# Statistical Efficacy of Distance Matrix Anonymization

**Coleman Harris**

## INTRODUCTION

•Disease mapping important in revealing spatial distribution of disease outbreak

- Potential privacy issues

•Location anonymization is one solution, but knowledge of pairwise distances between locations has privacy risks

- **Graph theoretic linkage attacks**

•Kroll & Schnell introduce method to anonymize distance matrices using **Lipschitz embedding**[1]

- Suggests reasonable privacy protections while upholding statistical properties

•Is **geographically weighted regression (GWR)** robust to using the Lipschitz embedded distance matrix?

## DATA

Allegheny County Property Assessment[2]

- Over 575k observations, 86 features

•Used sample of `n = 5000`

## METHODS

1.Fit OLS model for COUNTYTOTAL to inform GWR parameters

2.Fit GWR model for all locations, generating coefficients for:

- Original dataset
- Lipschitz embedded dataset

3.Compare GWR from two models on:

- $R^2$ (e.g. model fit)
- Empirical distribution of model coefficients

VANDERBILT UNIVERSITY MEDICAL CENTER

# **Geographically weighted regression** is robust to **data privacy** protections implemented via **Lipschitz embedding**

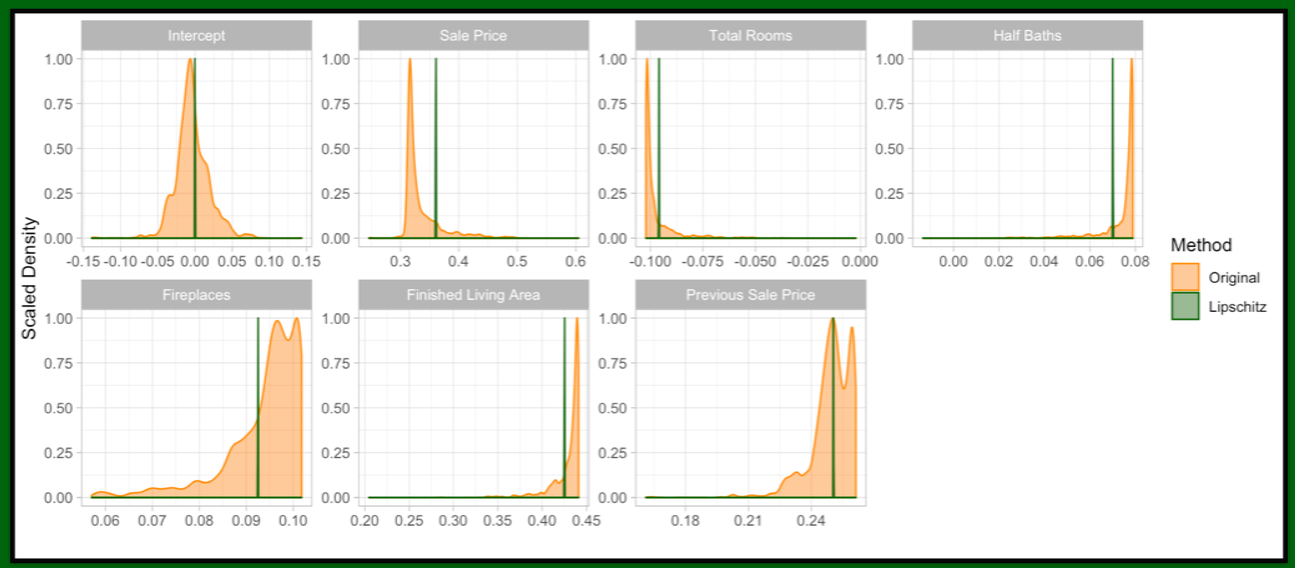


Figure 1: Empirical distribution of model coefficients, by data-generating method

Take a picture to view the code & analysis

## FULL MODEL

$$\begin{aligned}\texttt{County Total} = \hat{\beta}_0 &+ \hat{\beta}_1(\texttt{Sale Price}) + \hat{\beta}_2(\texttt{Total Rooms}) + \hat{\beta}_3(\texttt{Half Baths}) \\ &+ \hat{\beta}_4(\texttt{Fireplaces}) + \hat{\beta}_5(\texttt{Finished Living Area}) + \hat{\beta}_6(\texttt{Prev. Sale Price})\end{aligned}$$

## RESULTS

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 |

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 |

## DISCUSSION

- Coefficient estimates are similar, but embedded estimates have larger variability in empirical distribution
  - GWR model SEs potentially incorrect
- Tradeoff in privacy protections offered by embedding and reduction of ~2.5% in $R^2$
- Given that the Lipschitz procedure is compatible with GWR, this implies it can be implemented in other spatial techniques
  - Moran's I, PCNM
- Lipschitz parameters were set using estimates from Kroll & Schnell – different parameter sets may impact GWR models
- Because GWR relies on bandwidth, dataset size may play a role in its compatibility with the Lipschitz procedure

## REFERENCES

1. Kroll, M., & Schnell, R. (2016). Anonymisation of geographical distance matrices via Lipschitz embedding. International journal of health geographics, 15(1), 1.
2. https://catalog.data.gov/dataset/allegheny-county-property-assessments