Multidimensional Point Transform for Public Health Practice

P. AbdelMalik; M. N. Kamel Boulos
Faculty of Health and Education, University of Plymouth, Plymouth, UK

Keywords
Public health, geographic location, privacy, spatial transform, anonymisation

Summary

Background: With increases in spatial information and enabling technologies, location-privacy concerns have been on the rise. A commonly proposed solution in public health involves random perturbation, however consideration for individual dimensions (attributes) has been weak.

Objectives: The current study proposes a multidimensional point transform (MPT) that integrates the spatial dimension with other dimensions of interest to comprehensively anonymise data.

Methods: The MPT relies on the availability of a base population, a subset patient dataset, and shared dimensions of interest. Perturbation distance and anonymity thresholds are defined, as are allowable dimensional perturbations. A preliminary implementation is presented using sex, age and location as the three dimensions of interest, with a maximum perturbation distance of 1 kilometre and an anonymity threshold of 20%. A synthesised New York county population is used for testing with 1000 iterations for each of 25, 50, 100, 200 and 400 patient dataset sizes.

Results: The MPT consistently yielded a mean perturbation distance of 46 metres with no sex or age perturbation required. Displacement of the spatial mean decreased with patient dataset size and averaged 5.6 metres overall.

Conclusions: The MPT presents a flexible, customisable and adaptive algorithm for perturbing datasets for public health, allowing tweaking and optimisation of the trade-offs for different datasets and purposes. It is not, however, a substitute for secure and ethical conduct, and a public health framework for the appropriate disclosure, use and dissemination of data containing personal identifiable information is required. The MPT presents an important component of such a framework.

1. Background

Privacy, as related to identifiable health information, has been a subject of contention within public health and health research: the literature is littered with comments and complaints [1–3], surveys have sought to assess the perceptions of public health professionals and the general public [4–8], and both privacy advocates and public health professionals appeal to a vaguely painted patchwork of legislation [1, 9]. The issue does not generally lie with direct identifiers such as name or an identifying number, but rather with attributes that can be used in combination to re-identify individuals. These are referred to as key attributes [10] or quasi-identifiers [11]. For example, age and sex are commonly used public health quasi-identifiers that have been characterised as having “high utility to an intruder” attempting to re-identify individuals from a dataset [12].

As public health methods advance with ever-evolving technology to better capture the entire context within which health events occur, requirements for privacy-protective methods also increase. In an attempt to address the privacy issue, algorithms for anonymisation and privacy enhancing techniques (PETs) have been proposed and implemented [13], and calls for public health professionals to challenge policies and lobby legislators have been made [2]. Anonymisation algorithms are often measured as a function of indistinguishable records and re-identification probability. The term \( k \)-anonymisation refers to the concept where every record becomes indistinguishable from \( k - 1 \) other records [14, 15]. While no standards for acceptable anonymity thresholds have been established for public health, \( k \) values of 5 and 20 (representing re-identification probabilities of 20% and 5% respectively) have been suggested and used in the literature [12, 16].

An area that has seen a dramatic increase in concern is location privacy, particularly given the importance of spatial information in public health [17, 18]. With the ubiquitous use of Global Positioning Systems (GPS), online mapping applications that provide high-resolution aerial images, and the increasing use of spatial intelligence in public health, location privacy is becoming increasingly contentious – perhaps more so than with other information technologies [9, 18]. Over a decade ago, Armstrong et al published a paper on various mathematical transformations to mask original point location [19]. Of the methods described, random perturbation was found to perform best overall as measured by retention of pair-wise relations, event-geography relations, clusters, trends and anisotropies [19]. Other public health studies have continued to build on this type of spatial transform [20–23] and a good overview can be found in [18].

Methods Inf Med 1/2012
In a classical random perturbation, a circle of radius \( r \) is drawn around the point to be masked such that sufficient population is captured to render the point anonymous, and the point is randomly displaced within the area. This is repeated for each point, resulting in a series of points that are difficult to trace to their original locations due to the stochastic nature of the transform. Not all random perturbations are created equal, however, and advances in their development and implementation have been slow. Ideally, the displacement as measured by the perturbation distance should be minimised, and generally, the more densely populated the area in which a point (case) falls, the less it has to be spatially perturbed to meet a desired anonymity threshold. Adjustments to random perturbation therefore create dynamic radii dependent for each point on its underlying population [21]. This “context-sensitive approach” can be further improved by stratifying on other attributes, such as age and sex, to give a more accurate displacement that minimises information loss [21].

More elaborate revisions of random perturbation have been developed in recent years, including the use of Linear Programming (LP) [22] and a “donut” method of geomasking [23]. However, all of the proposed versions of these transforms modify location-based information almost as an afterthought or secondary anonymisation technique, either assuming that all other identifying information – including important quasi-identifiers such as age and sex – has already been anonymised or stripped, or adjusting the transform accordingly using generalised weighting schemes. Instead, what is needed is a transform that operates discreetly on multiple attributes, in concert with location as part of the overall anonymisation algorithm.

2. Objectives

The current methodology refines the random perturbation approach by combining new and previously studied methods to propose a flexible, dynamic and customisable multidimensional point transform (MPT) acting on attribute data. In this context, attributes of interest – such as location, age, sex, education, etc. – are referred to as dimensions since they define the scope of the transform. Like previous context-sensitive studies [20, 21, 23], the approach presented is an adaptive geomask. However, unlike others, it allows these other dimensions to be incorporated into the anonymisation algorithm directly based on custodian and user tolerances and requirements.

3. Methods

3.1 Algorithm: Overview

The proposed algorithm is dependent on the availability of a base population (real or synthesised) matrix, \( A \), of \( N \) individual records with \( Q \) attributes. The dimensions of interest must be elements of the attribute set, and given the spatial nature of the transform, must include a location attribute – ideally the geographic coordinates of the individual’s relevant address. Given a list of patients, \( B \), from this base population \( A \), the goal of the algorithm is to randomly “move” each patient in \( B \) within a maximum perturbation distance \( \Delta \), while controlling on all dimensions of interest for a defined anonymity threshold, \( k \). “Move” in this case means selecting an alternate record from \( A \) to represent the patient; in this way, the locations are realistic and non-random, but the selection is random.

Consider the example where the dimensions of interest are location, age and sex (other dimensions can be added, provided they are elements of both datasets). The algorithm ensures that the anonymity threshold \( k \) is maintained based on these dimensions and sequentially perturbs them as required based on pre-defined conditions and perturbation tolerances. Location perturbation is measured as the distance moved from the original point, and its maximum tolerance is defined by \( \Delta_L \); the age perturbation tolerance allows the dimension to be categorised, for example in 1-year increments, up to a maximum number of categories; and the sex perturbation tolerance is binary, either requiring a perfect match on gender or not.

The acceptable anonymity threshold is defined by \( k \). For example, \( k = 5 \) means that a given patient is indistinguishable from at least \( k - 1 = 4 \) other individuals within the selection area, which in turn translates to a 20% chance of correctly identifying the patient. The maximum perturbation distance \( \Delta \) is the maximum acceptable threshold for spatial displacement. This does not mean that all eligible records for displacement will be up to \( \Delta \) away from the original point, only that this is as far as the algorithm is allowed to go to achieve the desired \( k \). The actual maximum perturbation radius, \( R \), will depend on the data and defined \( k \).

Given the patient dataset \( B \) with \( j = 1 \) to \( n \) patients, all patients in \( B \) are removed from \( A \) to give the complement non-patient base population, \( C \). Removing the patient dataset individuals from the base population records at the onset of the algorithm has two key effects: it prevents selection of one patient in place of another, and it reduces re-identification risk by forcing \( k - 1 \) to consist entirely of non-patients. Next, for each record in \( B \), all records in \( C \) matching \( B_j \) on sex and age are isolated and the distance between each one and \( B_j \) is calculated. If fewer than \( k - 1 \) matching records are found within \( \Delta \), then the sex and age dimensions are perturbed (i.e. grouped or categorised), based on the pre-defined conditions and in parallel both in the case dataset and the population dataset, and the matching is redone. This is repeated until at least \( k - 1 \) matching records are found. If the algorithm is unable to reach the desired \( k \)-anonymity, then the record is non-transformable within the current requirements, is flagged as such, and the algorithm proceeds to the next record. Otherwise, the algorithm continues.

Of the matching records, the closest \( k - 1 \) records are identified, and a small random distance, \( \delta_r \), is added to the farthest \( k - 1 \) match distance, \( \delta_{\text{max}} \), defining the perturbation radius \( R \). The addition of this random distance ensures inclusion of the point used to set the farthest match distance in case any rounding occurs and guarantees a minimum \( k \)-anonymity. It also adds a small stochastic aspect that complicates re-identification of the original patient location, as not only is the selection of the transformed point different with each run, but so also is the underlying pool from which the point is selected. A record within \( R \) of \( B_j \) is then randomly se-
Fig. 1 Multidimensional Point Transform (MPT) flow
lected from C, and its location assigned as the perturbed point. This is repeated for the next record until all patient records in B have been transformed or flagged. The algorithm flow and components are illustrated in Figure 1.

3.2 Data

Synthesised population data for New York County were acquired from the MIDAS project [24] by request. The dataset contained synthesised records at the individual level, with the dimensions of interest being age, sex and residential location (latitude and longitude in decimal degrees). For each record, latitude and longitude were converted from decimal degrees to radians prior to algorithm execution for use in extent and distance calculations.

New York county was specifically chosen as the study area to allow for comparison with existing published methods – namely the results of the LP approach [22] – on distances required to achieve specified \( k \)-values when additional dimensions are taken into consideration. The two approaches are also similar in that they both seek to minimise perturbation distance and both rely on the presence of underlying spatially-referenced population data.

3.3 Algorithm: Preliminary Proof-of-Concept Implementation

Preliminary testing of the algorithm was completed using one thousand iterations...
for each of 25, 50, 100, 200 and 400 patient dataset simulations, generated by randomly selecting records from the synthesised New York County population.

The controlled dimensions were sex, age and distance. The anonymity threshold $k$ was set to 5 and the maximum perturbation distance $\Delta$ to 1 kilometre. An exact match to sex was required (i.e. no perturbation allowed), and 5 levels of age categorisation were permitted (including exact age). Age categories were created by increasing the age range by one year for each successive level: for the first level, age range is 0, so it is the exact age; for level 2, the range is 1 to give age categories 0–1, 2–3, 4–5, etc. A simplified illustration of the implementation of the age categorisation is given in Figure 2. Note that level 5 matches the age categories used in the American Community Survey for age and sex stratified population counts [25].

Extent-limiting steps were also added to the algorithm to improve computational performance. At the beginning of each iteration and after creation of the patient dataset, one kilometre was added to the maximum and minimum latitudes and longitudes of the patient dataset and used to constrain the extent of the base population. This method was also used when determining the eligible population for each record.

The small random distance added to create $R$ was restricted to a range of 1 and 10 metres to minimally impact geographic displacement. Distance was measured using the great circle formula.

Cumulative descriptive statistics (mean and standard deviations, as well as the minimum, maximum and median) were calculated for successive iterations to assess the effects of the transform on the perturbation distance. Analysis of the age dimension sought to identify the proportion of records requiring categorisation on age to achieve the required minimum $k$. The effect of adding the small random distance $\delta$, on $k$ was also described through descriptive statistics, as was the final perturbation radius. The displacement of the spatial mean of each patient dataset was also calculated in terms of perturbation distance.

The algorithm was coded and run in SAS v9.1; the results were also analysed in SAS v9.1 and graphed using Microsoft Of-

### Table 1 Results of the Multidimensional Point Transform (MPT) algorithm with different patient dataset sizes for New York County

<table>
<thead>
<tr>
<th>SETTINGS</th>
<th>PATIENT DATASET SIZE (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25</td>
</tr>
<tr>
<td>Total records</td>
<td>25,000</td>
</tr>
<tr>
<td>k-anonymity setting</td>
<td>5</td>
</tr>
<tr>
<td>Maximum $\Delta$ setting (kms)</td>
<td>1</td>
</tr>
<tr>
<td>Unique Individuals</td>
<td>24,569</td>
</tr>
</tbody>
</table>

### MEASURES

#### Perturbation Distance (kms)

<table>
<thead>
<tr>
<th></th>
<th>Location Only</th>
<th>Age-Perturbed records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.046</td>
<td>0.046</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.035</td>
<td>0.034</td>
</tr>
<tr>
<td>Minimum</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Median</td>
<td>0.039</td>
<td>0.039</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.975</td>
<td>0.979</td>
</tr>
</tbody>
</table>

#### Perturbation Radius (kms)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Median</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.070</td>
<td>0.042</td>
<td>0.004</td>
<td>0.060</td>
<td>0.994</td>
</tr>
<tr>
<td>Location Only</td>
<td>0.070</td>
<td>0.070</td>
<td>0.070</td>
<td>0.070</td>
<td>0.070</td>
</tr>
<tr>
<td>Age-Perturbed records</td>
<td>0.042</td>
<td>0.042</td>
<td>0.003</td>
<td>0.006</td>
<td>0.997</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.004</td>
<td>0.003</td>
<td>0.002</td>
<td>0.061</td>
<td>0.983</td>
</tr>
<tr>
<td>Median</td>
<td>0.039</td>
<td>0.039</td>
<td>0.039</td>
<td>0.039</td>
<td>0.980</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.975</td>
<td>0.979</td>
<td>0.914</td>
<td>0.980</td>
<td>0.992</td>
</tr>
</tbody>
</table>

#### Actual $k$-anonymity level

<table>
<thead>
<tr>
<th></th>
<th>Mean (rounded down)</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Median</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>17</td>
</tr>
<tr>
<td>Location Only</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Age-Perturbed records</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Minimum</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Median</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Maximum</td>
<td>17</td>
<td>15</td>
<td>66</td>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>

#### Spatial Mean Displacement (kms)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Median</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.010</td>
<td>0.004</td>
<td>0.0004</td>
<td>0.009</td>
<td>0.057</td>
</tr>
<tr>
<td>Location Only</td>
<td>0.007</td>
<td>0.003</td>
<td>0.0002</td>
<td>0.007</td>
<td>0.027</td>
</tr>
<tr>
<td>Age-Perturbed records</td>
<td>0.005</td>
<td>0.003</td>
<td>0.0001</td>
<td>0.005</td>
<td>0.016</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.003</td>
<td>0.002</td>
<td>0.0002</td>
<td>0.003</td>
<td>0.011</td>
</tr>
<tr>
<td>Median</td>
<td>0.002</td>
<td>0.001</td>
<td>0.0001</td>
<td>0.003</td>
<td>0.007</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.003</td>
<td>0.001</td>
<td>0.0001</td>
<td>0.003</td>
<td>0.007</td>
</tr>
</tbody>
</table>
The code was run in RAM using RAMDisk software [26], as preliminary tests showed this to be approximately ten times faster than using a SATA 7200 RPM hard drive.

4. Results

In total, 775,000 records were randomly chosen from the synthesised New York population of 1,482,104 unique individuals and tested with this algorithm, representing 601,790 unique individuals or 41% of the New York county population (Table 1).

Time taken to complete the algorithm ranged from about 5 minutes per iteration for the 25-patient dataset size, to just under two hours per iteration for the 400-patient dataset size.

The age dimension was seldom transformed, as summarised in Table 1. Only one record required an age-transform in each of the 50, 100 and 400-patient datasets, representing 0.0005% of the tested unique individuals.

The mean and median perturbation distances (46 metres and 39 metres respectively), as well as the mean and median perturbation radii (70 metres and 60 to 61 metres respectively), were consistent irrespective of the patient dataset size (Table 1). Cumulative means of the distance between the original and the transformed points are presented in Figure 3 for successive iterations, showing a plateau within 1 metre after less than 200 iterations.

The actual k-anonymity achieved across all runs averaged five individuals, matching on sex and age within the defined perturbation radius as prescribed by the pre-defined k requirement. Ignoring age and sex, the average number of individuals within the perturbation radius from which the random selection was made was just over 800 individuals.

The overall spatial mean of the transformed points was within 5.6 metres of the original spatial mean across all runs, and was inversely related to the patient dataset size (Table 1).

5. Discussion

This study describes a multidimensional point transform (MPT) for anonymising data for public health use that includes location perturbation as a core component of the overall anonymisation algorithm. By perturbing the location dimension in concert with other user-defined dimensions,
the MPT offers a more comprehensive and valid anonymised dataset than existing proposed random perturbation transforms.

Preliminary testing of the algorithm was completed on a synthesised New York County population using three dimensions: sex, age, and location. Since the sex perturbation tolerance was 0, there was no loss of information on the sex dimension; the age dimension was unaltered for 99.9995% of all runs; and the mean spatial displacement of transformed records was 46 metres from original patient points, irrespective of patient dataset size. Fifty percent of points being displaced across all runs were moved less than 40 metres and all transformed records were spatially accurate, representing actual household locations within the population.

Since New York County has a very high population density, it is not surprising that a \( k = 5 \) anonymity level was achieved within such a small distance and with no impact on the age dimension. To further illustrate the multi-dimensional aspect of the algorithm, ad hoc analyses were completed using 100 iterations of sample sizes of 25 cases with \( k = 20 \) and \( \Delta = 50 \) m. With these tightened constraints, the MPT performed precisely as expected: 19 of the total 2500 records were transformed with no age-categorisation required; 148 were transformed with a level 2 age categorisation (i.e. a 1-year interval; recall level 1 is the exact age with no categorisation); 274 were transformed with a level 3 age categorisation; 292 were transformed with a level 4 age categorisation; and 290 were transformed with a level 5 age categorisation. The total successfully transformed records of the 2,500 was thus 1,023 (41%) with a mean perturbation distance of 31 metres. The remaining 1,477 records could not be transformed within 50 metres of the original point at a \( k = 20 \) level, and all 1,477 reached level 5 age-categorisation as expected.

Previously described algorithms treat other record attributes separately from location, sometimes with a weighting to account for the generalised underlying population demographics. Armstrong et al assume that all other potentially identifying attributes such as health information, age, sex, and so on are sufficiently “non-individual specific” [19]. Kwan et al [20] acknowledge the arbitrary nature of their weight factors, and that their results are specific to the unique and particular combination of their underlying population and the case-data. Similarly, Cassa et al. [21] used generalised age-based population density weights – which they refer to as “multipliers” – at the census block group level to implement a probabilistic Gaussian-skewed random perturbation transform. \( K \)-anonymity calculations were also probabilistically based, without taking into consideration other individual dimensions (e.g. age and sex). One could reasonably argue that, given the general importance of and requirement for these dimensions in public health practice, they should be included in the implementation and assessment of anonymisation algorithms.

As with the LP approach of Wieland et al [22], the MPT seeks to minimise perturbation distance and relies on an existing spatial population to do so. However, by accomplishing this in the context of multiple dimensions, it reveals a significant impact on the implications of the transform. Wieland et al. [22] suggest that aggregating to zip code in New York County yields a \( k \) value of about 884 and a corresponding perturbation distance of 519 metres; their LP method perturbation distance using census blocks is only 3.3 metres for the same \( k \). A similar \( k \)-value acting on non-grouped individual points using the MPT requires a mean perturbation of 46 metres. However the actual \( k \) when taking the other dimensions into consideration drops dramatically to only five.

More recently, Hampton et al. [23] added a minimum perturbation distance to existing context-sensitive approaches [21] to implement a “donut” effect. The authors argue that allowing a case to be perturbed to its original location presents a re-identification risk, since an intruder will know that a few individuals may still be correctly identified, and also prevented cases from being perturbed outside their original administrative boundaries. The authors also suggest that their approach is adaptive because it adjusts for the underlying population density and also for minimum and maximum \( k \)-anonymity, whereas other random perturbation methods are only “semiaadaptive” because they fail to be bound by minimum anonymity constraints. However, the donut method as described is also semiaadaptive because it fails to consider the details of the population demographics, such as age and sex. Given the aggregate use of underlying population density, the authors’ suggestion for the incorporation of these dimensions would have to rely on weighting mechanisms similar to those previously described [20, 21]. The donut-algorithm also does not address the possibility of randomly generating a point in a residentially-improbable or impossible location, such as a river or park, and is subject to a re-identification risk associated with multiple iterations [27, 28]. In contrast, the MPT can easily exclude original patient locations, incorporate a “donut-like” effect if desired, and retain points within defined geographies.

5.1 Re-Identification Risk

An issue with random perturbation algorithms is that repeated iterations on the same dataset increase the likelihood of re-identification [27, 28]. The MPT reduces this risk by selecting points from a defined base population instead of generating random points from a uniform distribution within a defined area. This avoids inaccurate or unrealistic placement and can skew the point pattern. The stochastic distance added to create \( R \) in the MPT also adds selection variability with each iteration. Therefore, the spatial mean of repeated iterations will depend on the variable spatial distribution of the underlying population and will not necessarily approximate the original location, unless the population is uniformly distributed around the patient.

A possible weakness is the prevention of selection of one case as a transformation of another. Given two cases of identical age and sex within \( R \) of one another, it could be discovered that each is excluded from the transformed options for the other in favour of a more distant point, allowing potential re-identification of both original points. Removing the preventative selection criterion can resolve this, though it may also allow re-identification since repeated iter-
ations will result in case location selection twice as many times as others. Yet another potential for re-identification exists if the perturbation tolerances and thresholds used are known, though this would require extensive time and computing power. The multidimensional nature of the transform helps complicate re-identification efforts – the more dimensions are permitted to be perturbed, the more difficult re-identification becomes – while exact dimension matching is mitigated by the anonymity threshold.

The MPT as described does not anonymise records relative to themselves as is the general case with current $k$-anonymity techniques. In other words, it anonymises case records to an external dataset – in this case, a population dataset – and not to the case dataset itself. While this does not address potential “prosecutor re-identification” risk [29], the MPT can be configured to anonymise a dataset relative to itself if so desired as noted briefly below under Strengths. This approach, however, was not tested and is not the focus of the current study since in the context of location privacy it only works with generalised or aggregated locations.

It should be noted that meaningful privacy preservation is also a function of prevalence or incidence. For example, given 50 HIV cases in a population of 1,000, random selection of any one individual has a 5% chance of correctly identifying an HIV case. If all information were stripped for the 50 cases except for their location, each individual would be unidentifiable from 999 other individuals (i.e. $k = 1,000$). However, while this has achieved the maximum possible $k$ for this population, it still remains that correctly identifying an individual as having HIV has an effective $k$-value of only 20.

### 5.2 Limitations

The MPT relies on the presence of an underlying base population containing the same dimensions as those required by the data-user, with at least $k - 1$ non-patients for each dimensionally-matched patient. As the number of dimensions increases, available matches decrease, potentially necessitating dimensional compromises which can be controlled by increasing the allowable perturbation of the individual dimensions. The MPT allows exploration of the optimum context-specific combinations for appropriate data release and use.

Some issues were encountered that impacted overall performance, including periodic file locks, competing background applications, power outages, and system resources. Although using RAM allowed faster completion, future implementations may be limited by the amount available for allocation and machine specifications. Other performance-enhancing factors may include use of solid state drives, multithreading and multiple processors, and coding and implementation within an environment other than SAS. Performance will also be a function of the underlying population matrix size.

No amount of masking, de-identification or anonymisation will prevent the misuse of data. Their release must therefore consider other factors such as the user’s trustworthiness, the purpose and scientific or applied merit for which they will be used, implemented security measures and so on.

### 5.3 Strengths

Strengths of the MPT algorithm are its powerful flexibility and customisability, easily allowing criteria to be set on appropriate dimensions relevant to both the study and the target population. For example, in the current implementation, every individual is indiscriminately associated with five age classes. However, some scenarios may require minimum age classes to be set, such that information deemed more sensitive is only released if age is categorised within the appropriate classes. Base population files containing only age classes based on census information can still be used, allowing ages to be classed differently based on the population distribution within the region of interest. The MPT can even be used to transform non-spatial dimensions of the base population for use in future implementations of the transform.

Another advantage of the MPT is its use of a granular base population which can be assigned to increasingly coarser geography. For example, given only postcodes, points can still be approximated using the base population and other dimensions provided in the patient record. Therefore, a 32-year-old female patient in postcode XIXIXI can be assigned to any point within that postcode matching on age and sex within the allowable perturbations, and the MPT applied. This will incorporate a maximum error approximated by the sum of the maximum distance between residential points within the postcode area and the perturbation radius $R$, further confounding potential re-identification.

The MPT does not apply blanket rules to the entire patient dataset; rather, it anonymises each record individually for its own optimum transform. This allows release of the data with the best possible configuration. It also allows for more sophisticated integration of contextual information, facilitating comparison and calculation across datasets. For example, the algorithm could include distance to the nearest school as an added dimension. A maximum spatial displacement of each record from its closest school can be specified in combination with other relevant dimensions, preserving the relative spatial distance to schools. The algorithm is also not independent of the underlying geography. Because it uses pre-established locations for the random selection to meet the required anonymity threshold, knowledge of the existence of non-inhabitable regions or features will not increase re-identification potential (a noted issue with random perturbation techniques [19, 22]). These factors allow the spatial aspect of the transform itself to be bound by multiple, contextually appropriate rules.

If specific dimensions are not known a priori, such as education and income, an areal dimension can be added as part of the control to allow retention of the patient within the specified political or administrative boundary. This allows the flexibility to use as little or as much data as are available to achieve optimum results. The advantage to including additional dimensions beyond administrative or political boundaries is the incorporation of actual
contextual variables as opposed to potentially artificially-related areal units. As mentioned, New York county is extremely population dense making it relatively easy to achieve reasonable anonymity with very little spatial displacement, even when multiple dimensions are considered. As the ad hoc analyses show, however, the MPT allows users and custodians to identify this and modify the parameters in order to achieve acceptable results. In this case, for example, the custodian may agree to lower the \( k \)-value if acceptable or pending certain requirements on the part of the user (e.g. use restrictions, security requirements, etc.). Conversely or simultaneously, the user may accept additional perturbations (i.e. of sex or age) or increased spatial perturbation. The same decisions would have to be made for a sparsely populated rural area; either way, population density does not impair the MPT. By allowing the user and custodian to have control over the various aspects of the transform, including the appropriate or acceptable anonymity threshold, the MPT provides a “user-sufficient mask” [19].

### 5.4 Using Synthesised Populations

Health data are most valuable and informative in their most granular form, and developing a transform that works on individual point-level data at the address level is highly beneficial. However, such a transform would require knowledge of the underlying population – also at the individual point-address level. Although available through population registries, these data are themselves subject to privacy and confidentiality restraints, and are therefore generally not accessible for public health use. Instead, public health practitioners rely on aggregated census data to infer various population demographics. This is where synthesised populations may play a role. For example, a “synthesised, geospatially explicit” US population based on the year 2000 census has been generated to facilitate agent-based infectious disease modelling for the Modeling of Infectious Disease Agents Study (MIDAS) [30]. This population “correctly and appropriately” describes the age and sex demographics by household, and accurately reflects the actual US population. Details on the methodology and population characteristics have been published [30].

Since the MPT makes use of a synthesised population, its validity depends on how well the synthesised data mirror reality on the dimensions of interest. Since the population is based on the year 2000 census, it may inadequately reflect population demographics for earlier or more recent studies. However, given the recurring nature of the census, algorithms used to build the synthesised populations can be re-run to generate new and relevant populations with each census year [30]. A synthesised population may also be invaluable in exploring the relationships between perturbation distance and a variety of quasi-identifiers as illustrated through this study. Their use also allows for the creation of realistic, non-circular disease clusters for investigation – an issue that impacts other studies in this field [23].

Synthesised populations for the US and several other countries have been produced for MIDAS and are available by request. These populations were developed for epidemiological modeling, not for de-identification algorithms, further highlighting their general utility in public health. As such, the development of representative synthetic populations would be highly beneficial. Indeed, development of a synthetic 2010 US population is currently underway by MIDAS scientists, as are tools to allow researchers to generate custom populations based on demographic variables of interest [24].

### 5.5 Algorithm Refinement

Further refinement of the MPT could allow the user to set priority levels for the various dimensions. In the current example, the priority is given to age; age is perturbed only if the anonymisation threshold is not met within the prescribed maximum distance. Instead, the algorithm can be modified to prioritise minimum distance moved within a maximum age perturbation (i.e. the algorithm could begin with the maximum age perturbation to minimise distance and work backwards to achieve the optimum result). This provides maximum flexibility in exploring the optimal transform for a given dataset and context, as minimising changes in one dimension will necessarily impact the effect of other dimensions.

As an example, assume our dimensions of interest are distance, age, sex and race with decreasing priority assignment. In this case, the MPT as illustrated in Figure 1 will first search for \( k – 1 \) exact matches on age, sex and race. In the absence of meeting this requirement, it will generalise race within the defined generalisation threshold and look for \( k – 1 \) exact matches on age, sex and generalised race. Assuming it still fails, it will then generalise sex, and look for \( k – 1 \) exact matches on age and generalised sex and race. And so on. Based on the current design, it will only move on to generalising the next dimension once it has reached the maximum designated generalisation of the previous dimensions with failure to identify \( k – 1 \) matches in the population, since the loop is intra-dimensional. The loop can also easily be changed to allow several dimensional generalisations within an iteration – i.e. across dimensions. In this case, and using the same example, race would be transformed to its first generalisable level, followed by sex if required, then age; assuming failure, it would then loop back and generalise race to its next level, etc. The intent is to minimise loss on those dimensions deemed by the user to be more important to retain closer to their original value, as opposed to finding an overall perceived “optimal” solution, while minimising spatial disturbance (i.e. distance perturbed).

The MPT settings can also be informed by other research in this area. For example, the maximum number of combinations (MaxCombs) of variables of interest is a good predictor of uniqueness [12, 31] and can be used to determine appropriate “geographic area population size” (GAPS) [31] for privacy preservation. This can be used to inform preliminary decisions on setting \( k \) and \( \Delta \) for the MPT; for example, one can begin by setting \( D \) to the approximate mean radius of the census geography most closely corresponding to a calculated GAPS cutoff. MaxCombs can also be used to inform the dimensional categorisation levels, particu-
larly since it is dependent solely on the number of response categories and not the types of the quasi-identifiers.

It has been shown that \( k \)-anonymity can, in some cases, be “over-protective”, particularly for smaller sampling fractions, resulting in unnecessary information loss [29]. The current methodology helps reduce such information loss by incorporating the relevant dimensions directly into the anonymisation algorithm, allowing the user to set permissible categorisation and priority levels, and performing local recoding (i.e. allowing observations to have different and overlapping response intervals [29]).

Appropriate \( k \) values should be a function of the user and the use of the data, as well as the governance structures in place. Some general criteria for setting this threshold have been proposed [12] and should be incorporated into a more comprehensive framework for the disclosure of data.

Preliminary MPT testing was conducted on three dimensions: sex, age and location. Additional dimensions, larger patient data-sets and different base populations with varying population densities should also be explored, as should the effects on common spatial statistics used in public health. Since random perturbation techniques generally increase Type II error probability (e.g. cluster dilution) and do not affect Type I error probability [19], further studies on appropriate thresholds and applications of this algorithm are required in various contexts and with different base populations. Additional analyses quantifying the relationship between the anonymity level achieved and the distance displaced for specific contexts and base populations can also serve as part of a framework for assessing appropriate uses. Currently, privacy legislation applies to “identifiable individuals”. However, with the growing literature around anonymisation, one can now ask the question “at what \( k \)-value does an individual cease to become identifiable under the legislation”? Acceptable anonymity thresholds therefore need to be set and standardised, and the legislation needs to be revised to better reflect this in privacy definitions.

Sophisticated software agents [19, 32] could be used to combine the ingredients required (e.g. the base population from a municipal population registry, the health data from the custodian, and the user requirements) and return an appropriately and optimally transformed dataset (or null result, if no adequate transform is feasible given the data and user specifications). This allows the user to explore analyses that may only become evident after visual exploration of the data’s distribution. A graphical user “front” would be highly beneficial for this purpose, and an image of such an interface is suggested at http://www.personplacetime.org/tools/MPTinterface.jpg.

6. Conclusions

The multidimensional point transform proposed in this study works concertedly on multiple attributes, including the spatial attributes, to give a more complete and appropriate transform that builds location privacy into the anonymisation model from the beginning. Unlike previous studies, this algorithm does not leave other attributes “untouched”, but it does result in a transformed matrix with the same dimensions of the original matrix [19].

The ideal transform preserves the confidential and private nature of individual health records, as well as the geographic integrity of the data, to facilitate public health practice [19]. The optimal approach depends not only on the purpose for the data use and the acceptable disclosure risk [19], but also on the characteristics of the data. Acceptable disclosure risk by the custodian is also a multifaceted consideration based on acceptable anonymity thresholds, trust in the user, adequate security measures, and so on. However, such algorithms cannot substitute for secure and ethical conduct, and a framework for the appropriate disclosure, use and dissemination of data containing personal identifiable information is required [1]. There are also instances in which the release and use of identifiable information in public health are essential [33], and consideration must be made within a developed framework to allow for such cases. The proposed algorithm in this study presents a multidimensional approach that allows one to tweak and optimise the trade-offs for any given dataset and purpose, presenting a necessary component of the much-needed public health framework.

Acknowledgements

Many thanks to Professor Ray Jones at the University of Plymouth, Dr. Khaled El Emam at the CHEO Research Institute, Mr. Philip Ng at the Public Health Agency of Canada, Mr. Sami AbdelMalik, Mrs. Janet Honig and Miss Miriam Rawson for providing support, resources, insight, editing and sounding boards.

Competing Interests

The authors declare that they have no competing interests.

Authors’ Contributions

PA conceived and drafted the transform and manuscript, and conducted the preliminary implementation and analysis. MNKB critically revised the manuscript and contributed to its content and flow. Both authors read and approved the final manuscript.

References


