549
https://doi.org/10.1080/15230406.2020.1794975

Taylor & Francis
Taylor & Francis Group

ARTICLE

True spatial k-anonymity: adaptive areal elimination vs. adaptive areal masking

Laure Charleux[a] and Katherine Schofield[b]

[a]Department of Geography, University of Minnesota, Duluth, MN, USA; [b]Department of Mechanical and Industrial Engineering, University of Minnesota, Duluth, MN, USA

ABSTRACT
Spatial anonymization of address points is critical to fields such as public health. There have been recent concerns about applications of geomasks that did not guarantee the level of k-anonymity theoretically expected. An analysis of the problem and a potential solution were previously proposed: Adaptive Areal Elimination (AAE). The present paper expands on AAE and proposes a modified version, Adaptive Areal Masking (AAM). A benchmark comparison of both methods is conducted, which shows that AAM outperforms AAE in most configurations tested. The discussion attempts to identify the application cases for which AAE might still be preferable and addresses documentation needs with both methods.

ARTICLE HISTORY
Received 31 December 2019
Accepted 8 July 2020

KEYWORDS
Spatial anonymization; geomasking; privacy; k-anonymity; spatial aggregation

Introduction

The need for geomasking

The analysis of individual-level spatial data has important applications in a variety of fields, ranging from criminology to phenology. In the fields of public health and epidemiology, such data are used, for instance, to detect spatial patterns in the spread of disease (Chua et al., 2019) or to measure access to care (Schofield & Charleux, 2018). Visualizing and understanding spatial distributions of populations, cases, or patients can help implement novel approaches to resource allocation, care, and intervention (Ajayakumar et al., 2019; Blatt, 2015; Mazumdar et al., 2014). Data aggregated within zip codes and census block areas can be useful but often do not provide enough granularity (spatial resolution) for many research and public health issues. The spatial aggregation of individual data points to large areas affects the ability to detect clusters or other spatial relationships (Armstrong et al., 1999; Fefferman et al., 2005; Kwan et al., 2004).

However, spatial data are strong identifiers, and a location such as the place of residence or work can easily lead to the identification of individuals, especially if further demographic information is disclosed in the dataset. This is especially true in areas of low population or case density, but even in areas of greater density, fine spatial resolution can identify individuals or let them be "reverse identified" via maps, social media, or hacking (Brownstein et al., 2005, 2006). Geographical masking methods are procedures to protect private or sensitive data. They displace points in the vicinity of their original location, with the double goal of preventing the identification of individuals while maintaining relevant spatial properties of the dataset (Armstrong et al., 1999; Zandbergen, 2014).

Many types of geomasks have been proposed over the years. Zandbergen (2014) offers a comprehensive survey and Kounadi and Leitner (2015) a useful classification. Many of these proposals displace points within variations of a circular neighborhood (including torus or doughnut masks). They have the advantage of easily allowing weighted implementations, where the radius of the neighborhood varies based on the underlying population density (i.e. Cassa et al., 2006; Hampton et al., 2010). This reflects the fact that points need to be displaced further to guarantee privacy in areas that are less densely populated. To determine the radius used locally, these adaptive geomasks rely on the concept of k-anonymity. It has been defined as indistinguishability from a k-1 number of other records or individuals (Samarati & Sweeney, 1998; Spruill, 1983; Sweeney, 2002; Zandbergen, 2014). This means that, with all record information taken into account, there are at least k-1 other records of individuals with the same characteristics. A k-anonymity level of 20 means that one target individual or record cannot be distinguished from 19 others. This corresponds to a 5% risk of re-identification.

The problem of k-anonymity overestimation

It has come to light, however, that some applications of adaptive geomasks overestimate the level of spatial k-anonymity provided for specific individuals, as they rely on the false assumption that the underlying population density is uniform within the areal units through which the data is

CONTACT Laure Charleux lcharleu@d.umn.edu

provided (e.g. census tracts; Allshouse et al., 2010; Kounadi & Leitner, 2016). To be more precise, the geomasking procedures, as designed, are theoretically sound in themselves. But they are designed to be used with underlying population data provided as punctual objects, which can be aggregated on-the-fly within adaptive euclidean neighborhoods, through count or sum. Instead, underlying population data are most commonly available as pre-aggregated attributes of areal enumeration units. This is a difference in measurement frameworks such as GIS analysts frequently encounter and that can often be satisfactorily dealt with through transformation operations (Chrisman, 1999). However, Chrisman also underlines that GIS software and analysts tend to understand GIS transformations "in terms of geometric primitives at the expense of the attribute rules implicit in different measurement frameworks," focusing "on syntax, not semantics" (1999, p. 632). It is precisely this problem that led to the flaw identified by Allshouse et al. (2010) and Kounadi and Leitner (2016). Analysts tried to adapt these geomask procedures to area-based population data by using polygon overlays between circular euclidean neighborhoods and population enumeration units. But this was done without paying close enough attention to the treatment of attributes, apportioning based on intersection area, and its underlying assumption: that the density be uniform within the original polygons.

Of course, this assumption is rarely true, and similar overlay operations are routinely conducted without inducing major flaws in analysis conclusions, as long as it is understood that the results of the attribute treatment in the overlay transformation are approximative and it is acceptable that the error may be very large in a few locations. This is clearly not the case with adaptive geomasks, however, where the goal is to displace points within neighborhoods that are just large enough to reach a desired k-anonymity level in order to minimize displacement. Approximation is not acceptable in that case. It will lead to local over-estimation of k-anonymity, as identified by Kounadi and Leitner (2016), in places where the underlying population is overestimated. It will also lead to excessive loss of information from the original point dataset, through greater than needed point displacement, in places where the underlying population is underestimated.

Reliable k-anonymity with aggregated reference population data

As Kounadi and Leitner (2016) explain, the only way to calculate k-anonymity accurately is to use a masking process at a resolution level equal to or lower than the resolution at which the underlying population data are available. Therefore, instead of displacing points within Euclidean neighborhoods, they propose to displace them within polygons of known reference population, pre-aggregated to guarantee that the reference population in each polygon is equal to or greater than k. They call this merging process "Adaptive areal elimination" (AAE). The next step is to displace points either to the centroids of the merged polygons ("Adaptive Point Aggregation" – APA), which is, in fact, an aggregation method, or randomly within the merged polygons ("Adaptive Random Perturbation" – ARP), which corresponds to the traditional definition of a geomask and which will be studied in the rest of this paper.

While we concur with Kounadi and Leitner's demonstration of the superiority of AAE over traditional adaptive geomasks to preserve k-anonymity (2016), we have two concerns with the method they propose:

- The first concern is excessive loss of information, as the aggregation process stops only once the entire study area is partitioned in non-overlapping polygons that all reach the k-anonymity threshold, which may lead to some points being displaced more than what is strictly necessary. For instance, a polygon whose population is large enough to ensure k-anonymity might still get merged with another one whose population is not, to guarantee the k-anonymity of points located in the latter. As a result, points located in the first polygon might be displaced to the second one, even though displacement within their original polygon would have been enough to maintain their k-anonymity.
- The second concern relates to processing time: the aggregation process consists of iteratively aggregating polygons that do not reach the k-anonymity threshold with their adjacent neighbor that shares the longest border. Topological queries typically use more resources than queries based on euclidean distance. Furthermore, the method requires all polygons in the study area to be part of the aggregation process, even if they do not contain points to anonymize. Finally, the process is intrinsically serial and cannot parallelize well for computation. Kounadi and Leitner (2016) do not address performance issues, and their example runs on a few thousand polygons only. A state-scale problem based on census blocks might involve millions, most of which may not contain any point to anonymize.

To address those concerns while maintaining the main advantage of the AAE method, i.e. guaranteeing a true level of spatial k-anonymity, we want to propose a modified version of the method. In our version, a specific anonymization polygon is built for each individual point by iteratively aggregating the population polygon it falls within with the polygon whose centroid is the closest in Euclidean distance, then the second closest, etc., until the k-anonymity threshold is reached. This process preserves information better than AAE because the aggregation stops when the k-anonymity threshold is reached, and because smaller polygons tend to be merged first, therefore minimizing the potential displacement of points. The process does not guarantee that the merged polygons are adjacent (even if they will be most of the time), but it uses fewer resources than computing adjacency. It also saves resources by not processing the polygons that do not contain any point to anonymize, which can be the vast majority of them, depending on the application. Finally, each point being treated in turn, this process parallelizes well. We propose to call this modified version of AAE "Adaptive Areal Masking," or AAM.

In the rest of the paper, we will compare, through sample applications, Kounadi and Leitner's AAE-ARP (2016) with our modified AAM, both in terms of processing performance and in terms of information loss (both approaches guarantee true spatial k-anonymity). It is expected that AAM will outperform AAE in most configurations. We will conclude by discussing guiding principles for choosing one method over the other for different applications, with different needs.

Methods

Description of sample application

We want to compare the computational performances of AAE and AAM across datasets of various sizes. The experimental setup is based on Census blocks in Minnesota. It includes three census-block datasets, providing both polygon geometries and 2010 population attribute data:

- A state-wide dataset, with almost 259,777 polygons
- A rectangular extract centered on the Twin-Cities metropolitan area, with 26,220 polygons
- A smaller rectangular extract centered on Minneapolis, with 2646 polygons.

For each of these, random datasets of 200, 2,000 and 20,000 points were generated, for a total of 9 point datasets. The generation procedure was intended to somewhat mimic the actual population distribution: 100 points were created at random locations within each inhabited polygon, and then the desired number of points for each sample was randomly selected, without further consideration for location. This resulted in denser sample points where the density of census blocks is higher.

AAE and AAM procedures were performed on each of these 9 point datasets and the corresponding polygon datasets for k-anonymization thresholds of 50, 500, and 5000, for a total of 27 runs for each procedure. K = 20 is a popular threshold, corresponding to a 5% risk of re-identification. However, this threshold is typically applied to populations controlled by various demographic attributes. To simplify our experiment plan, we do not consider any attribute, but the higher k-values applied to the total population are a good proxy for lower (and more realistic) k-values applied to populations controlled by increasingly complex combinations of attributes.

AAE and AAM implementations

The implementation of AAE described by Kounadi and Leitner (2016) was not made publicly available. A new implementation in Python using the arcpy package and embedded in an ArcGIS toolbox was created and is available at z.umn.edu/AAAnonymization (and also at http://doi.org/10.5281/zenodo.3906998). Two procedur es are used sequentially:

- aggregating original polygons to create anonymization polygons – this is the AAE procedure per se, and
- displacing points randomly within anonymization polygons – this procedure is optional and needed only for ARP.

The aggregation procedure includes the following steps:

(1) Create a list of polygon IDs for polygons with population >0 and < k-anonymity threshold
(2) Iterate over this list of IDs, in decreasing population order:
    a. Try to select the polygon with this ID (this will fail if the polygon has already been merged with another one in a previous iteration)
    b. If a polygon with this ID is found, and while the population remains less than the threshold, aggregation loop:
        i. Find the neighboring polygon with the longest shared boundary (target polygon)
        ii. Append the target polygon to the geometry of the polygon being processed and add their populations up
        iii. Delete the target polygon
(3) After all loops have run, delete polygons with population = 0 which might remain in the dataset, not having been used in mergers.

The care taken to not use uninhabited polygons as seeds for mergers is to ensure that points will not be displaced further than necessary within areas that do not improve the level of k-anonymity. Some polygons with population = 0 will be used in mergers as target polygons though, if they are located between inhabited polygons that need to be aggregated together. With the same concern for minimizing mergers, our implementation differs slightly from the original proposal in that, in the case of equilateral polygons that have same length borders with neighboring polygons, only one of the neighbors is picked randomly for aggregation (Figure 1).

The point displacement procedure (only for ARP) unfolds as follows:

For each point in the dataset:

```
(1) Identify the anonymization polygon it falls within
(2) Compute min and max X and Y coordinates of this polygon
(3) Loop, while new point not in polygon:
    a. Generate new point at random X and Y coordinates between their respective min and max
    b. Check whether new point is in polygon or not
(4) Replace the original point by new point.
```

Several point datasets can be anonymized using the same anonymization polygon dataset without having to run the polygon aggregation procedure again, provided that the population taken into consideration and the k-anonymization threshold are the same (Figure 2).

A python implementation of the AAM procedure, using the arcpy and numpy packages, is available in the same ArcGIS toolbox as the AAE implementation. The procedure executes as follows:

(1) Create an array containing the coordinates of each polygon centroid and the polygon population, as well as place holders for distance and cumulative population

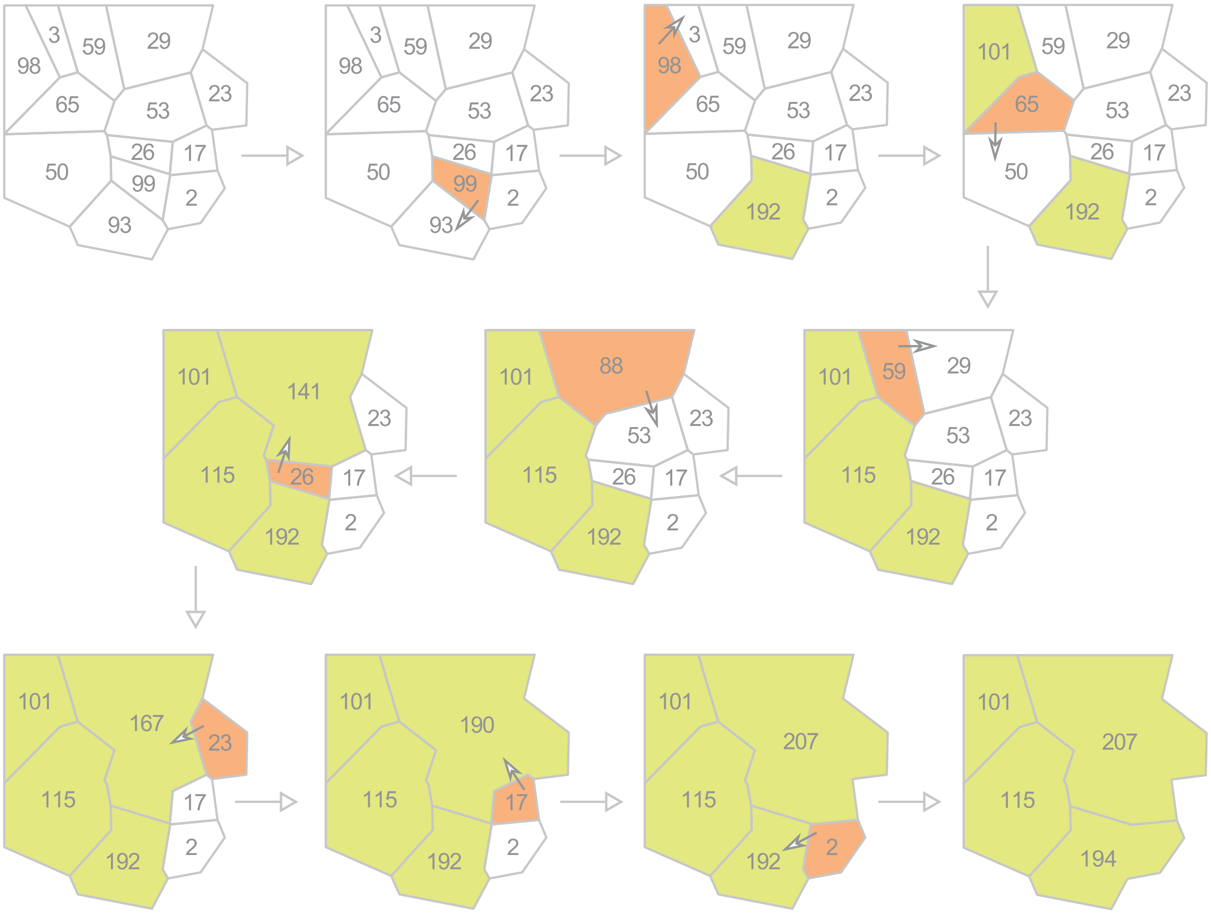


Figure 1. Illustration of the AAE aggregation process (k = 100). In each step, the orange polygon will be merged according to the arrow. Population counts are shown inside each polygon.

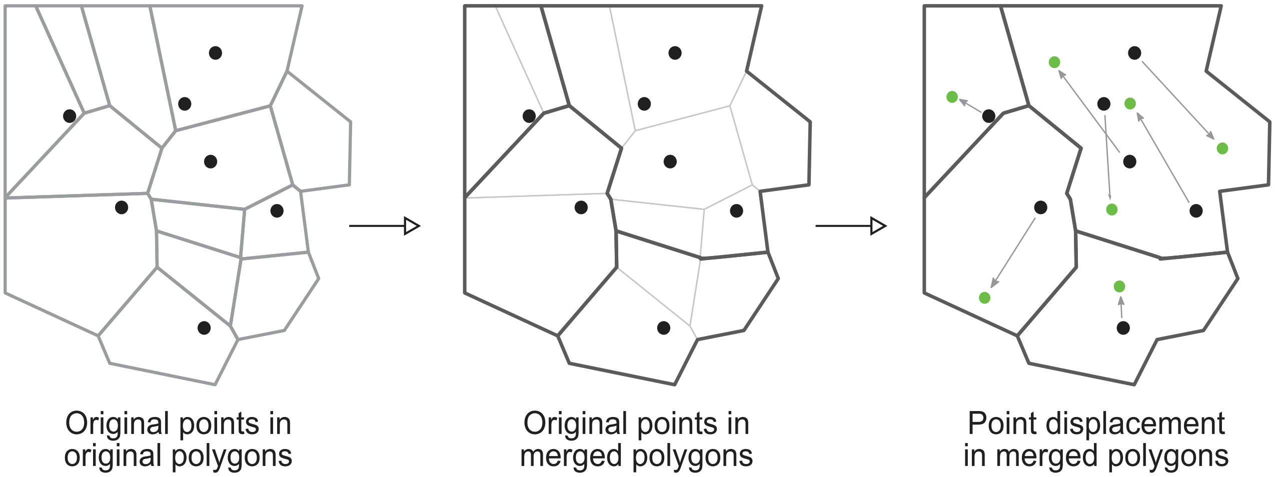

Original points in original polygons
Original points in merged polygons
Point displacement in merged polygons

Figure 2. Illustration of the ARP point displacement process.

```
(2) Identify the polygon that contains each point (spatial join)
(3) For each point in the dataset:
    a. Retrieve the geometry and the population of the containing polygon
    b. If the population is less than the k-anonymization threshold:
        i. Calculate Euclidean distance of point to all polygon centroids, using the numpy array previously created
        ii. Sort the array by increasing distance
        iii. Compute the cumulative population
        iv. Identify the index of the polygon where the cumulative population reaches the k-anonymization threshold
        v. Merge all the polygons with less or equal index value
    c. If the population reaches the k-anonymization threshold or after a merged anonymization polygon is created, displace the point:
        i. Compute min and max X and Y coordinates of the polygon
        ii. Loop, while new point not in polygon:
            1 Generate new point at random X and Y coordinates between their respective min and max
            2 Check whether new point is in polygon or not
        iii. Replace the original point by new point (Figure 3).
```

Measure of computational performance

The only computational performance that is considered here is the relative time of execution. All the procedures were performed on a standard personal computer equipped with an Intel(R) Xeon(R) CPUE E5-1603 v3 @ 2.80 GHz, 16 GB of RAM and running ArcGIS Pro 2.3 on Windows 10 Enterprise. The resulting running times will be examined to answer the following questions:

- How do running times increase with the number of polygons and points and with the k-anonymity threshold for both AAE and AAM?
- How do the running times of AAE and AAM compare for different scenarios?

Additionally, some preliminary tests were conducted to gauge the potential of improving the running time of AAM through parallel computing.

Measures of geographic performance

Not all geographical applications require the same level of locational precision. The geographical performances, that is the analytical impact of the loss of information resulting from the anonymization process, are assessed and compared for AAE-ARP and AAM in three different ways.

- Point displacement: the Euclidean distances between each sample point and its anonymized versions are calculated for each sample application. Descriptive statistics are examined to assess the impact of the anonymization method and threshold on point displacement.
- Density estimation: kernel density estimation procedures at various bandwidths are run on the original sample point distributions and on their anonymized versions for each sample application.

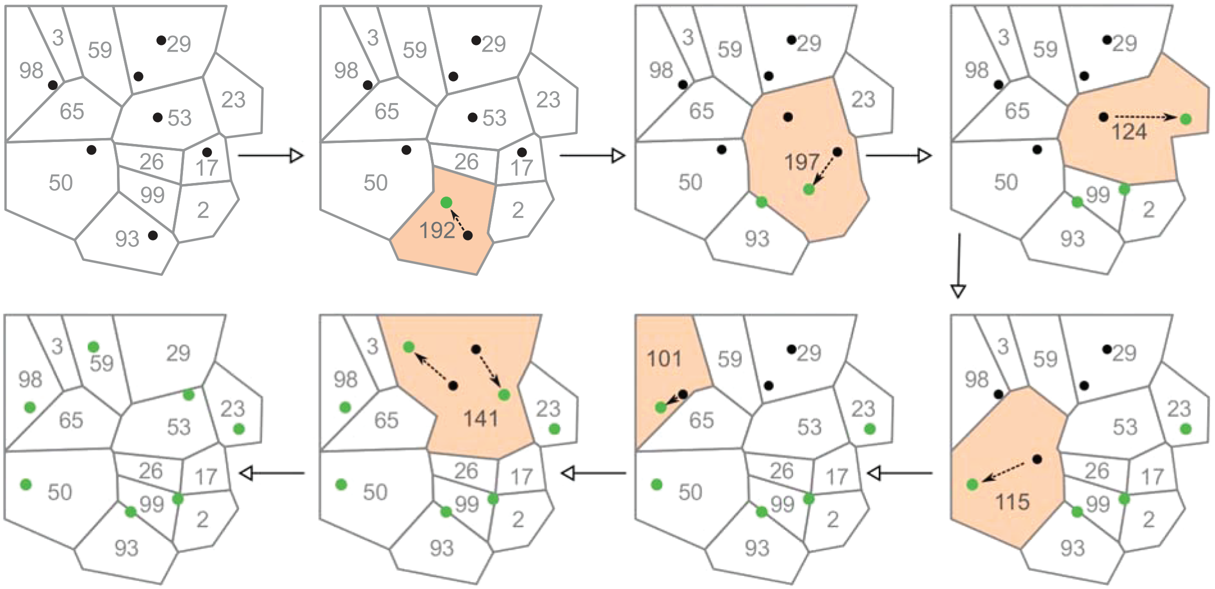


Figure 3. Illustration of the AAM point displacement process (k = 100).

The correlations between original and anonymized images are examined to assess the impact of the anonymization method and threshold on density estimates. Identification of clusters or "hotspots" of infectious disease is of increasing importance in epidemiology (Lessler et al., 2017). Different methods are used to identify different types of clusters but, at the fundamental level, they all rely on the density of occurrences relative to the density of a population. Therefore, the better an anonymization method preserves density patterns, the better it can be expected to preserve hotspot patterns.

- Accessibility: three destination point datasets are created by picking 50 random road intersections for each of the three test areas. The driving times to these destinations are computed for the original sample point distributions and their anonymized versions. For each sample application, both the average travel time to all 50 destinations and the travel time to the nearest destination are considered. The correlation between the results obtained from original and anonymized points is examined to assess the impact of the anonymization method and threshold on accessibility measures. Access to care is a major field of investigation in public health and includes a spatial component (Guagliardo, 2004). While a variety of spatial accessibility measures have been used, they all rely on some estimates of travel time to care locations. The better an anonymization method preserves travel times, the better it can be expected to preserve accessibility measures.

Results

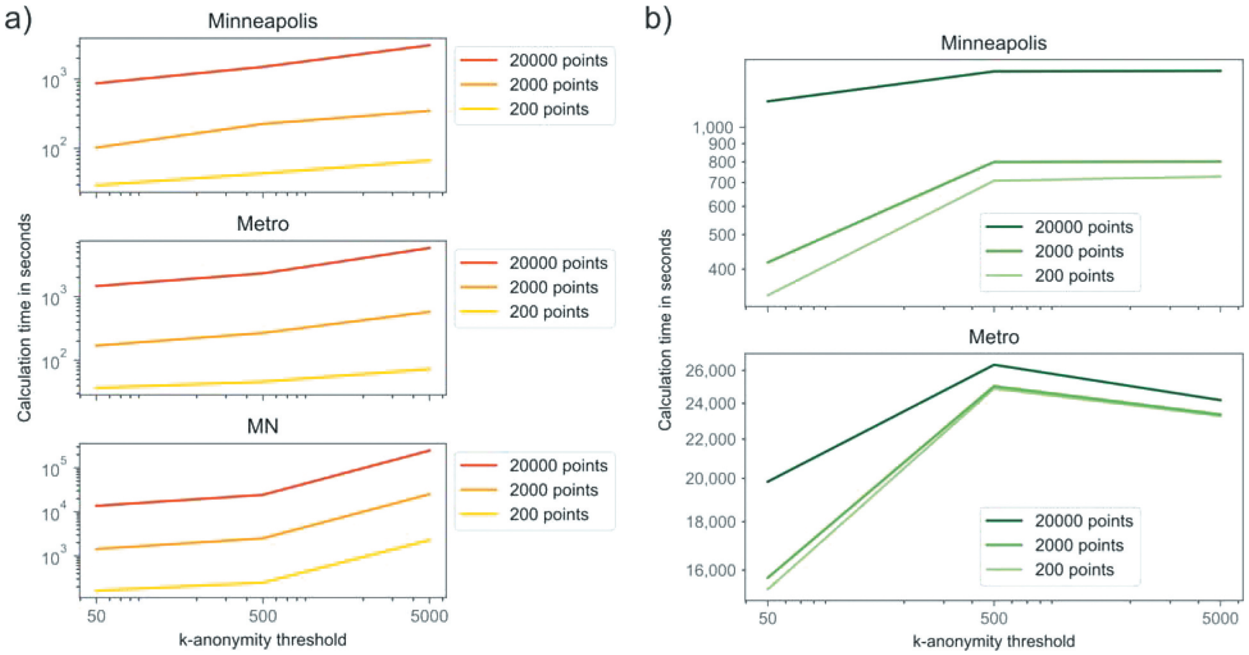
Computational performance

While AAM is a one-step procedure, calculation times for the AAE procedure are obtained by adding the time taken by the polygon merging procedure and the time taken by the random point displacement procedure (ARP) (Figure 4, Tables 1 and 2).

For AAE, the number of polygons to merge is the principal factor impacting performance, with calculation times multiplied by 30–50 between the calculations on the Minneapolis area polygons and the calculations on the Metro area polygons (10 times more polygons). Running this merging procedure on the state-wide polygon dataset turned out to be impractical on the desktop computer: the procedures were interrupted after five days, with estimated run times reaching dozens of days.

The number of points to displace is the second factor impacting performance, with a non-linear effect. The displacement procedure is 3–6 times longer with 2,000 points compared to 200, and 6–9 times longer for 20,000 points compared to 2,000.

The influence of the anonymity threshold is ambivalent. On the one hand, a higher threshold requires to merge more polygons and the merging procedure takes about twice as long when going from k = 50 to k = 500. But there is a plateau effect and the merging procedure takes a similar time between k = 500 and k = 5000. On the other hand, displacing points in bigger polygons go faster, and the displacement procedure takes 1.1 to 4 times longer for k = 50 compared to k = 5000. As a result,



Logarithmic scales for both axes

Logarithmic scales for both axes

Figure 4. Calculation times for AAM (a) and AAE (b).

the total calculation times increase significantly between k = 50 and k = 500 for both the Minneapolis area and the Metro area samples, but they plateau between k = 500 and k = 5000 for the Minneapolis area and even decrease for the Metro area.

Calculation times are a lot more stable and predictable with AAM. A regression model on logged variables shows that the number of points, the number of polygons, and the anonymity threshold explain 90% of the variance of the calculation times, with t-values of 11, 8, and 5, respectively (Figure 5).

For 9 of the original 27 test configurations, on state-wide census-blocks, AAE could not run in a practical time on the desktop computer, while AAM took up to almost three days. In the 18

Table 1. Calculation times for AAM in minutes.

	Sample (all values in minutes)		
No. of points and k values	Minneapolis	Metro	MN
k = 50			
200 points	0.48	0.62	2.78
2000 points	1.78	2.83	23.97
20,000 points	14.47	24.42	228.37
k = 500			
200 points	0.73	0.77	4.13
2000 points	3.77	4.48	41.85
20,000 points	25.02	38.65	306.75
k = 5000			
200 points	1.12	1.22	38.18
2000 points	5.80	9.62	7.60
20,000 points	51.20	96.78	4112.35

Table 2. Calculation times for AAE in minutes.

	Sample (all values in minutes)					
	Minneapolis			Metro		
No. of points and k values	*Merging*	*Displacement*	*Total*	*Merging*	*Displacement*	*Total*
k = 50						
200 points	4.85	0.78	5.63	253.20	1.35	254.55
2000 points		2.12	6.97		8.58	261.78
20,000 points		14.87	19.72		77.45	330.65
k = 500						
200 points	11.10	0.70	11.80	410.88	0.80	411.68
2000 points		2.23	13.33		3.20	414.08
20,000 points		12.82	23.92		25.50	436.38
k = 5000						
200 points	11.45	0.68	12.13	387.18	0.75	387.93
2000 points		1.92	13.37		2.08	389.26
20,000 points		12.55	24.00		15.95	403.13

configurations for which we can compare calculation times between AAE and AAM, AAE beats AAM only twice, with low polygon numbers, high point numbers, and high anonymity thresholds. Because AAM can merge the same polygons again and again when the dataset includes multiple points in similar locations, it can be outperformed by AAE, which merges polygons only once, in configurations with a high density of points per polygon.

This does not take into consideration, however, that the AAM algorithm can be implemented using parallel processing, which is possible only for the point displacement procedure of AAE but not for the merging procedure. The gain in calculation time with parallel processing depends on the number of cores available. A few of the testing configurations were run with 2, 3, and 4 cores (the maximum available on the test machine) and the algorithm behaved as expected, with some overhead processing time to set up the parallelization and merge the results (of the order of a few minutes, depending on the configuration), and the anonymization processing time divided by the number of cores. This means that the larger the problem and the number of cores, the more time is saved with parallelization. Parallelizing a small problem with a small number of cores actually loses time.

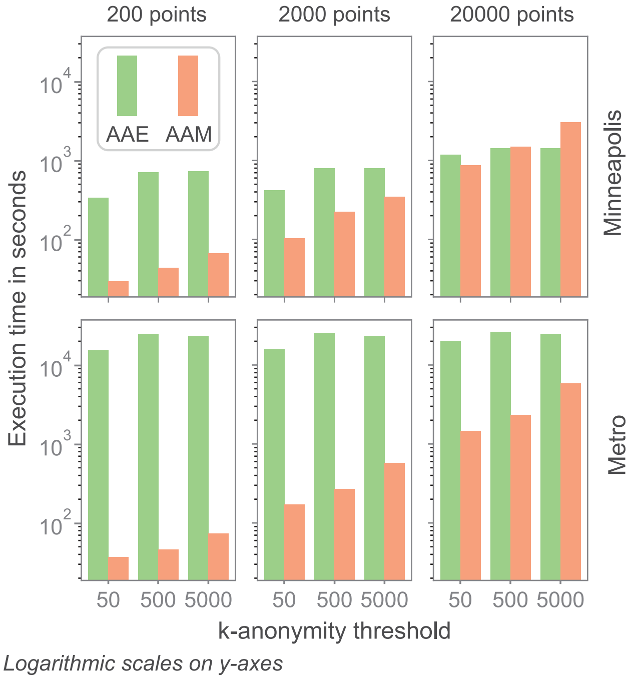


Figure 5. Direct comparison of calculation times for AAE and AAM.

Geographic performance

Point displacement

The k-anonymity threshold has a clear impact on the average point displacement (see Figure 6), as more polygons need to be fusioned to reach higher thresholds, leading to wider areas where points can be displaced. As expected, the number of points to be anonymized has no bearing on the point displacement: for a given study area, the lines depicting different numbers of points are almost undistinguishable. The study area matters though: for both AAE and AAM, points in the Metro samples were displaced further away than points in the Minneapolis sample. And for AAM, points in the statewide samples were displaced even more. Rather than with the number of polygons in each study area, this has to do with the differences in population density in each study area. Anonymization polygons for points located in rural areas need to be larger than those for points located in urban areas to reach an identical anonymization threshold.

As for the anonymization algorithm, AAE-ARP systematically led to larger average point displacement

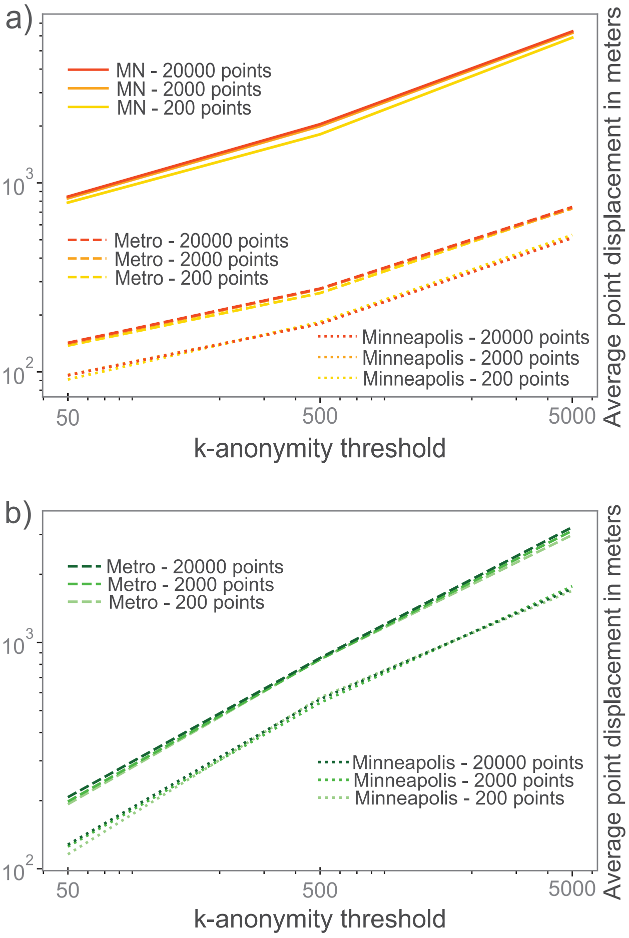


Figure 6. Average point displacement for AAM (a) and AAE (b).

than AAM (see Figure 7). Differences get larger when the anonymity threshold increases and the population density decreases, from 1.3 to 1.4 times larger for the Minneapolis samples with k = 50, to 32–35 times larger for the Metro samples with k = 5000. Interestingly, the length of displacement is also more consistent with AAM than with AAE-ARP, with coefficients of variation 6%-108% larger for AAE. Differences in consistency are greater with lower anonymity thresholds and with lower population densities.

Density estimation

The graphs in Figure 8 show that, as expected, the level of correlation between kernel density estimations produced from original points and those produced from anonymized points increases with the bandwidth used and decreases as the k-anonymity threshold increases, for all spatial configurations tested. They also show that correlation levels are better for higher sample point densities. Finally, they show that, for a same level of k-anonymity, AAM systematically provides a higher level of correlation than AAE-ARP. This is especially critical for higher k-anonymity thresholds: AAM curves for k = 5000 tend to be aligned with the AAE-ARP curves for k = 500.

Accessibility

As expected, the correlation between accessibility indicators computed from original datasets and indicators computed from anonymized datasets diminishes as the k-anonymity threshold increases. Correlation levels appear largely independent of the number of sample points, and they increase with the size of the geographic area considered, as distances traveled increase. Correlation levels are higher when calculated on 50 destinations than when calculated on the closest destination only. In both cases, however, AAM performs systematically better than AAE. Differences can be dramatic in the worst scenarios: for example, datasets in the Minneapolis area anonymized at the 5000 k-anonymity threshold: ~0.95 vs. ~0.72 for 50 destinations, and ~0.58 vs. ~0.27 for the closest destination only (Figure 9).

Discussion

Performance

When choosing between AAE and AAM for specific applications, both computing and geographical performances should be considered. While AAM performs systematically better than AAE from a geographical perspective, differences are sometimes small:

- Difference of performance in point displacement is smaller when point density is high and k-anonymization threshold is low
- Likewise, density estimations are more similar when point density is high and k-anonymization threshold is low
- Accessibility indicators are more similar when the k-anonymization threshold is low and the study area is large.

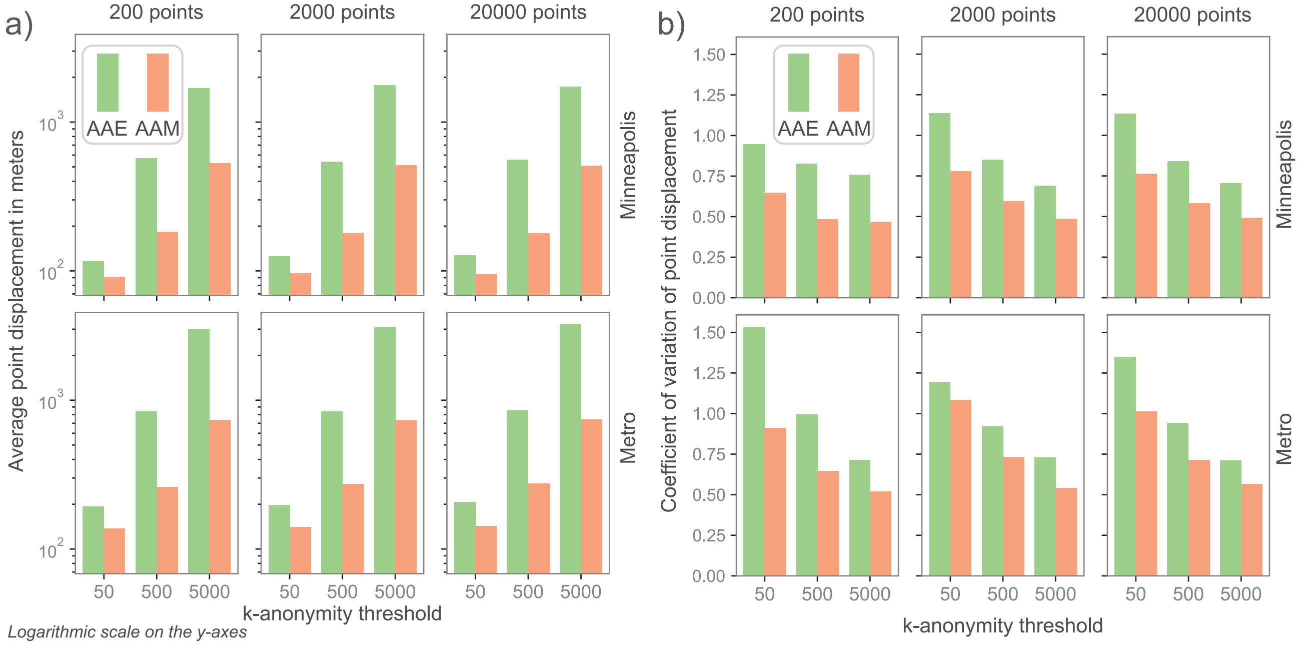


Figure 7. Direct comparison of point displacement for AAE and AAM: averages (a) and coefficients of variation (b).

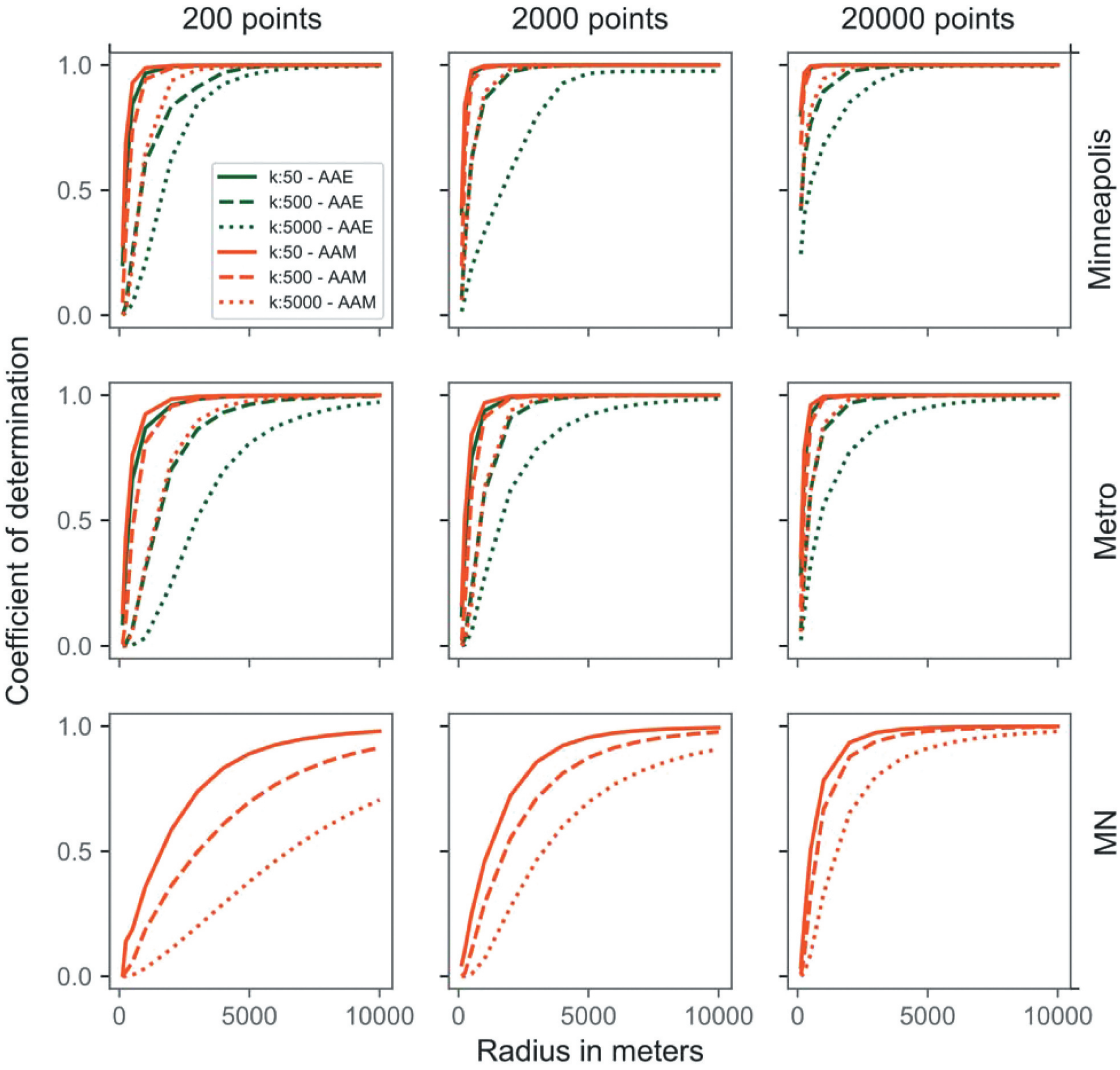


Figure 8. Correlation of density estimations between original and anonymized data sets.

On the computing side, things are a bit more complex, because of the different behaviors of the two procedures involved in AAE. Overall AAM performs better than AAE in most configurations tested, but in configurations where the number of points to anonymize is high enough and the number of population polygons is low enough, AAE can outperform AAM, even taking potential parallelization into account. And since those are high-density configurations, the point displacement or density estimation performances will be only slightly worse than with AAM. But accessibility calculations could be significantly worse, which means AAM might still be preferable for that purpose.

An additional consideration is that some anonymization tasks might need repetition on new point datasets while the reference population does not change. For instance, one might want to anonymize weekly cases of influenza using the same underlying population data. AAE provides the possibility to re-use anonymization polygons previously calculated, while AAM does not (see uncertainty considerations below), which could make AAE the preferred option in that case, provided that the number of population polygons is not too large to start with.

Indeed, our experiment has shown that the resources needed for the aggregation phase of AAE can quickly become unreasonable: a good desktop computer was not able to process a quarter million polygons. Of course, more powerful computers might be available, but so-called “supercomputers” derive their power

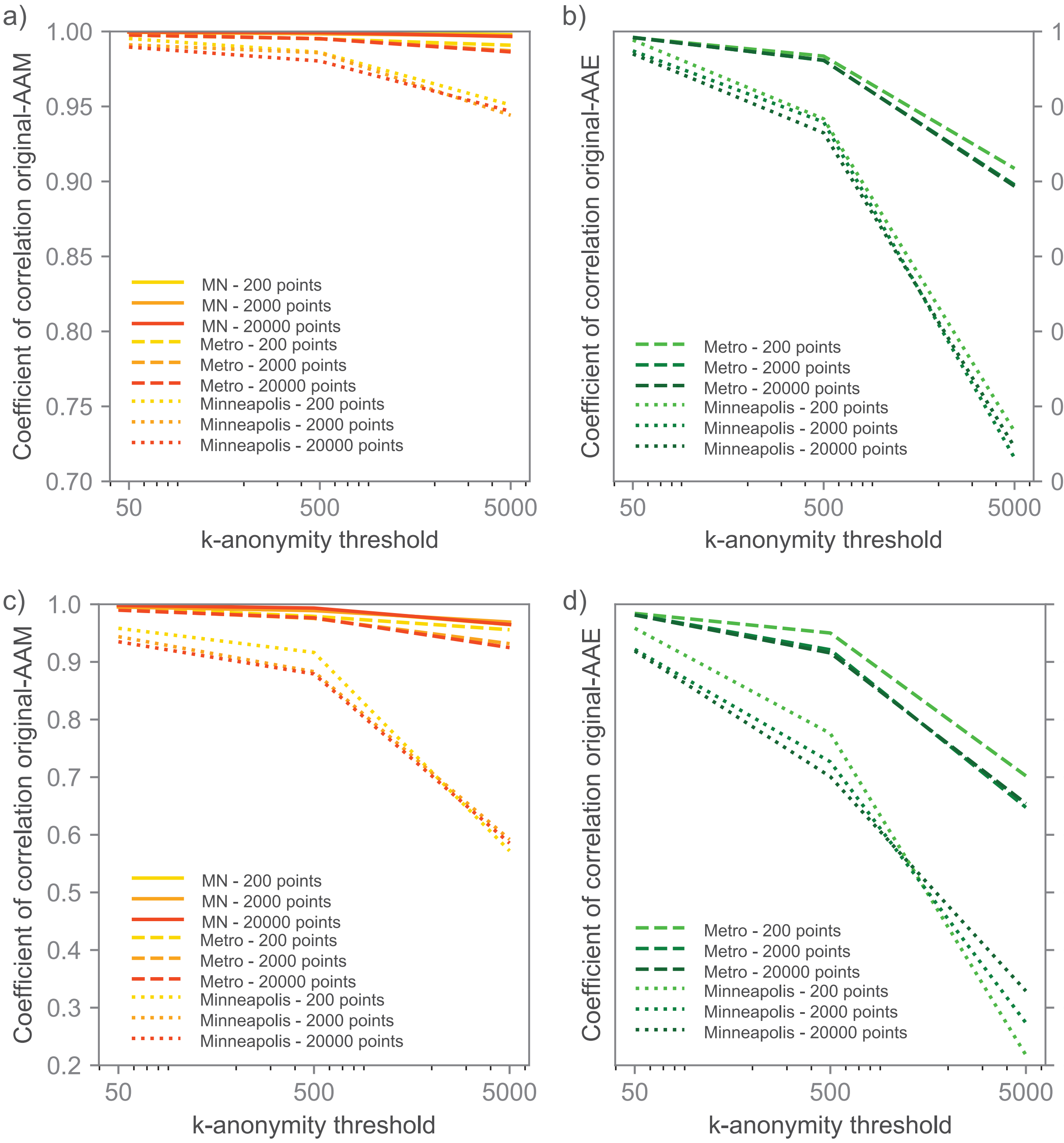


Figure 9. Correlations of driving times to 50 destinations between the original datasets and datasets anonymized with AAM (a) or AAE (b), and of driving times to the closest destination, for AAM (c) or AAE (d).

mostly from parallelization, which cannot apply to this intrinsically serial procedure. One could also try to partition the input population polygons and to process polygon aggregation for each partition in parallel (for instance, we could have processed state-wide census blocks county by county). This would lead to suboptimal polygon aggregation near partition lines and further degrade geographic performance, however. Overall, considering that applications with high numbers of population polygons are more likely to involve a large proportion of polygons not containing any point to anonymize, it is most probable that such applications will be better processed through AAM, which also has better geographical performances.

Uncertainty considerations

Both AAE and AAM displace points, introducing locational uncertainty. A major advantage of the AAE method is that this uncertainty can be "transparently" documented by disclosing to end users the dataset of non-overlapping anonymization polygons, as well as their population attributes, without increasing the risk of re-identification.

This is not possible for AAM however, as it produces individual anonymization polygons for each point, which may overlap. Imagine two anonymized point locations with their corresponding anonymization polygons, the first one contained by the second. This would mean that the actual location of the second point is within the area of the second polygon that lies beyond the first one. Since the population of this area has to be well below the k-anonymity threshold (otherwise the area corresponding to the first polygon would not have been appended), the risk of re-identification is increased well below the desired k-anonymity level. Anonymization polygons generated by the AAM procedure should therefore not be disclosed to end users. Furthermore, characteristics such as their population or area should not be disclosed either, as this information also increases the risk of re-identification if the population data are publicly available, as is the case for census data.

Nevertheless, it should be a best practice for anonymizers using AAM to calculate and provide to end users a few aggregate indicators describing the uncertainty introduced by the anonymization procedure. It is very simple for instance, to compute a few statistics describing the distribution of the point displacement distances. Furthermore, these statistics should be provided for each combination of attributes used to control the reference population. Indeed, points denoting the locations of people with characteristics that are locally rare will end up being displaced further away on average than those corresponding to people with more common characteristics. This is akin to working at a lower resolution for under-represented populations.

To be clear, health data challenges related to under-represented populations are not specific to this procedure, not even to anonymization in general. For instance, calculating a simple prevalence or death rate might prove impossible for very specific populations, as ironically illustrated by the overwhelming number of maps featuring grayed-out counties with "insufficient data" in a hyper-specialized atlas of heart disease-related mortality in women from diverse racial and ethnic backgrounds (Casper, 2000). But at least the data resolution problem is blatantly visible to readers in this case. In the case of anonymized point datasets featuring different subpopulations, however, this difficulty could easily be overlooked and analysts could reach flawed conclusions when trying to compare these subpopulations.

It is therefore critical that anonymized datasets be provided with extensive documentation of the point displacement characteristics for each subpopulation so that analysts can take into account different uncertainty levels when fitting models, assessing probabilities, etc. Still, attention should be paid to the degrees of freedom for each subpopulation, in order to restrict the number of statistics reported when needed, to prevent the reconstitution of the actual point displacement distances, which would gravely compromise the level of k-anonymity.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The work for this paper was supported by a Pilot Project Grant through the Midwest Center for Occupational Health and Safety (MCOHS) Education and Research Center, University of Minnesota (UMN), Subaward NIOSH T42OH008434.

ORCID

Laure Charleux http://orcid.org/0000-0001-6476-2715
Katherine Schofield http://orcid.org/0000-0001-8561-9379

Data Availability

The data and software that support the findings of this study are available at the following permanent links:

- ArcGIS Toolbox: http://doi.org/10.5281/zenodo.3906998
- OGC geopackage of the sample data: http://doi.org/10.5281/zenodo.3906976

References

Ajayakumar, J., Curtis, A. J., & Curtis, J. (2019). 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

Armstrong, M. P., Rushton, G., & Zimmerman, D. L. (1999). Geographically masking health data to preserve confidentiality. *Statistics in Medicine*, *18*(5), 497–525. https://doi.org/10.1002/(SICI)1097-0258(19990315)18:5<497::AID-SIM45>3.0.CO;2-%23

Blatt, A. J. (2015). *Health, science, and place: A new model*. Springer International Publishing. https://doi.org/10.1007/978-3-319-12003-4

Brownstein, J. S., Cassa, C. A., Kohane, I. S., & Mandl, K. D. (2005). Reverse geocoding: Concerns about patient confidentiality in the display of geospatial health data. *AMIA … Annual symposium proceedings. AMIA symposium* (pp. 905).

Brownstein, J. S., Cassa, C. A., Kohane, I. S., & Mandl, K. D. (2006). An unsupervised classification method for inferring original case locations from low-resolution disease maps. *International Journal of Health Geographics*, *5*(1), 56. https://doi.org/10.1186/1476-072X-5-56

Casper, M. L. (2000). *Women and heart disease; an atlas of racial and ethnic disparities in mortality (cdc:12169)* (CDC report). https://stacks.cdc.gov/view/cdc/12169

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

Chrisman, N. (1999). A transformational approach to GIS operations. *International Journal of Geographical Information Science*, *13*(7), 617–637. https://doi.org/10.1080/136588199241030

Chua, J. L., Ng, L. C., Lee, V. J., Ong, M. E. H., Lim, E. L., Lim, H. C. S., Ooi, C. K., Tyebally, A., Seow, E., & Chen, M. I.-C. (2019). Utility of spatial point-pattern analysis using residential and workplace geospatial information to localize potential outbreak sources. *American Journal of Epidemiology*, *188*(5), 940–949. https://doi.org/10.1093/aje/kwy290

Fefferman, N. H., O'Neil, E. A., & Naumova, E. N. (2005). Confidentiality and confidence: Is data aggregation a means to achieve both? *Journal of Public Health Policy*, *26*(4), 430–449. https://doi.org/10.1057/palgrave.jphp.3200029

Guagliardo, M. F. (2004). Spatial accessibility of primary care: Concepts, methods and challenges. *International Journal of Health Geographics*, *3*(1), 3. https://doi.org/10.1186/1476-072X-3-3

Hampton, K. H., Fitch, M. K., Allshouse, W. B., Doherty, I. A., Gesink, D. C., Leone, P. A., Serre, M. L., & Miller, W. C. (2010). Mapping health data: Improved privacy protection with donut method geomasking. *American Journal of Epidemiology*, *172*(9), 1062–1069. https://doi.org/10.1093/aje/kwq248

Kounadi, O., & Leitner, M. (2015). Spatial information divergence: Using global and local indices to compare geographical masks applied to crime data. *Transactions in GIS*, *19*(5), 737–757. https://doi.org/10.1111/tgis.12125

Kounadi, O., & Leitner, M. (2016). Adaptive areal elimination (AAE): A transparent way of disclosing protected spatial datasets. *Computers, Environment and Urban Systems*, *57*, 59–67. https://doi.org/10.1016/j.compenvurbsys.2016.01.004

Kwan, M.-P., Casas, I., & Schmitz, B. (2004). 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. https://doi.org/10.3138/X204-4223-57MK-8273

Lessler, J., Azman, A. S., McKay, H. S., & Moore, S. M. (2017). What is a hotspot anyway? *The American Journal of Tropical Medicine and Hygiene*, *96*(6), 1270–1273. https://doi.org/10.4269/ajtmh.16-0427

Mazumdar, S., Konings, P., Hewett, M., Bagheri, N., McRae, I., & Del Fante, P. (2014). Protecting the privacy of individual general practice patient electronic records for geospatial epidemiology research. *Australian and New Zealand Journal of Public Health*, *38*(6), 548–552. https://doi.org/10.1111/1753-6405.12262

Samarati, P., & Sweeney, L. (1998). *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression* (Technical Report SRI-CSL-98-04). Computer Science Laboratory, SRI International. http://www.csl.sri.com/papers/sritr-98-04/

Schofield, K., & Charleux, L. (2018). 1019 Injury severity, return to work, and disability in occupational injury – Does community health play a role? *Occupational and Environmental Medicine*, *75*(Suppl 2), A540–A540. https://doi.org/10.1136/oemed-2018-ICOHabstracts.1530

Spruill, N. L. (1983). The confidentiality and analytic usefulness of masked business microdata. *Proceedings of the American statistical association, section on survey research methods* (pp. 602–607).

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/S0218488502001648

Zandbergen, P. A. (2014). Ensuring confidentiality of geocoded health data: Assessing geographic masking strategies for individual-level data. *Advances in Medicine*, *2014*, 1–14. https://doi.org/10.1155/2014/567049