Statistical Efficacy of Distance Matrix Anonymization

Coleman Harris, Department of Biostatistics
Vanderbilt School of Medicine

INTRODUCTION & BACKGROUND

It is often of interest to include geographical and spatial representations in epidemiological and medical studies, particularly to determine evidence of spatial clustering of a disease or outbreak. Much work has been done[1,2] to explore the privacy concerns associated with disease mapping, and the anonymization methods to alleviate these concerns.[3,4] Recently, Kroll & Schnell[5] look beyond geographic data into the disclosure of distance matrices that leave data susceptible to a graph theoretic linkage attack.[6]

Linkage attacks involve an adversary who combines common attributes (quasi-identifiers) of two files in order to match records and get more information about an individual. The most famous example of this is when Sweeney connected a Massachusetts voter file[7] with health data to learn why then Governor Bill Weld was admitted to the hospital. To understand the graph theoretic linkage attack, consider that relative distances between observations can be calculated in both the target and identification files. An adversary's identification file may be similar to a voter file, in that it is quite easy to calculate relative distances between individuals that can be compared to a target file of interest.

An adversary will attempt to compare the pairwise relative distances in the target and identification files to determine which pairs of quasi-identifiers are compatible. Essentially, since distances are estimable for both datasets, we use the distance values as edge weights[8] to increase the probability of matching individuals in the two datasets. This problem in the graph theory space is dubbed the maximum clique problem, an NP-hard problem in which approximations are used to solve the distance between vertices and edges of a compatibility graph like the one described above.

Kroll & Schnell's proposed anonymization of the distance matrix implements Lipschitz embedding, which provides useful properties that preserve some statistical measures of the geographical data while maintaining a small disclosure risk. The authors investigated the privacy implications in the following simulation:

1. Distort the distance between 400 geocoded English hospitals via Lipschitz embedding
2. Form compatibility graphs between the datasets
3. Implement maximum clique search algorithm[9]
4. Produce matches between the files

The expected value of the precision was then taken over 20 simulations, resulting in an average precision of roughly 25% (*k*-anonymity terms: $k < 4$). It is further shown that the anonymized data maintains its nearest neighbors classification rate and preserves relative orderings compared to the untransformed data; further statistical robustness is proposed but not implemented.

Hence, the purpose of this study is to further examine the flexibility of the Lipschitz anonymization method, particularly in regards to the statistical properties of geographical weighted regression, which were not demonstrated in the Kroll & Schnell paper. These results will yield greater information about further analyses that can utilize this method for distance matrix anonymization, and will provide privacy-conscious analysts the ability to use relevant spatial techniques while maintaining location data privacy.

METHODS

Lipschitz Embedding

A brief description of the Lipschitz embedding method will provide useful for understanding this paper. The procedure takes a random sample of size *k* in a spatial area of chosen size around the points of interest, and maps the original points into *d* dimensions based on distances to the random sample. Hence, there are essentially 3 tuning parameters to consider: the random sample size *k*, the number of dimensions *d*, and the spatial area in which to randomly select points.

For example, in the English hospitals examples in the Kroll & Schnell paper, the spatial area sampled from was the geographic area of England, and the best tuning parameters were determined to be $k = 5$ and $d = 20$.

Nearest Neighbors

Before extending the Lipschitz embedding procedure to geographically weighted regression, I sought to confirm the author's initial findings in regards to the nearest neighbors classification. The expectations of this were two-fold: to determine the difficulty and efficiency of implementing the Lipschitz embedding method using R and to validate the authors' results. This required following the R code examples outlined in the appendix of the Kroll & Schnell paper, while also implementing the English hospitals dataset given in the paper[10] combined with a dataset used to geocode the UK postal codes[11].

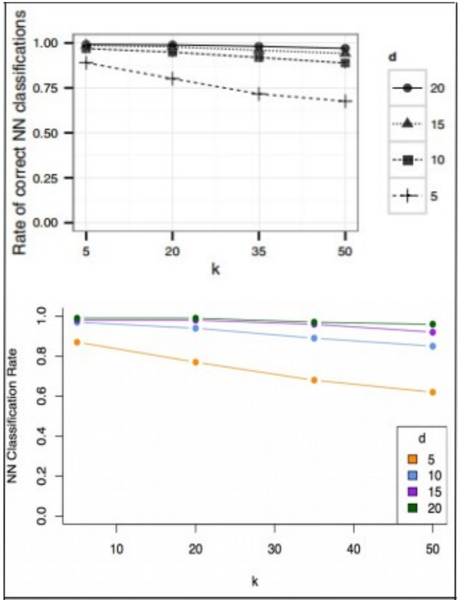

Figure 1: Comparison of nearest neighbor classification rates given in Kroll & Schnell (top) with my implementation (bottom), where parameter *k* is the size of the reference set of spatial coordinates used in the Lipschitz method, and *d* is the number of dimensions the Lipschitz method maps into.

Hence we can compare the NN classification results demonstrated in the study (Fig. 2 in the paper) to the approximate results found using my implementation. These can be found to the right in **Figure 1**. My results are a near-identical reproduction of the original estimates, bolstering the conclusions from the paper about the statistical properties maintained when using the Lipschitz embedding method. NN here is just the nearest graphical (e.g. geographic) neighbor in the set, which is mostly maintained under the embedding method. It is of note to point out that even though the NN classifications are maintained, the method itself seeks to distort the distance matrix between all points, which in turn impacts the effectiveness of the attack on the data. I include the full implementation of the Lipschitz method and nearest neighbor classification using R in **Appendix 1**.

Geographically Weighted Regression

Geographically weighted regression (GWR) is a spatial analysis method used to determine where locally-weighted coefficients in a regression model differ from the global values.[12] Essentially, it is an exploratory technique that seeks to determine geographic areas where fitted coefficient values from a global model do not necessarily capture local variations in the data. Kroll & Schnell note in the aforementioned paper that methods exist to calculate the spatial weight matrix used in GWR from a distance matrix. Hence, GWR *should* be robust to the Lipschitz-transformed distance matrix, suggesting that the method is still valid with the added benefit of location data privacy provided from the Lipschitz method.

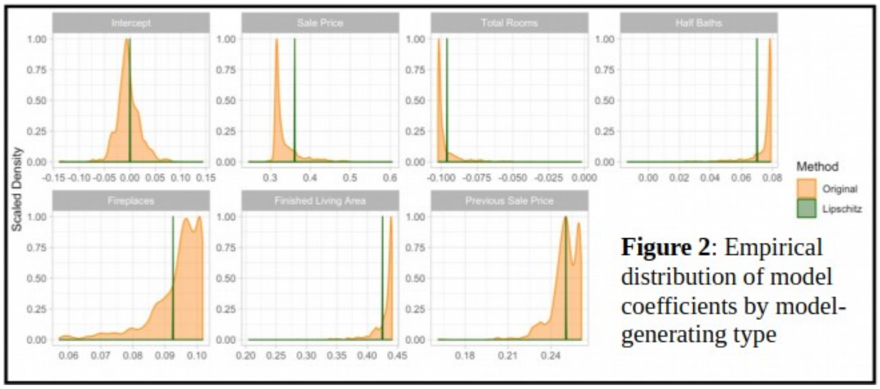

Figure 2: Empirical distribution of model coefficients by model-generating type

I sought to implement GWR in both the Lipschitz transformed and original datasets, using tools found in the spgwr R package[13] and Lloyd's *Local Models for Spatial Analysis*[14]. All analysis was done using R version 3.5.1[15]. Considering the difficulty of finding geocoded, healthcare datasets I opted to use the Allegheny County Property Assessment[16] to develop a relevant model to test the efficacy of GWR when using Lipschitz embedding. Hence the model building approach was as follows, built on a subset of 5000 samples from the dataset for the sake of computability:

1. Fit and model-build a traditional OLS model for the 'County Total' variable (Allegheny County's total property assessment value)
2. Fit the GWR model for all locations, yielding coefficients for:
   - Original dataset
   - Lipschitz embedded dataset, with embedding parameters held at most effective values per the Kroll & Schnell paper (referenced above)
3. Compare the GWR model in both sets on:
   - R^2 (e.g. model fit)
   - Empirical distribution of each model coefficient

The final model predicted 'County Total' with the following covariates: 'Sale Price', 'Total Rooms', 'Half Baths', 'Fireplaces', 'Finished Living Area', and 'Previous Sale Price'. The implementation of GWR for this study can be found in **Appendix 2**.

RESULTS

As we see in **Figure 2** and **Table 1**, the majority of coefficient estimates are within a rounding error of each other. The biggest discrepancy is for the 'Sale Price' covariate, whose estimate is off by less than 0.3, which itself is a large change in point estimate but not much practical change in terms of interpretability. Further, we see in **Figure 2** that since we are mapping the coordinates into more dimensions, we are by definition inducing more variability into the empirical distributions of the coefficients. For example, some estimates are essentially the same (Intercept, 'Previous

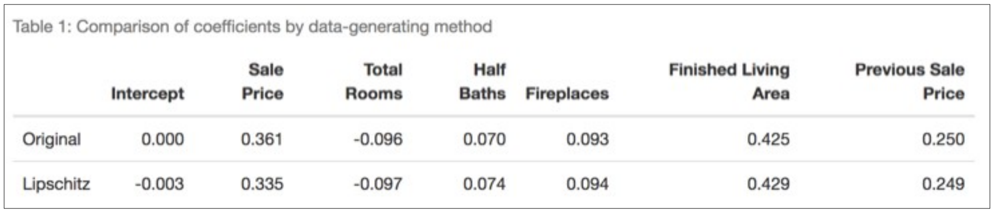

Table 1: Comparison of coefficients by data-generating method

	Intercept	Sale Price	Total Rooms	Half Baths	Fireplaces	Finished Living Area	Previous Sale Price
Original	0.000	0.361	-0.096	0.070	0.093	0.425	0.250
Lipschitz	-0.003	0.335	-0.097	0.074	0.094	0.429	0.249

Sale Price') suggesting that introducing we are randomness via both the GWR and Lipschitz process – a model and dataset dependent process.

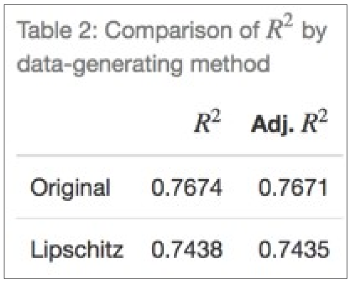

Table 2: Comparison of R^2 by data-generating method

	R^2	**Adj.** R^2
Original	0.7674	0.7671
Lipschitz	0.7438	0.7435

Consider also **Table 2** that compares the R^2 between the two models. In brief, for the privacy protections of the Lipschitz method, we are sacrificing roughly 2.5% of our model's predictability. This is not an insignificant loss, but consider also an analyst reporting GWR for exploratory analysis on a highly sensitive dataset. They may choose to utilize the Lipschitz method to allow their findings to be reproducible with far less privacy risk, rather than the more explanatory model that imposes significant privacy risk.

DISCUSSION

First, let's consider that this paper outlines a simple implementation to explore the utility of the Lipschitz embedding procedure. Based on the results above, it is reasonable to assume that for this simple case, GWR is robust to using the Lipschitz embedded distance matrix. However, as seen in **Figure 2**, there is non-negligible variability induced by the procedure, which in the asymptotic case may yield incorrect model standard errors. This could have a negative impact on model inference and prediction, and is something that needs to be examined further.

Another limitation worth noting is the initial privacy estimates in the Kroll & Schnell paper. Clearly, the estimates are dependent on the choice of tuning parameter, however, it is likely also dependent on the dataset used. If a robust privacy simulation were implemented, the resulting average precision estimates may be higher than suggested. Although the results are intentionally higher than in practice, given an assumption that the adversary has knowledge of the tuning parameters used, it is likely the true precision is larger than suggested because of the narrow window used to derive its estimate. Hence, this is another avenue to pursue in further research to validate the method's privacy in a more robust way.

There exist other spatial analysis methods that are similar to GWR in that they are dependent on weight and/or distance matrices. Thus, the results presented here may be applicable to other methods like the Moran's I test and PCNM (principal coordinates of neighbor matrices), given that the privacy benefits and asymptotic standard errors are investigated as noted above.

CONCLUSION

In conclusion, we recognize that this paper is one of the first to investigate the statistical utility of implementing a Lipschitz embedding anonymization method to provide privacy protections for spatial analysis methods. For the simple case, the embedding procedure works with GWR, with a few notable tradeoffs. Further research is needed to validate it's efficacy in GWR, in more complicated spatial analysis methods, and in terms of determining a robust privacy estimate of the Lipschitz procedure.

REFERENCES

1. Armstrong, M. P., Rushton, G., & Zimmerman, D. L. (1999). Geographically masking health data to preserve confidentiality. Statistics in medicine, 18(5), 497-525.
2. Cox, L. H. (1996). Protecting confidentiality in small population health and environmental statistics. Statistics in medicine, 15(17), 1895-1905.
3. Wieland, S. C., Cassa, C. A., Mandl, K. D., & Berger, B. (2008). Revealing the spatial distribution of a disease while preserving privacy. Proceedings of the National Academy of Sciences, 105(46), 17608-17613.
4. Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., & Vilhuber, L. (2008, April). Privacy: Theory meets practice on the map. In Proceedings of the 2008 IEEE 24th International Conference on Data Engineering (pp. 277-286). IEEE Computer Society.
5. Kroll, M., & Schnell, R. (2016). Anonymisation of geographical distance matrices via Lipschitz embedding. International journal of health geographics, 15(1), 1.
6. Kroll, M. (2014). A graph theoretic linkage attack on microdata in a metric space. arXiv preprint arXiv:1402.3198.
7. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.
8. Kroll M. A graph theoretic linkage attack on microdata in a metric space. Trans Data Priv. 2015;8(3):217–43.
9. Konc, J., & Janezic, D. (2007). An improved branch and bound algorithm for the maximum clique problem. proteins, 4(5).
10. https://www.whatdotheyknow.com/request/list_of_hospitals
11. https://www.freemaptools.com/download-uk-postcode-lat-lng.htm
12. Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., & Pebesma, E. J. (2008). Applied spatial data analysis with R (Vol. 747248717). New York: Springer.
13. Roger Bivand and Danlin Yu (2019). spgwr: Geographically Weighted Regression. R package version 0.6-32/r1748. https://R-Forge.R-project.org/projects/rspatial/
14. Lloyd, C. D. (2010). Local models for spatial analysis. CRC press.
15. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
16. https://catalog.data.gov/dataset/allegheny-county-property-assessments