Measuring The Impact of Spatial Perturbations on The Relationship Between Data Privacy and Validity of Descriptive Statistics

Kelly Broen (broenkelly@gmail.com)
University of Michigan School of Public Health https://orcid.org/0000-0002-2220-4026
Rob Trangucci
University of Michigan
Jon Zelner
University of Michigan School of Public Health

Research

Keywords: epidemiology, Data Privacy, HIPAA

Posted Date: September 18th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-77449/v1

Version of Record: A version of this preprint was published on January 7th, 2021. See the published version at https://doi.org/10.1186/s12942-020-00256-8.

Measuring the Impact of Spatial Perturbations on the Relationship Between Data Privacy and Validity of Descriptive Statistics

Kelly Broen[1,2,4] , Rob Trangucci, & Jon Zelner[1,2]

[1]Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA.

[2]Center for Social Epidemiology and Population Health, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA.

[3]Dept. of Statistics, University of Michigan, Ann Arbor, MI 48109, USA.

[4]Correspondence to: broenk@umich.edu

Abstract

Background: Like many scientific fields, epidemiology is addressing issues of research reproducibility. Spatial epidemiology, which often uses the inherently identifiable variable of participant address, must balance reproducibility with participant privacy. In this study, we assess the impact of several different data perturbation methods on key spatial statistics and patient privacy.

Methods: We analyzed the impact of perturbation on spatial patterns in the full set of address-level mortality data from Lawrence, MA during the period from 1911-1913. The original death locations were perturbed using seven different published approaches to stochastic and deterministic spatial data anonymization. Key spatial descriptive statistics were calculated for each perturbation, including changes in spatial pattern center, Global Moran's I, Local Moran's I, distance to the k-th nearest neighbors, and the L-function (a normalized form of Ripley's K). A spatially adapted form of k-anonymity was used to measure the privacy protection conferred by each method, and the its compliance with HIPAA privacy standards.

Results: Random perturbation at 50 meters, donut masking between 5 and 50 meters, and Voronoi masking maintain the validity of descriptive spatial statistics better than other perturbations. Grid center masking with both 100x100 and 250x250 meter cells led to large changes in descriptive spatial statistics. None of the perturbation methods adhered to the HIPAA standard that all points have a k-anonymity > 10. All other perturbation methods employed had at least 265 points, or over 6%, not adhering to the HIPAA standard.

Conclusions: Using the set of published perturbation methods applied in this analysis, HIPAA-compliant de-identification was not compatible with maintaining key spatial patterns as measured by our chosen summary statistics. Further research should investigate alternate methods to balancing tradeoffs between spatial data privacy and preservation of key patterns in public health data that are of scientific and medical importance.

Introduction

Researchers in public health, medicine, and the social sciences are facing a reproducibility crisis that continues to grow with the complexity of data collection, cleaning, and analysis pipelines. A reproducible study has been defined broadly as one from which a researcher can duplicate results using the data from the original analysis and the methods described in the study (1). In practice, meeting this standard can prove to be quite difficult. These issues are magnified in public health and medicine, where ethical and legal protections of patient and research subject privacy must be considered ahead of the public health and scientific benefits of reproducibility. These issues are particularly acute for spatially referenced disease and health data which may reveal not only the identity but the spatial location of individuals with sensitive health conditions, e.g. HIV infection, or behavioral risks such as injection drug use (2). These roadblocks to a consistently reproducible spatial epidemiology have limited the application of powerful spatiotemporal analytic tools in public health practice. This represents a significant loss to public health, as such data can provide insights into how to best intervene on a wide range of health conditions, ranging from those associated with exposure to environmental toxicants, spatially concentrated social inequality, and infectious disease transmission (3-6).

For example, as recent work in the area of vaccine-preventable diseases has shown, the scale at which such data are reported can determine the nature and the quality of inferences that can be drawn from them (7). In recent months, the COVID-19 pandemic has shown the crucial role of understanding the determinants of fine-scale spatial variation in infection outcomes, as such data are key for understanding differential risks of mortality by age, socioeconomic status and as a function of neighborhood environments. This has created an unprecedented amount of interest in making individual level case data publicly available, with multiple sources producing

maps of case and testing rates (8-11). As analysts produce maps for public release in the rapidly changing pandemic setting, maintaining individuals' privacy is increasingly essential as stigma-driven harassment also increases (12, 13). While all maps are using aggregated counts, the level to which data has been aggregated varies; some maps are providing data at as low a level as the zip code level while many only release information by county (8-11). More granular maps have suppressed data for zip codes with limited numbers of cases, but there are no standardized limits for data release (10).

A number of geomasking methods have been proposed to address the problem of identifiability in publicly released spatial health data. Geomasking algorithms shift the coordinates of a point of interest in a way that is intended to reduce the likelihood of identification of all individuals in the dataset to the point that it no longer presents a meaningful risk of identification. However, there has been relatively little attention paid to the amount of spatial information lost relative to privacy protection gained from each of these approaches. In this paper, we measured the tradeoff between increased privacy and spatial information loss provided by a wide variety of geomasking approaches applied to the same detailed dataset. We used an array of geographic perturbation methods described in the literature on spatial analysis and medical geography, which are commonly employed in the public release of sensitive spatial data, as well as a widely-used metric of anonymization known as *k-anonymity*(14).

A better understanding of the nature and extent of these tradeoffs is necessary to allow researchers, regulatory bodies such as IRBs, and data providers such as public health departments and hospitals, agree on spatial perturbation methods that can preserve patient or participant privacy, while understanding how they may result in potential biases that could limit the utility of such data for different types of analyses.

The acceptable ratio of information lost to privacy gain is likely to vary as a function of 1) the sensitivity of the underlying data, 2) the nature of the data sharing, e.g. with a trusted partner subject to a data use agreement vs. wide public release, and 3) the public health urgency of the problem the data may aid in solving. These questions have always been pertinent, but the COVID-19 pandemic has forced them towards the front of the conversation.

Privacy-first reproducibility

A commonly discussed standard for reproducibility in public health and medicine is that published analyses should include access to all underlying data, the exact methods employed from data processing to analysis and figure generation (including the code to run all analyses), and documentation sufficient to run the provided code on the provided data and obtain the published results (15). Finally, all of these components should be distributed in a way that makes them widely accessible (e.g. under a permissive software license, hosted on an open and visible platform such as *github*) (15). Done properly, this allows others to directly validate results, rapidly deploy new methods (16) and pursue alternate hypotheses using the original data. However, this maximally transparent approach is ethically and legally prohibited when the relevant data contain identifiable information including home addresses and key patient demographics. These are considered protected health information (PHI) under the Health Insurance Portability and Accountability Act (HIPAA) and therefore cannot be publicly released in an unmasked form (17). In this paper, we argue for and outline the contours of a *privacy first* approach to reproducibility that balances these ethical and legal obligations to individuals with potential benefits to public health. Although the HIPAA statute does not lay out specific standards for what constitutes an unacceptable level of identifiability, a common interpretation of

HIPAA requirements on data release is that each data point must be indistinguishable from at least 10 others in the same dataset (18).

Under HIPAA, data may be publicly released after all identifiable information is removed, with identifiability defined by 18 specific attributes (17). The unit of interest in geospatial epidemiology - an individual's location or set of locations visited over time - is clearly sensitive, identifiable information, and therefore methods for deidentification of spatial data must be robust to malicious reverse engineering. Despite the importance of these methods for completing privacy-respecting reproducible research, little is known about how to leverage different methods of spatial perturbation to accomplish the twin goals of 1) maximizing participant privacy (i.e. minimizing identifiability) while 2) maintaining key spatial patterns necessary for reproducibility and verification of published results (19). Because of this lack of guidance on how to best de-identify individual-level spatial health data to maintain HIPAA compliance, spatial epidemiologists and other health researchers face significant barriers to reproducibility. HIPAA outlines two approaches by which de-identification can be considered to have been achieved:

1) **Safe Harbor:** This method requires the removal of all identifiers. Only the first three digits of zip codes are maintained if "the geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people". If the geographic unit contains 20,000 or fewer people, all five digits of the ZIP code are removed.(20)
2) **Expert Determination:** Under this approach, "a person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable" implements a scientifically

> verified method on spatially identifiable data until there is "very small risk that [the] intended recipient could identify [the] individual." (17). Although HIPAA does not explicitly quantify this risk, it is commonly interpreted as each individual being indistinguishable from at least 9 other individuals in the dataset (18).

Despite efforts to develop geomasking methods that can address these issues, there is no consensus on the relative benefits of each approach. Instead, previous work in this area has tested only one or a small number of perturbation approaches on a specific dataset (21-23), making comparison to other perturbation methods infeasible. The primary measure of privacy employed in these studies is k-anonymity, though the implementation of this metric across studies has been inconsistent (21-23).

Data We geocoded the household location of each recorded death in Lawrence, Massachusetts, from 1911-1913. We chose to use a historical dataset because if all individuals in the dataset have been deceased for over 50 years, HIPAA allows for complete release of information considered private for living individuals (17). Using ArcGIS Version 10.6.1, multiple sheets mapping different sections of Lawrence, Massachusetts were configured to create a complete map of the city limits. Individual residences were labeled on the original maps, allowing each address to be located and geocoded. Each point represents a death between 1911 and 1913. Shapefiles of the city of Lawrence, Massachusetts, and the Merrimack River were obtained from Mass.gov (24, 25).

Methods

We employed seven different perturbation methods to assess the impact of different approaches to data masking on both privacy and spatial patterns in the underlying data. We examined both non-aggregating perturbations, which move points to unique locations, and aggregating perturbations, which agglomerate them into a single location:

Non-Aggregating Perturbations:

1) **Random Perturbation:** Each case is moved a randomly selected distance in a randomly selected direction. Case locations are not restricted by the bounds of the study area, but two maximum perturbation distances were employed, restricting points to locations within a 50- or 250-meter radius.
2) **Random Weighted Perturbation:** Same as random perturbation, but the maximum distance for each case is constrained to the distance to the point's k-th nearest neighbor. We implemented random weighted perturbation twice, with points moved within the distance of the 5th and 50th nearest neighbors.
3) **Donut Masking:** Each case is moved in a random direction within a random distance constrained to an interval defining a maximum and minimum distance. Donut masking was implemented twice, with points moved between 5-50 meters and 50-250 meters.
4) **Horizontal Shear:** Cases are perturbed using a linear transformation to shear the data horizontally. We shifted each point along its x axis until it was 45 degrees away from its original position relative to the center of the distribution of points. (23).
5) **Voronoi Masking:** This approach moves each case to a point on the nearest edge of its Voronoi tessellation, or the polygon around the original points where the lines are equidistant to the point and its nearest points (26). Although Voronoi masking does not always move points together, if two points are both each other's nearest neighbor, they

will be snapped together so Voronoi masking does have some degree of aggregating effect.

Aggregating Perturbations: Aggregating perturbations move multiple points to the same centroid of a cell within a user-defined grid, effectively hiding the individual within a larger population (27). We employed two methods of aggregation adapted from Seidl, et al, 2015:

1) **Grid Line Masking:** Points are moved to the nearest edge of their enclosing grid cell.
2) **Grid Center Masking:** Points are moved to the centroid of the cell within which they are located.

To understand how the resolution of the grid employed impacts our outcomes, both of these were performed using a fine-scale grid (100mx100m) and a coarser one (250mx250m),

Spatial Measures

To determine how much and which types of information were preserved by each approach, we compared each perturbed dataset to the original data using multiple spatial statistics:

1) **Point Center:** The center of the spatial distribution is calculated as the mean and median of the point coordinates, comparing each perturbation to the original data. The difference in mean and median from the unperturbed data was calculated as the Euclidean distance between the points. Changes in the center of the spatial distribution demonstrate the overall movement of points resulting from each perturbation.
2) **Global Moran's I:** This is a measure of spatial clustering ranging from -1 (complete separation) to 1 (complete clustering) (28). Points were aggregated to 200x200 meter cells and Global Moran's I was calculated to compare if the number of deaths in a cell is overall similar or dissimilar to the number of deaths in surrounding cells.

3) **Local Moran's I:** This is a measure of local spatial autocorrelation, indicating how similar a spatial unit is to its surrounding neighbors. As with Global Moran's I, values range from [-1, 1] (29). As with Global Moran's I, points were aggregated to 200x200 meter cells.
4) **Distance to Kth-nearest Neighbor:** For each perturbation, the average distance of a death to its 1st, 5th, 10th, and 20th neighbors was calculated and compared to the same distance in the unperturbed data as in (21). As points become more clustered in space, average distance to the kth nearest neighbor decreases. Examining the 1st, 5th, 10th, and 20th neighbor allows us to measure the magnitude of clustering or dispersion conferred by a perturbation.
5) **L-Function:** The last spatial metric computed is the L-function, a normalized form of Ripley's *K*. The L-function calculates the expected number of points within a multi-dimensional ball of radius *r*, divided by the volume of the ball (30).This is used to assess whether the points within a fixed distance of a given location demonstrate clustering or repulsion to an extent greater than would be expected by random chance alone.

Measuring de-identification

We used k-anonymity, which is a metric widely used to measure the degree of privacy conferred by a particular perturbation. Specifically, in a dataset with a k-anonymity of 10, each released record is indistinct from at least 9 (k-1) other records (14). For non-spatial data, this typically requires deleting or randomizing data fields until there are at least *k*-1 indistinct records for each case. In the context of spatial data, k-anonymity refers to the number of perturbed points closer to the unperturbed point than its own perturbation. An individual point's k-anonymity is measured using the number of newly perturbed points that fall within a circle around the point's new, perturbed location, with the radius of that circle equal to the distance the point was moved by the perturbation (22). K-anonymity is typically reported as both the average k across each

point in the dataset, as well as the minimum k. To ensure protection for all subjects, if the minimum k-anonymity for a point in the dataset is < 10, the perturbation is considered not to meet HIPAA de-identification standards. Because the k-anonymity provided by a perturbation is a function of the spatial density of the data, we performed perturbations on both the full dataset as well as down-sampled data, e.g. randomly sampling only 75% of the available points, to understand the impact of the density of the unperturbed data on the degree of anonymity conferred by each approach.

Results

In this section, we will review the impact of each of the different approaches to perturbation outlined above on the spatial characteristics of the perturbed datasets, as well as the degree of anonymization conferred by each approach.

Impact of Perturbation on Key Spatial Statistics

Point Center: The median center of the spatial distribution moved the farthest Euclidean distance with affine shear masking, followed by grid center masking with 100x100 meter cells, and grid center masking with 250x250 meter cells, which moved the mean 123 meters, 42 meters, and 33 meters, respectively. All other perturbations had little effect on the median of the spatial points, moving less than 12 meters in Euclidean distance, and none of the perturbations moved the mean center of the spatial distribution more than 5 meters in Euclidian distance.

Global Moran's I: The unperturbed data has a Global Moran's I of 0.58, indicating positive spatial autocorrelation between the numbers of deaths in each cell. Although all the perturbations

maintained a positive value of Global Moran's I, grid center masking with both 100x100 meters and 250x250 meters cells resulted in greatly decreased values of *I*, from 0.58 to 0.36 and 0.30, respectively. Grid line masking with 250x250 meter cells and random weighted perturbation within the 5th nearest neighbor also decreased the Global Moran's I to 0.52 and 0.55, respectively. All other perturbations increased the *I* value, with donut masking between 50 and 250 meters increasing the value the most to 0.79.

Local Moran's I: Because Global Moran's I is the average of all Local Moran's I, trends were similar. Assessing the number of deaths in 200x200 meter cells, each perturbations' distribution of Local Moran's I was compared to the unperturbed distribution of Local Moran's I. Voronoi masking meters and random perturbation at 50 meters followed the original distribution Local Moran's I most closely while donut masking between 50-250 meters created the most altered distributions. Choropleths of Local Moran's I demonstrate the change in spatial autocorrelation for the number of deaths per 200x200 meters cells.

Distance to kth-Nearest Neighbor: For each perturbation, as *k* increases, the average distance to the k-th nearest neighbor became more similar to the distances for the unperturbed data. Aggregating perturbations decreased the average distance at all values of k, while non-aggregating perturbations increased the distance to the k-th nearest neighbor. Voronoi masking, which has both aggregating and non-aggregating properties because some points are moved together, greatly decreased the average distance to the 1st nearest and neighbor but maintained the average distance to all other neighbors.

L-Function: The L-function was measured for each perturbation and compared to the original data. Voronoi masking had the least effect on the dispersion patterns while affine shear and grid center masking at both 100- and 250-meters cells had the greatest effect.

Impact of perturbation on data privacy

Using the complete dataset, there was no perturbation that met the HIPAA standard of not leaving any points with k-anonymity < 10. For clarity, we denote k-anonymity as ρ and average k-anonymity as $\bar{\rho}$. Affine shearing provided the greatest privacy protection, with only 6.5% of cases (265 cases) with $\rho < 10$ and 3.3% of cases (134 cases) with $\rho < 5$. Grid center masking with 250x250 meter cells resulted in 8.8% of cases (357 cases) with $\rho < 10$ and 3.9% of cases (159 cases) with $\rho < 5$. All other approaches left at least 623 cases (or 15.4% of all cases) with ρ < 10. Voronoi masking conferred the least anonymity, with $\bar{\rho} = 1.90$ and all points having ρ < 10. When using a random sample of 75% of the cases, none of the perturbations met the HIPAA standard of all points having a k-anonymity greater than or equal to 10. As the percent of points released decreases, anonymity for those points continued to decrease, indicating that high spatial density increases individual privacy as measured by k-anonymity.

These results indicate that, to obtain the level of de-identification required by HIPAA standards using the perturbation methods we employed, significant alterations of some key spatial patterns were required. Affine shearing provided the greatest K-anonymity but greatly moved the spatial center of the distribution and altered the Local Moran's I patterns. Grid-center masking with 250-meter cells provides the next greatest K-anonymity, but also significantly alters values of key statistics such as Global Moran's I, Local Moran's I, and Ripley's K/L.

Discussion

Our results show that, using the published perturbation methods applied in this analysis, HIPAA-compliant de-identification was not compatible with maintaining key spatial patterns as measured by our chosen summary statistics. This highlights the challenges of ensuring privacy while releasing datasets that maintain key spatial patterns. Affine shear provided the greatest anonymity using the k-anonymity metric and maintained some spatial patterns, but points could be easily de-identified if the angle of the shearing can be determined. Spatial features, such as the Merrimack River in this dataset, would indicate where the true locations of cases could not be, and reverse-engineering around these empty spots could determine the angle of the shear, which could then easily be undone to obtain the original data. Grid center masking with cells of 250x250 meters produced large changes in Global Moran's I values and dramatically altered the distribution of local clustering indicators (e.g. local Moran's I) but also provided the greatest de-identification as measured by k-anonymity that is not as vulnerable to reverse engineering as easily as affine shearing. However, grid center masking with cells of 250x250 meters still did not meet HIPAA standards for privacy (minimum $\rho \geq 10$ for the entire dataset) with 357 cases with $\rho < 10$

Voronoi masking, random perturbation, and random weighted perturbation had the least impact on underlying spatial patterns, but also provided minimal de-identification with hundreds of points having $\rho < 10$ and a minimum $\rho = 1$. Voronoi masking was either the first or second closest to the original value for all measures of spatial aggregation, indicating that while Voronoi masking may not provide de-identification thorough enough to meet HIPAA standards, it does maintain underlying spatial patterns better than other methods of geomasking. This suggests that efforts to build on Voronoi-based approaches to increase their impact on privacy may be fruitful.

This might include using multiple iterations of the Voronoi tessellation algorithm, known as Lloyd's algorithm, as well as combining another perturbation technique with Voronoi masking (31). Another iterative possibility is to employ a hybrid approach to maximizing k-anonymity, e.g. by applying a stronger perturbation with individual $\rho < 10$ after the first application of an approach that works well for the large majority of points.

Although closer to the HIPAA standard than all other perturbations except affine shear, grid center masking with cells of 250x250 meters strongly impacted all of the spatial measures employed. Because grid center masking is an aggregating perturbation, it necessarily decreased the distance to kth-nearest neighbors as well as Global Moran's I. Although grid center masking with such large cells may not provide high fidelity for spatial statistics at the fine scale examined here, the deterministic nature of the of perturbation results in predictable biases of the underlying statistics. A further analysis of these relationships may be helpful for estimating correction factors that can be used to adjust estimates derived from perturbed data so that they are closer to those derived from the underlying data.

Our analysis has a number of strengths. Unlike previous research, the anonymity metric used to measure de-identification was specifically derived from the HIPAA standard. This provides a realistic measure of the likelihood that a given approach will produce data that accord with U.S. health privacy laws. Additionally, our direct comparisons of a variety of perturbation measures using a single policy-relevant anonymization metric may aid in the development of a consensus around how and when these different approaches should be applied.

Despite these strengths, these results also have several important limitations. For example, they are limited by the use of a single spatial dataset characterized by strong spatial clustering representative of data from a densely populated urban neighborhood or small city. It is

likely that different perturbations will have different implications when the underlying data have different spatial characteristics, e.g. the presence of multiple distinct spatial clusters, lower density of points over a larger spatial area, etc. In addition, the original mortality data demonstrated significant spatial autocorrelation with a statistically significant Global Moran's I of 0.58. Because aggregating perturbations will always move points together and create empty spaces where points previously were, they will always bias Moran's I towards greater dispersion given the true underlying distribution. If the true data were less clustered, aggregating methods of perturbation might produce different biases. An important next step towards developing a set of broadly-applicable best practices for privacy-first reproducibility is performing the analyses presented here on datasets characterized by different densities and spatial scales. Future studies should investigate the effect that differences in the underlying data have on the tradeoff between de-identification and maintenance of spatial patterns.

Despite its broad use as a measure of spatial anonymity, k-anonymity may in fact not be ideal for this purpose. For example, in the context of non-spatial data, ensuring that an individual cannot be distinguished from *k* other individuals in the same dataset may be reasonable. However, k-anonymity for spatial data is heavily influenced by the point density of the original data: if points are very close together, the k-anonymity conferred by a perturbation may be large even though the actual distance between the original and perturbed locations is very small. The risk posed to privacy becomes clear when the availability of other sources of spatial information is available, e.g. in census data or via projects such as WorldPop. This means that individuals outside the original dataset may be at risk of identification as a result of linkages between spatial data and key publicly available metadata elements (e.g. population density, age distributions, race/ethnicity, sex/gender breakdowns). Consequently, even if a perturbation increases within-

dataset anonymity, it may little to no impact on privacy at the population level if it provides information on risk in the underlying population that can be extracted via approaches such as a kriging and other methods of spatial interpolation and smoothing.

Future studies should investigate alternative approaches to spatial de-identification that address the limitations of within-dataset k-anonymity. In addition, these questions become more complex when additional information beyond the spatial location of a case is included in a dataset, e.g. age, sex, comorbidity status, etc. Resolving these technical, ethical and legal issues will have positive benefits for researchers, patients, and policymakers across the health sciences. The urgency of these questions is clear: as the response to COVID-19 has shown, high-resolution data can be helpful for informing both short-term tactics and long-term strategies in public health response (32). But the benefits of more granular data may not outweigh the downsides if individual privacy cannot be protected. The urgency of the COVID-19 pandemic also underscores the need for a set of well-defined and agreed-upon privacy and technical standards for spatial epidemiology that can be rapidly deployed in an emergency while maintaining high ethical standards and legal compliance. Finally, although this is outside the scope of the present analysis, approaches focused on optimizing both the informational content and privacy protection afforded by perturbation may be fruitful.

Additionally, although we have used HIPAA as a benchmark, the approaches described here have clear relevance to other types of data not necessarily subject to HIPAA protection, but for which ethical and legal barriers to full reproducibility still exist. For example, effective intervention to prevent human trafficking and other forms of exploitation may aided by geospatial data, while the underlying location of reported events is clearly sensitive. Fine-scale demographic data are increasingly used for a broad range of social science applications, but their

availability is limited because of the risk of disclosing individual-level information. Consequently, methods that facilitate openness and reproducibility – while complying with ethical and legal standards set out in HIPAA and other regulations to safeguard privacy – are sorely needed to advance the impact of the spatial sciences across public health, medicine, and the social sciences, while safeguarding the privacy and safety of patients, study participants, and the population as a whole.

Declarations

Ethics Approval

This study was not subject to review by the Institutional Review Board because it used publicly available data.

Availability of Data and Materials

Data and code are available at https://github.com/broenk/Spatial_Perturbation.

Competing Interests

The authors declare that they have no competing interests.

Funding

KB was funded by the Targeted Research Training Program through the University of Michigan Center for Occupational Health & Safety Engineering (COHSE) and a grant from the rOpenSci foundation. JZ was funded by a grant from the rOpenSci foundation.

Authors' Contributions

All authors are responsible for this manuscript and have been involved in the conception and design; analysis and interpretation of the data; or drafting and revising of the manuscript

Acknowledgements

The authors acknowledge the Lawrence History Center for providing historical records and Dr.

Chris Muller for accessing and digitalizing these historical death records. We also acknowledge

Dr. Veronica Berrocal for providing critique and feedback on early drafts.

References

1. Cacioppo JT, Kaplan RM, Krosnick JA, et al. Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science. 2015.
2. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature News* 2016;533(7604):452.
3. Ostfeld RS, Glass GE, Keesing F. Spatial epidemiology: an emerging (or re-emerging) discipline. *Trends in ecology & evolution* 2005;20(6):328-36.
4. Gray SC, Edwards SE, Miranda ML. Race, socioeconomic status, and air pollution exposure in North Carolina. *Environmental research* 2013;126:152-8.
5. Hixson BA, Omer SB, del Rio C, et al. Spatial Clustering of HIV Prevalence in Atlanta, Georgia and Population Characteristics Associated with Case Concentrations. *Journal of Urban Health* 2011;88(1):129-41.
6. Liu H-Y, Skjetne E, Kobernus M. Mobile phone tracking: in support of modelling traffic-related air pollution contribution to individual exposure and its implications for public health impact assessment. *Environmental Health* 2013;12(1):93.
7. Brownwright TK, Dodson ZM, van Panhuis WG. Spatial clustering of measles vaccination coverage among children in sub-Saharan Africa. *BMC Public Health* 2017;17(1):957.
8. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*.
9. Coronavirus Disease 2019 (COVID-19): Cases in the U.S.: Centers for Disease Control and Prevention; 2020. (https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html). (Accessed).

10. Florida's COVID-19 Data and Surveillance Dashboard. Florida Department of Health, Division of Disease Control and Health Protection 2020.

11. Times TNY. Coronavirus in the U.S.: Latest Map and Case Count. The New York Times, 2020.

12. Tavernise S, Oppel Jr. RA. Spit On, Yelled At, Attacked: Chinese-Americans Fear for Their Safety. The New York Times, 2020.

13. Elassar A. Armed vigilantes blocked a neighbor's driveway with a tree to force him into quarantine. CNN; 2020. (https://www.cnn.com/2020/03/29/us/maine-coronavirus-forced-quarantine-trnd/index.html). (Accessed).

14. Sweeney L. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 2002;10(05):557-70.

15. Peng R. The reproducibility crisis in science: A statistical counterattack. *Significance* 2015;12(3):30-2.

16. Wicherts JM, Veldkamp CL, Augusteijn HE, et al. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in psychology* 2016;7:1832.

17. The Health Insurance Portability and Accountability Act of 1996. 1996.

18. Zerbe J. Geospatial data confidentiality guidelines. 2015.

19. Zandbergen PA. Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. *Advances in medicine* 2014;2014.

20. Services UDoHaH. Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. *US Department of Health and Human Services, Washington, DC)*

Available at: https://www hhs gov/hipaa/for-professionals/privacy/special-topics/de-identification/index html Accessed September 2012;26:2018.

21. Seidl DE, Paulus G, Jankowski P, et al. Spatial obfuscation methods for privacy protection of household-level data. *Applied Geography* 2015;63:253-63.
22. Hampton KH, Fitch MK, Allshouse WB, et al. Mapping health data: improved privacy protection with donut method geomasking. *American journal of epidemiology* 2010;172(9):1062-9.
23. Armstrong MP, Rushton G, Zimmerman DL. Geographically masking health data to preserve confidentiality. *Stat Med* 1999;18(5):497-525.
24. MassGIS. Hydrography (1:100,000). Massachusetts: MassGIS (Bureau of Geographic Information), 2019.
25. MassGIS. County Boundaries. Massachusetts: MassGIS (Bureau of Geographic Information), 2019.
26. Voronoi G. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième mémoire. Recherches sur les parallélloèdres primitifs. *Journal für die reine und angewandte Mathematik* 1908;134:198-287.
27. Allshouse WB, Fitch MK, Hampton KH, et al. Geomasking sensitive health data and privacy protection: an evaluation using an E911 database. *Geocarto international* 2010;25(6):443-52.
28. Moran PA. Notes on continuous stochastic phenomena. *Biometrika* 1950;37(1/2):17-23.
29. Anselin L. LOCAL INDICATORS OF SPATIAL ASSOCIATION - LISA. *Geographical Analysis* 1995;27(2):93-115.
30. Dixon PM. Ripley's K Function. *Wiley StatsRef: Statistics Reference Online* 2014.

31. Tang C, Monteleoni C. On Lloyd's algorithm: New theoretical insights for clustering in practice. Presented at Artificial Intelligence and Statistics2016.

32. Raskar R, Schunemann I, Barbar R, et al. Apps gone rogue: Maintaining personal privacy in an epidemic. *arXiv preprint arXiv:200308567* 2020.

Figures

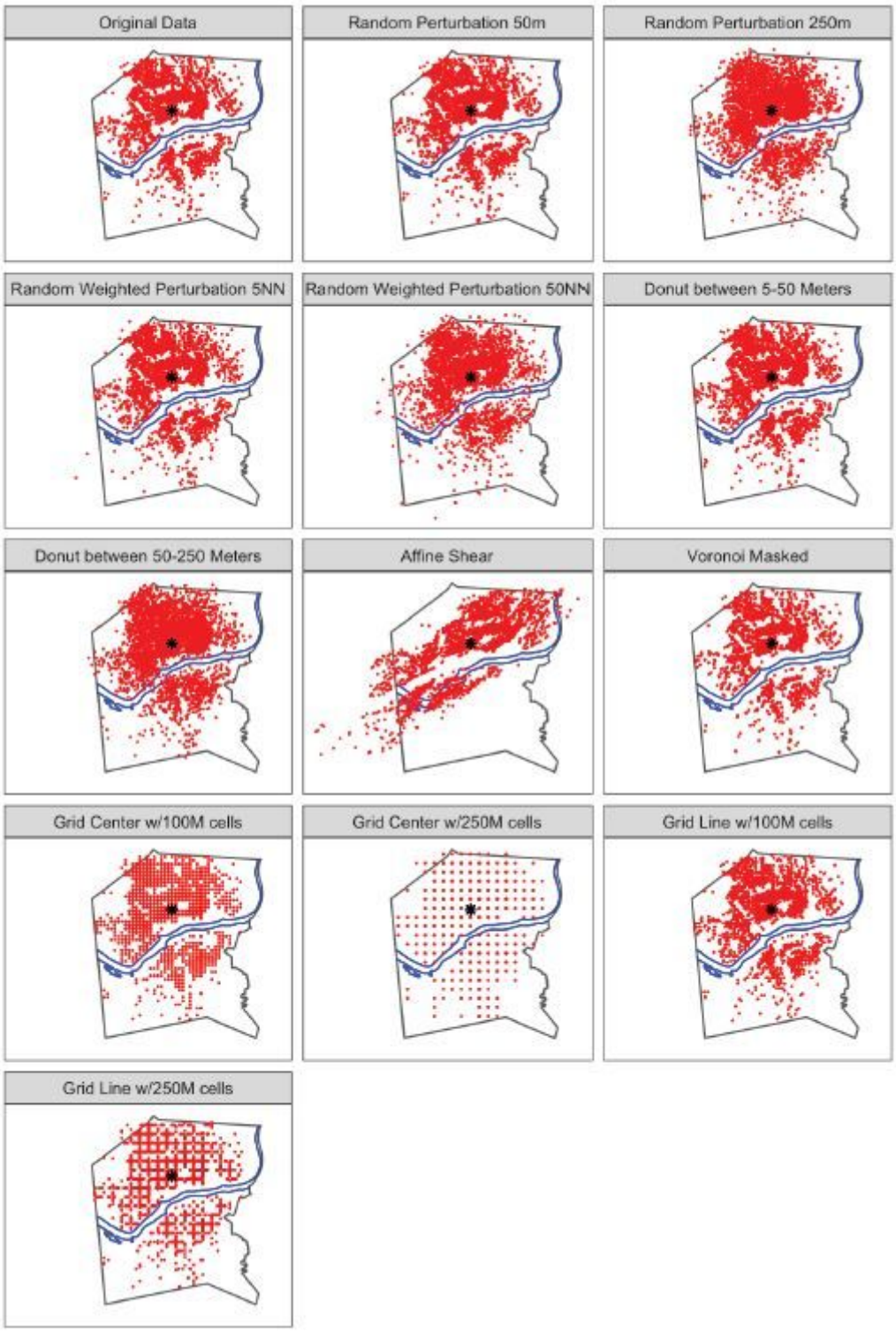

Figure 1

Maps of Each Perturbation Each death is recorded in red, with the spatial center represented by the black center. The blue lines represnt the boundaries of the Merrimack River.

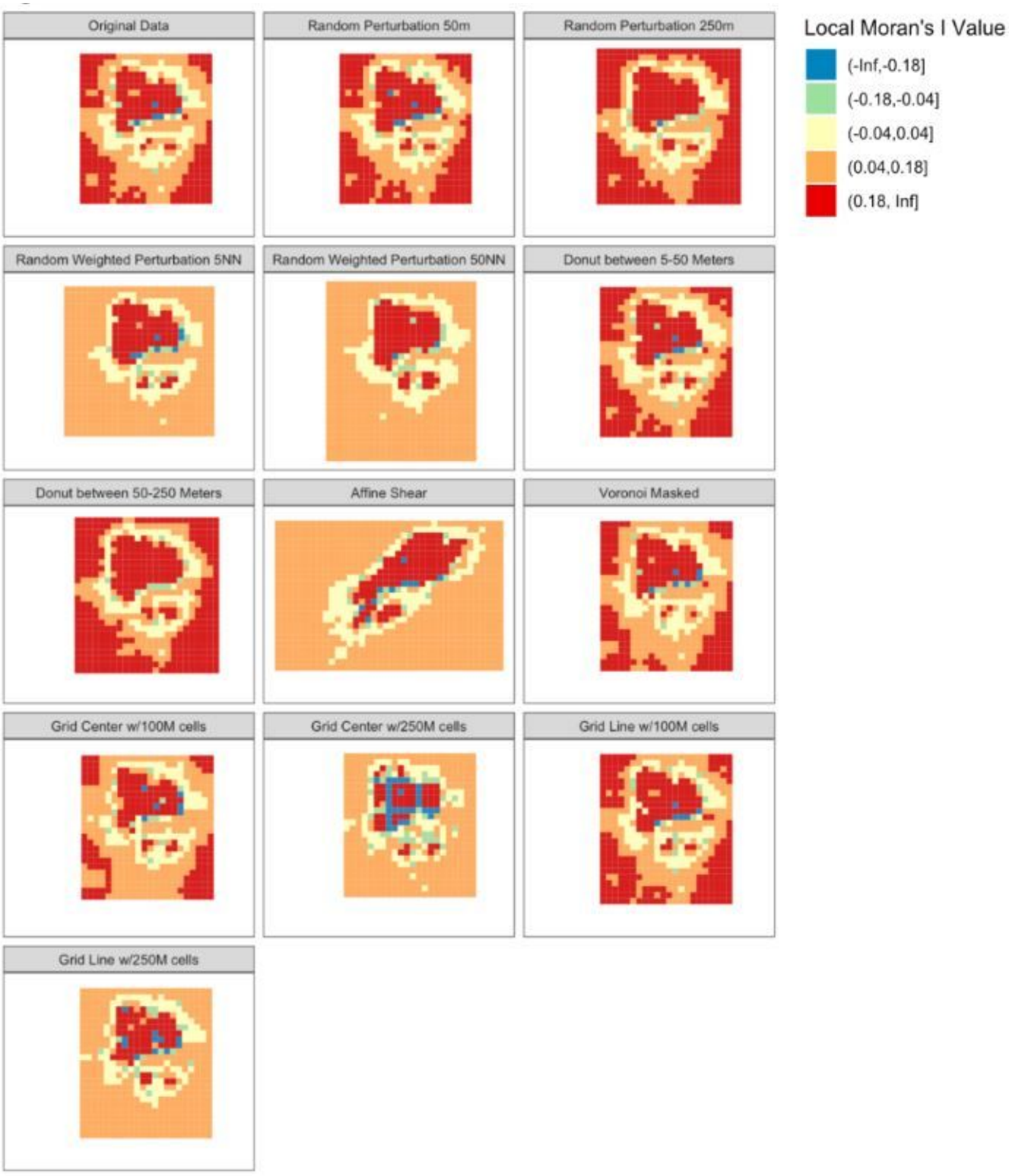

Figure 2

Local Moran's I For each perturbation, Local Moran's I was calculated using 200x200 meter grid celis.

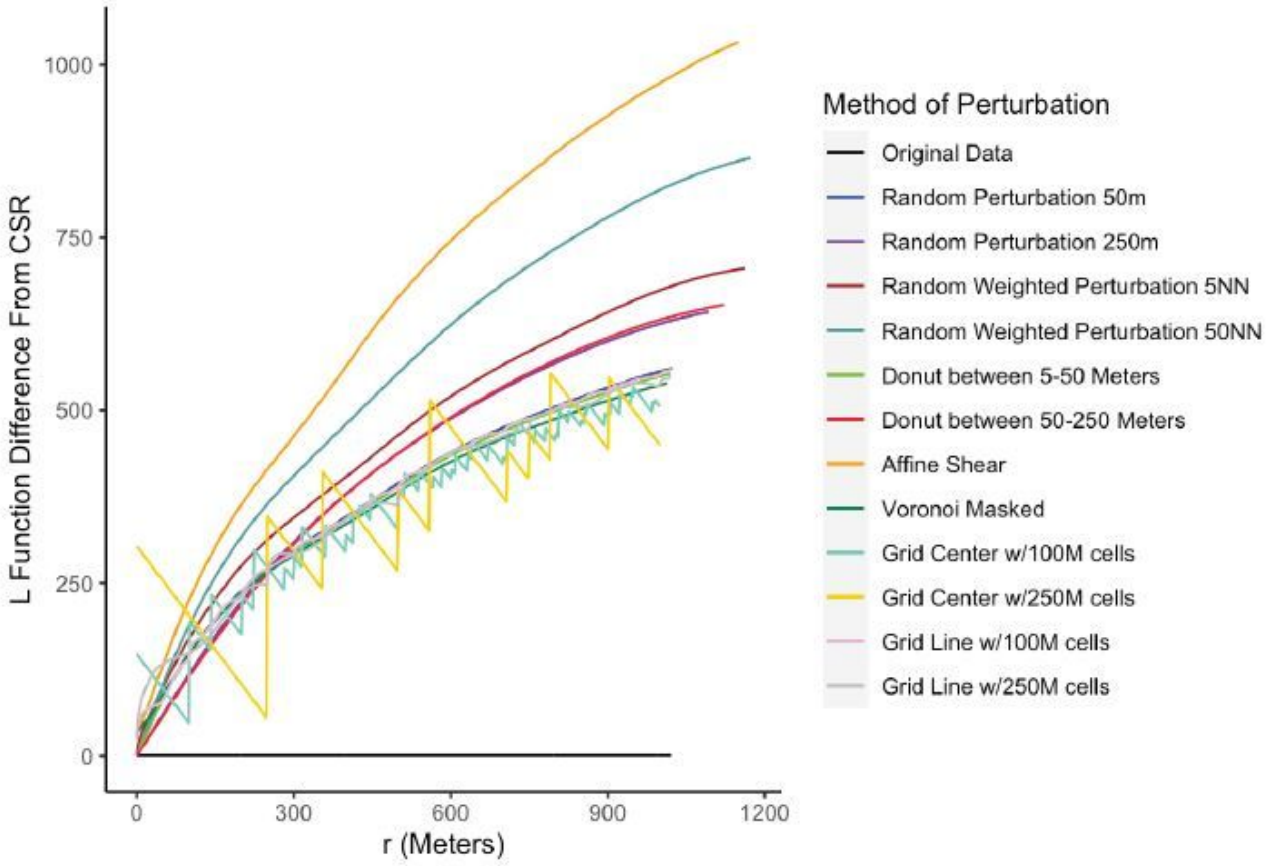

Figure 3

Ripley's L Each line represents the difference from the original data's variation from complete randomness, with the original data centered at 0