Predicting Homicide Rates from Census Tracts in Brazil

By: Anahita Hassan, Prina Doshi, Elizabeth Szeto, Anika Joshi

Project Supervisor: Dr. Gonzalo Mena *Project Advisor: Dr. Dani Nedal*

Background & Introduction

- Minha Casa, Minha Vida (MCMV) program, launched in 2009, is a social housing initiative in Brazil
- MCMV aims to reduce the significant housing deficit by providing affordable housing for low income families through subsidized loans
- MCMV neighborhoods have been linked to increased violence and homicides
- We explore relationship between MCMV housing units and homicide rates in the municipality of Salvador in Bahia state

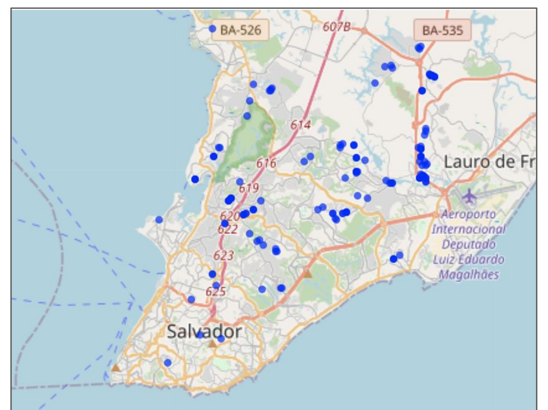

Fig 1. MCMV housing locations

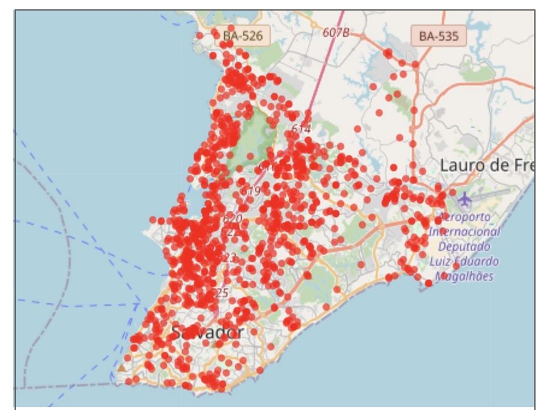

Fig 2. Homicide locations, 2022

Goal of this study: Develop a robust predictive model to forecast homicide rates using geographical and census-tract-level covariate data. Insights into homicide predictors will inform recommendations on future safer MCMV housing locations.

Data Processing & Methods

Data

- MCMV Housing data: Shapefile data containing geolocated coordinates of 2067 housing units built between 2009 and 2021 in Salvador
- Covariate data: 480 variables including demographic and housing characteristics for 5085 census tracts in Bahia from 2010 Brazil Census
- Homicide data: Individual-level records of 1159 homicide incidents in 2022, including temporal, geographic, and victim information

Data Feature Engineering

- Calculate distance between homicide location to nearest MCMV location
- Calculate homicide rate by census tract

Methods

- Predict homicide rate using neighboring census tract values
- *Generalized Linear Model (GLM) with regularization*
  - Lagged Covariates & Zero-Inflated Model

$$Y_{(x,y,t)} \sim \text{Poisson}(\lambda_{x,y,t})\; e^{\lambda_{x,y,t}} = e^{\sum_i C_{i,x,y,t}\cdot\alpha_i} \text{ for some rate } \lambda\text{, covariates } C_{1:N} \text{ over some space x,y and at time t}$$

- *Spatial Lag Model*
  - Captures local spatial dependencies

$$\text{Homicide count} = lagsarlm(\text{homicide\_rate} \sim \text{Longitude} + \text{Latitude} + \text{homicide\_MCMV\_dist} + \text{D1\_006} + \text{D1\_009} + \text{D1\_016})$$

* *D1_006* = Permanent private households owned and paid off; *D1_009* = Permanent private households ceded by employer; *D1_016* = Permanent private households with bathroom for the exclusive use of residents or toilet

- Model Comparison
  - GLM: Identifies general trends, but underpredicts in general
  - Spatial lag model: Underpredict in some areas, but these areas differ from the GLM model predictions
  - Largest residual tracts differ between the two models

Analysis & Results

GLM

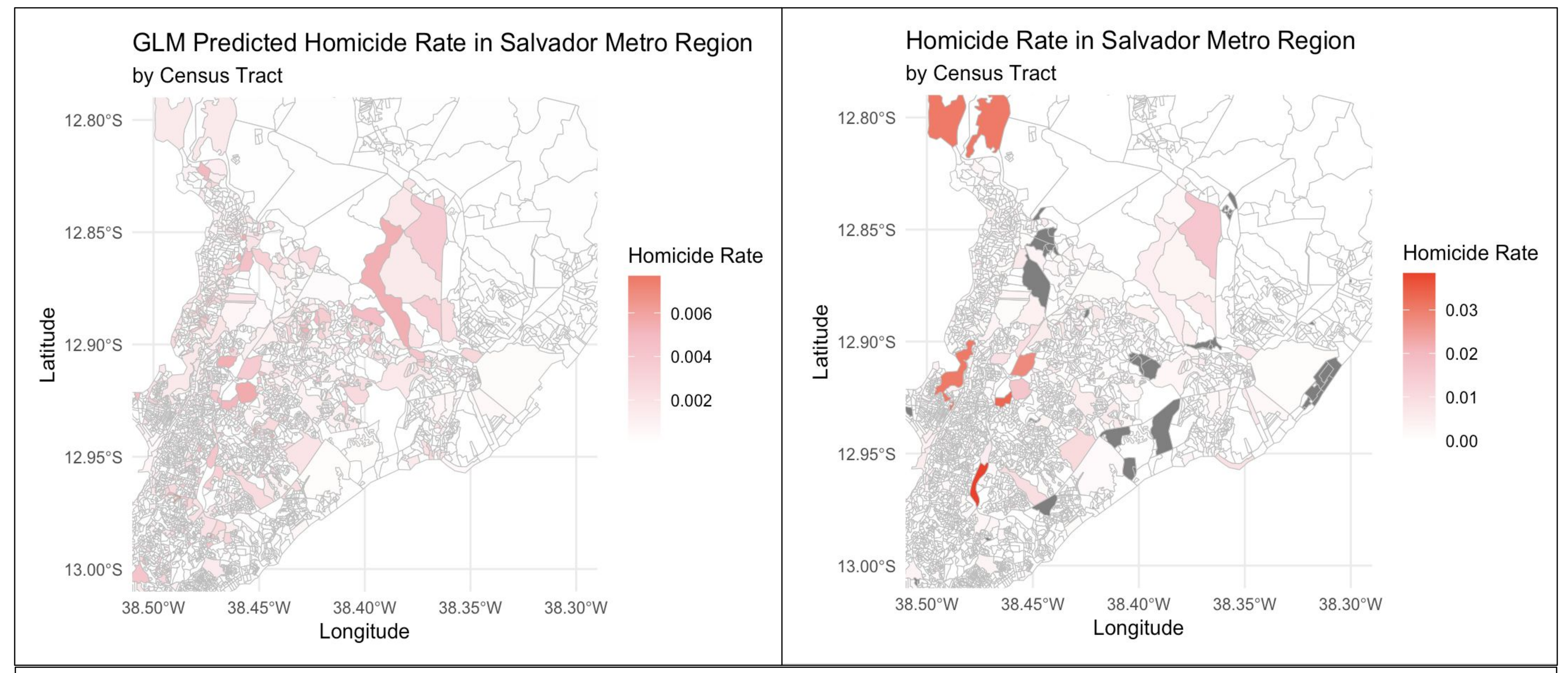

Fig 3. Side-by-side comparisons of GLM-predicted homicide rates and true homicide rates by census tract. Note the different color scales, which are necessary to ensure visibility.

- Model does a fair job of estimating areas with higher homicide rates, but consistently under-estimates rates in the tracts
- Income, location and population indicators impacted predictive power as shown in Figure 4

Spatial Lag Model

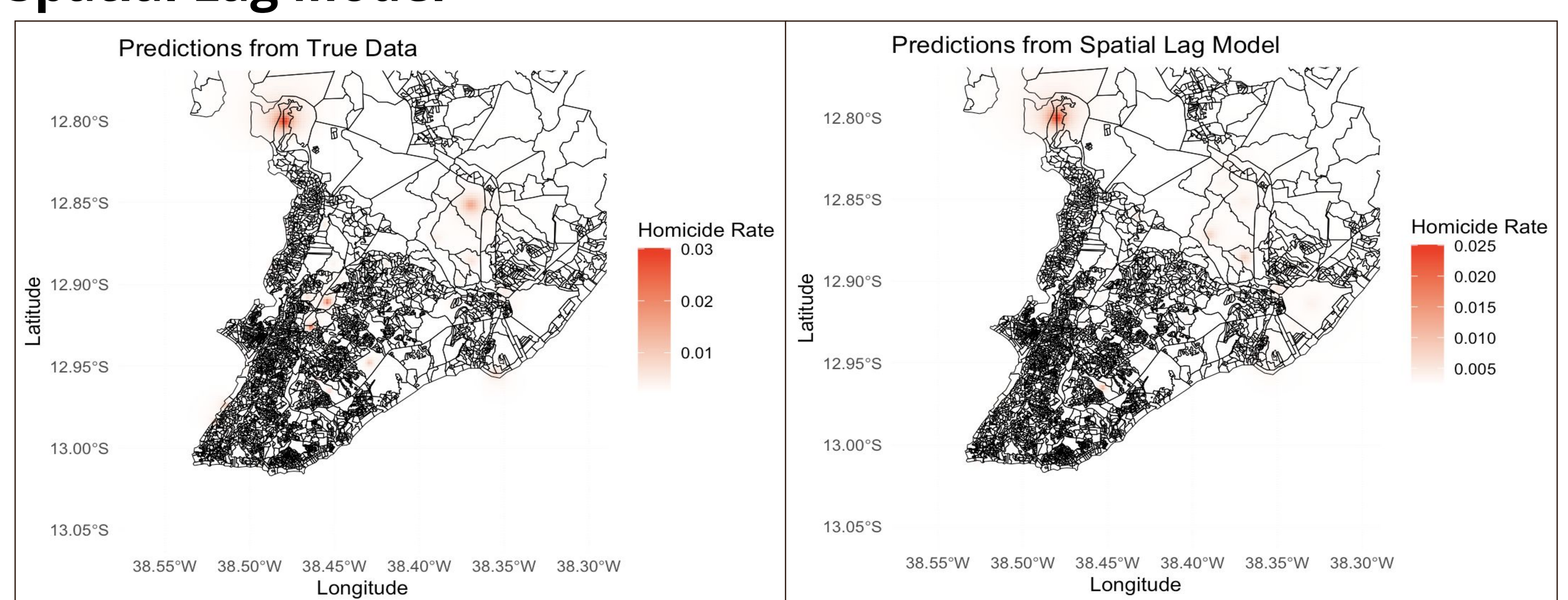

Fig 4. Side-by-side comparisons of spatial lag model-predicted homicide rates and true homicide rates by census tract. Note the different color scales.

- Mean homicide rate across both maps stays roughly the same, but there are high outliers with predictions from the true data
- Model predictions are more consistent and less extreme

Results

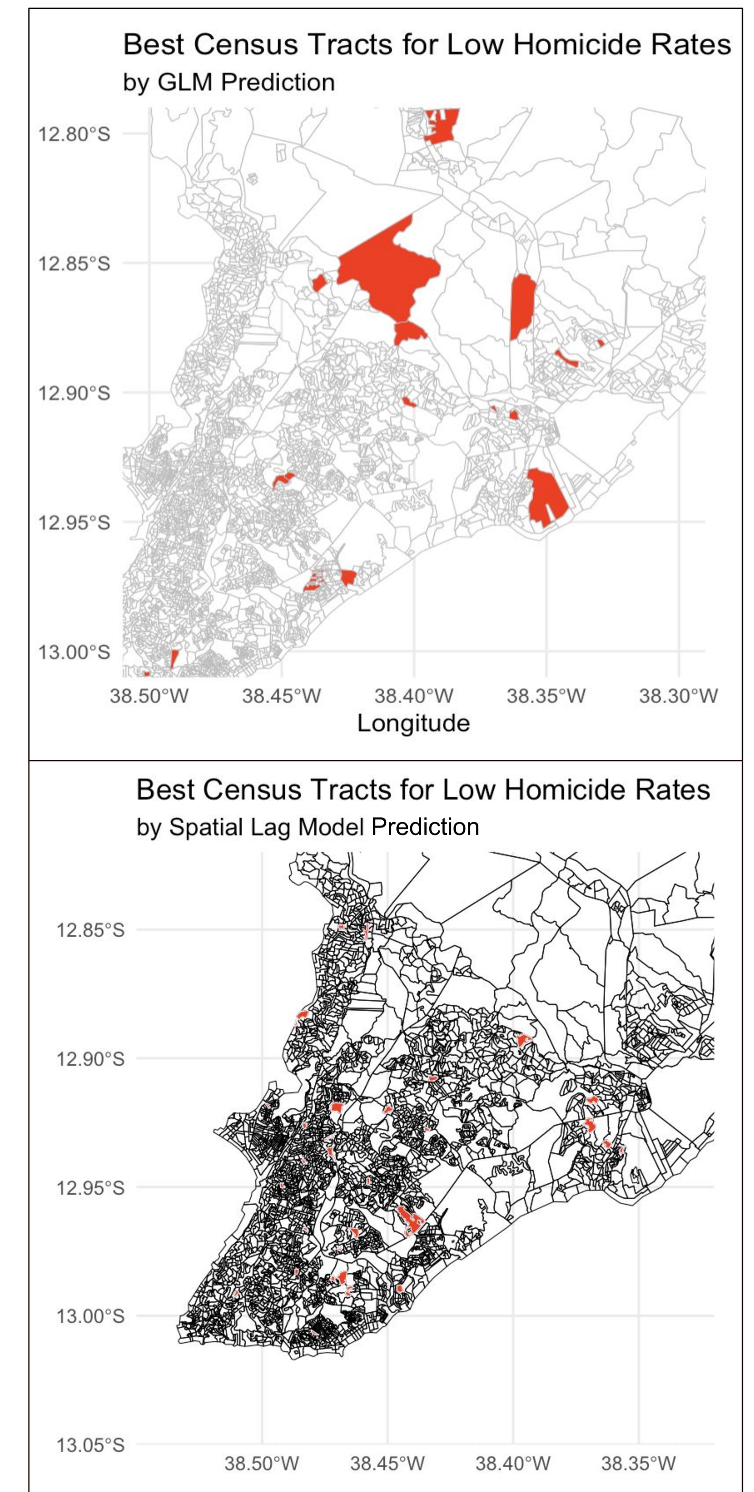

Fig 4. Census tracts projected to have low homicide rates, predicted by GLM (top) and Spatial Lag Model (bottom)

- Both models reveal distinct patterns in projected low-homicide census tracts, but with differences in size and characteristics of predicted areas
- Recommended GLM tracts are larger with varying homicide rates
- Spatial lag model highlights smaller, localized census tracts
- Coastal tracts, populated by affluent, educated, predominantly white populations, project lower homicide rates → trend is reflected in **both** models

Limitations of Models

- Temporal limitations with reported year of homicides
- Unreliability of covariate reporting
- Feedback loop in police response to homicides

Conclusions

- GLM model is more suitable for broad-regional wide analysis, but is more sensitive to socioeconomic factors
- Spatial lag model is better for localized targeted interventions
- For future work, we could refine predictive models to gain deeper insights into the dynamic between MCMV units and violent crime
  - Enhance spatial modeling techniques by experimenting with alternative methods for smoothing spatial lag effects
  - Incorporate supplementary datasets (eg. police activity records or other socio-environmental covariates)

References

1. Belchior, C, A., Gonzaga, G. & Ulyssea, G. (2023). Unpacking the Impacts of Low-Income Housing Programs: Evidence from Brazil. *IZA Discussion Paper No. 16113*. Retrieved from https://docs.iza.org/dp16113.pdf
2. Klotz, R., & da Costa, C. S. (2018). Moving to Opportunity and Inequality in Latin America: The Case of the Minha Casa Minha Vida Program. *Cities*, 74, 205–214. https://doi.org/10.1016/j.cities.2017.11.008
3. RioOnWatch. (2015). Minha Casa Minha Vida: A Look at Brazil's Social Housing Program. *RioOnWatch*. Retrieved from https://rioonwatch.org/?p=21175

Acknowledgements

Special thanks to Dr. Gonzalo Mena, Dr. Dani Nedal, and Dr. Aaditya Ramdas for their guidance and feedback throughout the research process.