Published in final edited form as:

Health Place. 2018 May ; 51: 61–67. doi:10.1016/j.healthplace.2018.02.013.

Predictors of firearm violence in urban communities: A machine-learning approach

Dana E. Goin¹, Kara E. Rudolph^{1,2}, and Jennifer Ahern¹

¹Division of Epidemiology, School of Public Health, University of California, Berkeley, California

²Department of Emergency Medicine, School of Medicine, University of California, Davis, Sacramento, California

Abstract

Interpersonal firearm violence is a leading cause of death and injuries in the United States. Identifying community characteristics associated with firearm violence is important to improve confounder selection and control in health research, to better understand community-level factors that are associated with firearm violence, and to enhance community surveillance and control of firearm violence. The objective of this research was to use machine learning to identify an optimal set of predictors for urban interpersonal firearm violence rates using a broad set of community characteristics. The final list of 18 predictive covariates explain 77.8% of the variance in firearm violence rates, and are publicly available, facilitating their inclusion in analyses relating violence and health. This list includes the black isolation and segregation indices, rates of educational attainment, marital status, indicators of wealth and poverty, longitude, latitude, and temperature.

INTRODUCTION

Homicide was the 16th leading cause of death in the United States in 2015, and the 8th leading cause of death among black Americans¹. The vast majority, 72.9%, of homicides were committed with firearms¹. Although firearm violence is a major public health challenge, the literature on predictors or consequences of community firearm violence is in its infancy. In particular, to our knowledge there has been no utilization of machine learning methods to identify predictors of firearm violence at the community level.

Epidemiologic study of firearm violence is relatively new. Studies that have considered violence as either an exposure or outcome have typically examined it at the individual-level, despite increasing evidence of the importance of contextual measures of violence^{2,3,4,5,6,7}. While individual-level studies help identify people at risk of violence perpetration or victimization, identifying features of communities that predict firearm violence rates is equally relevant for public health surveillance. An increasing number of studies have examined the associations of community violence with health outcomes using a multilevel

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

framework. These studies have controlled for possible confounders such as race/ethnicity, female-headed households, unemployment, housing quality, car ownership, public assistance income, median home value, educational attainment, and poverty but did not provide rationale for their inclusion^{8,9,10}. The selection of these variables is likely based on theory developed by sociologists and criminologists about the root causes of firearm violence. While theory is necessary to guide quantitative analysis, it can be complemented by empirical approaches that use machine learning on observed data to identify potential confounding variables. Furthermore, it is unknown how variables usually thought of as relevant to community violence—like employment, demographics, different measures of poverty, joblessness, or housing tenure—interact or jointly affect firearm violence rates. Machine learning methods are ideal for this sort of analysis but have never been used to consider a broad array of potential predictors of firearm violence.

We aim to fill this research gap by identifying a set of community level variables that best predict firearm violence levels at the community level, using the LASSO and random forest machine learning algorithms. We geographically link population-based firearm violence data to more than 300 community characteristics that are publicly available through the Census and other sources. We collected variables in the following categories: geography, climate, demographics, employment characteristics, commuting patterns, educational attainment, indicators of economic hardship, marital status, citizenship status, housing characteristics, the built environment, and urbanicity. These factors have been discussed as important factors in previous work on firearm violence^{11,12,13,14,15,16}. We then use a flexible, machine learning approach to identify the set of covariates most predictive of community firearm violence. This approach adds empirical evidence to our understanding of the community characteristics associated with firearm violence. Furthermore, these publicly available variables may be used to control for confounding in a wide range of studies analyzing the effects of community firearm violence on health, economic, or other social outcomes.

METHODS

Data

Violence—We calculated rates of community-level interpersonal firearm violence (hereafter referred to as firearm violence) by Zip Code Tabulation Area (ZCTA) using emergency department, hospitalization, and mortality data from California's Office of Statewide Health Planning and Development (OSHPD) and Vital Statistics. Violent incidents were geocoded based on the victim's home address at the time of hospitalization or death. We restricted our analysis to urban ZCTAs because the drivers of rural interpersonal firearm violence are very different in nature, frequency, and etiology from firearm violence in urban areas. Urban ZCTAs were defined as those with > 90% of the population living in an urban area as determined by the U.S. Census. This includes 85.9% of California's population. Interpersonal firearm violence totals by ZCTA were created by summing the total number of deaths and injuries attributable to assault or homicide by firearms from the Emergency Department records and Patient Discharge and Inpatient Hospitalization records from OSHPD and the death records from Vital Statistics. The injury-related records from OSHPD were classified from ICD-9 codes while the death records used ICD-10. The codes used to

identify firearm-related injuries were E9550 – E9554, E9650 – E9654, E970, and E9850 – E9854. The codes used to identify firearm-related deaths were X72 – X73, X93 – X95, Y350, and Y22 – Y24. Assaults were identified using codes E960 – E969 and E970 – E977. Homicides were identified from codes U01*, U02*, X85 – Y09, Y871, Y35*, and manner of death classified as homicide. To estimate yearly rates, the number of cases in each ZCTA was divided by the estimated number of people living in each ZCTA. These denominators were created by interpolating population counts at the ZCTA level from the 2010 Census using the intercensal and postcensal population estimates from the U.S. Census Bureau as benchmarks. The rates were then averaged over the 2007–2011 period. We modeled the square root of the violence rate to reduce the skewness across communities.

This study was reviewed and approved by the California Health and Human Services Agency Committee for the Protection of Human Subjects, and the University of California, Berkeley Committee for the Protection of Human Subjects.

Covariates—The American Community Survey (ACS) is a part of the Census that collects information on hundreds of community-level indicators. We compiled over 300 measures of community characteristics from the ACS’s 5-year files for ZCTAs for 2007–2011. We also included ZCTA-level alcohol outlet density from the Census ZIP Code Business Patterns and county-level temperature and precipitation measures from WestMap and the PRISM Climate Program^{20,21}.

These measures comprise many ecologically interrelated factors including age distribution, citizenship, means of commuting, housing characteristics, earnings, educational attainment, employment characteristics of families, employment status of individuals, participation in SNAP, income, language spoken at home, marital status, poverty status, racial and ethnic segregation and isolation, tenure and vacancy, veteran status, and work status. We excluded any variables that were missing for more than 5% of the observations, which reduced the list from 340 to 242 covariates. Summary statistics for the variables considered in the final analysis are available in Supplemental Table 1. The dissimilarity and isolation indices were built using racial and ethnic distributions from ACS variables and calculated for each ZCTA using standard dissimilarity and isolation formulas from the U.S. Census²². Economic segregation was calculated using an information theory index^{23,22}. There were 1004 urban ZCTAs and 765 rural ZCTAs for a total of 1769 ZCTAs in California. After dropping 84 (8.4%) urban ZCTAs that were missing at least one of the covariates, we had a final sample size of 920 urban ZCTAs.

Statistical Analysis

Our primary goal was to identify a set of variables most predictive of firearm violence. These predictive variables are ranked in terms of “variable importance”, which is defined as the change in mean squared error when the values of a covariate are randomly permuted across the data versus when they take their actual values²⁴. This measures the “importance” of a variable in the sense that it captures the improvement in prediction error attributable to one variable, holding others constant. This definition of variable importance is often used in the context of random forest regression²⁵.

We considered a comprehensive set of publicly available variables and used a machine learning approach to identify the subset most predictive of firearm violence. We combined the least absolute shrinkage and selection operator (LASSO) with random forest algorithms, and assessed the optimal set of predictors using 10-fold cross-validation. Cross-validation is a process of splitting the data into training and test sets, so each observation is included in both training the algorithm and assessing its predictive performance while minimizing the risk of overfitting.

We used LASSO to reduce dimensionality because prediction models with many low-relevance covariates can be highly variable²⁴. This also helps reduce the possibility for multicollinearity in the final models. LASSO works by setting a constraint parameter for the size of the coefficients in an ordinary least squares regression, which shrinks the coefficients toward zero. The coefficients are determined by the following equation²⁴

$$\hat{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_i \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (1)$$

Because the penalty parameter is the sum of the absolute value of the coefficients, those least relevant to the outcome are set equal to zero and drop out of the analysis. Only the variables with nonzero coefficients from the LASSO on at least one of the training sets were kept for the random forest portion of our analysis.

The random forest algorithm uses de-correlated regression trees and bagging (bootstrap aggregation) to create predictions²⁵. We describe the algorithm briefly here but note that more detailed descriptions are available for interested readers^{24,25}.

Regression trees are attractive because they capture interactions within data, and have low bias if they are grown sufficiently deep. They are created by binary recursive partitioning. This is most easily visualized using two covariates, but generalizes easily to higher orders. First, assume there are two variables (x_1, x_2) we would like to use to predict the values of a third variable (y). We can think of x_1 and x_2 as the two axes of a plane. We split the plane according to a value in x_1 . Call the splitting value v_1 . The plane is now divided into two halves, and we can predict the value of y using the average of the y values for each side of v_1 . We can also split the plane again, selecting a value in x_2 . We call this second splitting value v_2 . Now we have four regions of the plane, called R_1, R_2, R_3 , and R_4 (Figure 1). The predicted value of y in each region is the average of all observations whose set of (x_1, x_2) values correspond to that region. Mathematically, the prediction function is

$$f(x) = \sum_{m=1}^M c_m I(x \in R_m) \quad (2)$$

where m is the number of regions created by the splitting. This can also be represented as a tree, where the terminal nodes represent the final regions of the plane (Figure 2).

A drawback to this approach is that small perturbations in the data can create very different splits and therefore while the tree is unbiased in expectation, the variance can be large. In order to reduce the variance of individual trees, a common approach is to use bagging (bootstrap aggregation). The bagging procedure is as follows: draw bootstrap samples of the training data, estimate new trees on each sample, and average the predictions. Since each regression tree is identically distributed, the average of the bootstrapped trees preserves their unbiasedness. The average has lower variance than a single tree. Random forest achieves even better variance reduction by de-correlating the trees. This is done by randomly selecting a small subset of the covariates for each tree.

We applied the random forest algorithm to the reduced set of variables determined by the LASSO. In addition to determining the set of variables that best predicted the violence rates, we also ranked each variable based on importance. We repeated the random forest fitting and ranking for each of the 10 cross-validation training sets and averaged the rankings across the 10 folds to create a final rank for each variable. We calculated predictive power of the set of variables identified above in terms of R-squared and MSE applying the random forests algorithm to data from the cross-validation test sets. To identify which set of covariates produced the model with the best prediction, we sequentially added variables to the random forest in order of their importance ranking. We identified the set of variables with the best predictive performance via the R-squared and MSE.

We used R version 3.3.1 for all analyses²⁶.

RESULTS

Rates of firearm violence vary across the urban areas of California (Figure 3). Rates of firearm violence during the 2007–2011 period were highest in the Bay Area, especially in Oakland, San Francisco, Richmond, and El Cerrito (Figure 4). Rates of firearm violence were also high in Los Angeles, Compton, and places bordering Compton including Lynwood, Willowbrook, and West Athens (Figure 5). For diagnostic plots and information regarding spatial autocorrelation of firearm violence rates, please refer to the Supplemental Material.

Of the 242 variables originally included, the LASSO selected 125 unique predictors across the 10 folds. This set of variables predicted community firearm violence with an average R-squared of 0.759 and an MSE of 1.33×10^{-5} , using the random forest algorithm. Predictive performance varied across the folds, with a minimum R-squared of 0.689 and maximum of 0.836. The minimum MSE was 9.24×10^{-6} and the maximum was 1.79×10^{-5} . The set of 125 variables with their average variable importance scores is available in Supplemental Table 3.

The maximum average R-squared (0.778) and the minimum average MSE (1.22×10^{-5}) were obtained with 18 variables included in the model. The change in R-squared and MSE with the addition of variables is shown in Figure 6 and Figure 7. As Figures 6 and 7 show, predictive performance improves rapidly up until 10 predictors are included in the model. Using the top 10 predictors, the average R^2 is 0.766 and the average MSE is 1.29×10^{-5} .

This means that the top 10 predictors can explain 76.6% of the variation in community firearm violence. Predictive performance increases more gradually from 10 to 18 predictors and slightly worsens thereafter.

The 18 variables and their variable importance scores are in Figure 8. The top 5 covariates with the highest variable importance – the black isolation index, the black segregation index, the percent of households receiving foodstamps, the percent of men age 65+ with high school education, and the percent never married – achieved 0.708 average R-squared and 1.60×10^{-5} average MSE alone.

Six of the top 18 variables were related to education, predominantly among men, although they also included the educational status of veterans. Two marriage status variables were also identified as important as well as variables related to geography and temperature, family employment characteristics, methods of commuting, income, and wealth. Absent from the list were factors related to the age distribution, rates of citizenship, female-headed households, language spoken at home, and rates of housing tenure and vacancy. The sources of these variables are listed in Supplemental Table 2.

Many previous studies analyzing rates of firearm violence have modeled the outcome using a Poisson or negative binomial regression with the log(count) of incidents as the response variable and log (population) as an offset²⁷. Our data were over-dispersed, meaning the variance was significantly larger than the mean, so to model the outcome as a count we would need to use a negative binomial model. Unfortunately, there is no random forest equivalent for negative binomial regression. However, it is well known that square root transformations stabilize the variance of variables with either Poisson or negative binomial distribution and transform them to approximate Gaussian distributions^{28,29}. In addition, we show in the Supplemental Material that the assumptions for ordinary least squares are satisfied³⁰.

DISCUSSION

We used machine learning to identify a set of predictors of firearm violence rates at the community level. Results confirmed some relationships suggested by the literature and identified several additional variables of relevance. Our approach used a combination of variable selection and prediction algorithms and benefited from consideration of a broad range of publicly available variables and data from the urban population of California for five recent years.

The ability of racial-ethnic geographic segregation and isolation indexes to predict firearm violence rates lends empirical evidence to sociological theory suggesting that spatial restrictions on where minority groups were permitted to live can compound multiple disadvantages into one environment, multiplying risk factors for violence³¹. Two marriage status variables were also important predictors. This is consistent with theory suggesting that family structure is an indicator of social cohesion, socioeconomic status, and social capital —constructs that have long been understood as relevant for community violence^{13,3,32}.

Other important predictors of community violence included education and related aspects of socioeconomic status. Education may have immediate and inter-generational impacts on the violence rate through the economic opportunities it engenders. This is especially illustrated by the high variable importance scores of educational attainment among men older than age 65. In addition, the intersection of veteran status and low educational attainment has not been discussed in previous literature. This population may be a key demographic for targeted support by non-profit and/or governmental organizations, especially those that support reintegration of veterans.

The high ranking of variables describing poverty and use of food stamps further affirm the degree to which violence is linked to economic characteristics of communities. The proportion of employed people commuting by car may indicate that certain areas have fewer economic opportunities and those who have jobs must travel further and longer to reach them. In addition, the wealth of community members, characterized by the percent with rental properties or investment income, appears to be an important factor in predicting violence. These findings could suggest that historical patterns of economic opportunity or exclusion have implications for community violence rates.

While violence is known to be higher during summer months, variations in average temperature across neighborhoods are not usually considered important predictors of community violence, despite well-known effects of temperature on health^{8,9,10,33}. It is also not usually considered by sociologists or other social scientists who focus primarily on social and economic contexts of violence^{11,15,14,34,13}. However, its importance in this study provides additional evidence of the ecological relationship between local climate and violence^{35,36}.

Latitude and longitude were both selected in the optimal set of important predictors of firearm violence. This suggests there are place-based characteristics important for predicting rates of firearm violence that are not included in our current set. It is possible these factors could include variations in housing policy, social norms regarding violence or gang activity, measures of social cohesion, or characteristics of police presence^{37,32}. However, to our knowledge these measures are not publicly available, and our focus was on identifying a set of publicly available variables easily accessible to researchers.

We note several limitations to this study. The variables identified as most predictive of urban firearm violence rates are not independent and likely have complex causal relationships with one another. This set of variables serves as a starting point for further investigation of the processes behind violence, identification of communities at risk, and identification of variables to potentially control in analysis of the health, educational, economic and other effects of firearm violence. In addition, this study is an ecological analysis and does not include individual-level variables. We identified several ecological factors that predict firearm violence rates across communities, but this is not a multilevel study that can identify community factors associated with individual risk of victimization by firearm violence, controlling individual confounders. Nevertheless, the design of violence reduction initiatives at the community level may benefit from considering the set of factors identified by this analysis in their program development.

We used an approximation of zip codes (ZCTAs) because this was a large enough community unit to have stability in violence rates, and was small enough to capture local dynamics. However, there are also other levels of Census geography—blocks, block groups, tracts, or places—which may be of interest to other researchers, for which these variables may not be the best predictors of violence. Furthermore, previous work suggests that zip codes may not be the ideal community unit for public health surveillance³⁸. We cannot assess how the results of this study may differ with more granular community units, as ZCTAs were the smallest level of geographic information available for the data used in this study. However, given the frequency of the outcome, smaller levels of geography may have suffered from data sparsity issues.

Multicollinearity is a challenge when evaluating the relationships of correlated variables. Two features of our machine learning approach are designed to reduce multicollinearity. First, the LASSO algorithm will drop extremely correlated variables (or linear combinations of the same variable) by shrinking one of the variables' coefficients to zero. However, because we have done 10-fold cross validation, we have 10 separate LASSO regressions from which we use any variable that is selected at least once. This reduces the chance that a key variable will be eliminated from the analysis. Second, the random forest algorithm chooses a subset of covariates for each regression tree, which reduces the chance that very correlated variables will be chosen simultaneously.

The data used for our study are emergency department, hospitalization, and death records. Inherently, these data capture only fatal events and those severe enough to require hospital care. In addition, the ZCTAs correspond to the home address, and so our measure captures the rate of resident victimization by firearm violence. This is distinct from a measure of rates of incident occurrence in geography, which may be important to consider for understanding locations (parks, certain blocks, or specific intersections) where firearm violence occurs. However, identification of places where the rate of victimization is high for residents is also relevant to public health.

It is also important to note that our results are specific to interpersonal firearm violence. Interpersonal firearm violence and self-directed firearm violence are two different types of phenomena and conflating the two would potentially obscure meaningful relationships^{39,40}. It is an area for future work to identify important predictors of self-directed firearm violence rates.

Strengths of our approach to identify predictors of community-level violence include its flexible incorporation of a high-dimensional list of covariates and the machine learning algorithm which captures complex interactions between variables. This analysis included ZCTAs for the urban population of California. Due to the diversity of California's cities, we believe the results from this analysis could generalize to other similar urban populations. The approach used in this analysis, however, is just one of many potential ways to generate a list of predictors of firearm violence rates. Other methods may include alternative machine learning algorithms, including ensemble methods. Comparing lists generated by alternative methods and including additional community characteristics is an area for future work.

In summary, we identified predictors of firearm violence using a machine learning approach. The optimal set of 18 variables identified captured aspects of contemporary poverty and indicate the continued relevance of historical deprivation, especially related to housing segregation, education, and economic opportunity. These factors are all consistent with previous work. Temperature was also identified as an important predictor, which suggests climate and other environmental variables are relevant to understanding firearm violence in California. This set of variables is publicly accessible, and we hope other researchers will consider these variables when studying the prevention or effects of community firearm violence in the United States. In addition, these findings affirm that poverty, wealth, and educational opportunities are important factors to consider when developing interventions related to violence.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health/NICHD Office of the Director DP2HD080350, the Robert Wood Johnson Foundation Health and Society Scholars Program, and the University of California, Berkeley, Committee on Research.

References

1. WISQARS C. Atlanta, ga: Us department of health and human services, cdc; p. 18 Available at <https://webappcdc.gov/sasweb/ncipc/leadcaus10ushtml2015>
2. Rankin BH, Quane JM. Social contexts and urban adolescent outcomes: The interrelated effects of neighborhoods, families, and peers on african-american youth. *Social Problems*. 2002; 49(1):79–100.
3. Sampson RJ, Wilson WJ. *Toward a theory of race, crime, and urban inequality. Race, crime, and justice: A reader*. 1995; 1995
4. Cockerham WC, Hamby BW, Oates GR. *The social determinants of chronic disease*. 2017
5. Ford JL, Browning CR. Effects of exposure to violence with a weapon during adolescence on adult hypertension. *Annals of epidemiology*. 2014; 24(3):193–198. [PubMed: 24530410]
6. Boynton-Jarrett R, Ryan LM, Berkman LF, et al. Cumulative violence exposure and self-rated health: longitudinal study of adolescents in the united states. *Pediatrics*. 2008; 122(5):961–970. [PubMed: 18977974]
7. Bingenheimer JB, Brennan RT, Earls FJ. Firearm violence exposure and serious violent behavior. *Science*. 2005; 308(5726):1323–1326. [PubMed: 15919997]
8. Masho SW, Cha S, Chapman DA, et al. Understanding the role of violence as a social determinant of preterm birth. *American journal of obstetrics and gynecology*. 2016
9. Beck AF, Huang B, Ryan PH, et al. Areas with high rates of police-reported violent crime have higher rates of childhood asthma morbidity. *The Journal of pediatrics*. 2016; 173:175–182. [PubMed: 26960918]
10. Messer LC, Kaufman JS, Dole N, et al. Neighborhood crime, deprivation, and preterm birth. *Annals of epidemiology*. 2006; 16(6):455–462. [PubMed: 16290179]
11. Kennedy BP, Kawachi I, Prothrow-Stith D, et al. Social capital, income inequality, and firearm violent crime. *Social science & medicine*. 1998; 47(1):7–17. [PubMed: 9683374]
12. Gilbert L. Urban violence and health—south africa 1995. *Social science & medicine*. 1996; 43(5): 873–886. [PubMed: 8870151]

13. Blau JR, Blau PM. The cost of inequality: Metropolitan structure and violent crime. *American Sociological Review*. 1982;114–129.
14. Crutchfield RD. Labor stratification and violent crime. *Social Forces*. 1989; 68(2):489–512.
15. Choc J. Income inequality and crime in the united states. *Economics Letters*. 2008; 101(1):31–33.
16. McMahon SD, Todd NR, Martinez A, et al. Aggressive and prosocial behavior: Community violence, cognitive, and behavioral predictors among urban african american youth. *American Journal of Community Psychology*. 2013; 51(3–4):407–421. [PubMed: 23229395]
17. Slovak K, Singer MI. Children and violence: Findings and implications from a rural community. *Child and Adolescent Social Work Journal*. 2002; 19(1):35–56.
18. Slovak K, Singer M. Gun violence exposure and trauma among rural youth. *Violence and Victims*. 2001; 16(4):389–400. [PubMed: 11506448]
19. Peek-Asa C, Zwerling C, Stallones L. Acute traumatic injuries in rural populations. *American journal of public health*. 2004; 94(10):1689–1693. [PubMed: 15451733]
20. Daly C, Halbleib M, Smith JI, et al. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous united states. *International journal of climatology*. 2008; 28(15):2031–2064.
21. Center, WRC. WestMap Climate Analysis. PRISM Climate Mapping Program; URL <http://www.cdfa.dri.edu/Westmap/Westmaphome.php>
22. Iceland, J., Weinberg, DH. Racial and ethnic residential segregation in the United States 1980–2000. Bureau of Census; 2002.
23. Iceland, J. The multigroup entropy index (also known as theil's h or the information theory index). Vol. 31. US Census Bureau; 2004. p. 2006Retrieved July
24. Friedman, J., Hastie, T., Tibshirani, R. The elements of statistical learning. Vol. 1. Springer; Berlin: 2001. Springer series in statistics
25. Breiman L. Random forests. *Machine learning*. 2001; 45(1):5–32.
26. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2015. URL <http://www.R-project.org/>.
27. Osgood DW. Poisson-based regression analysis of aggregate crime rates. *Journal of quantitative criminology*. 2000; 16(1):21–43.
28. Anscombe FJ. The transformation of poisson, binomial and negative-binomial data. *Biometrika*. 1948; 35(3/4):246–254.
29. Bartlett M. The square root transformation in analysis of variance. *Supplement to the Journal of the Royal Statistical Society*. 1936; 3(1):68–78.
30. Freedman, DA. Statistical models: theory and practice. Cambridge university press; 2009.
31. Shihadeh ES, Flynn N. Segregation and crime: The effect of black social isolation on the rates of black urban violence. *Social Forces*. 1996:1325–1352.
32. Sampson RJ. Collective efficacy theory: Lessons learned and directions for future inquiry. *Taking stock: The status of criminological theory*. 2008; 15:149.
33. Hajat S, O'Connor M, Kosatsky T. Health effects of hot weather: from awareness of risk factors to effective health protection. *The Lancet*. 2010; 375(9717):856–863.
34. Bhavnani R, Donnay K, Miodownik D, et al. Group segregation and urban violence. *American Journal of Political Science*. 2014; 58(1):226–245. URL <http://dx.doi.org/10.1111/ajps.12045>.
35. Hsiang SM, Burke M, Miguel E. Quantifying the influence of climate on human conflict. *Science*. 2013; 341(6151):1235367. [PubMed: 24031020]
36. Rotton J, Cohn EG. Outdoor temperature, climate control, and criminal assault: The spatial and temporal ecology of violence. *Environment and Behavior*. 2004; 36(2):276–306.
37. Reardon SF, Bischoff K. Income inequality and income segregation1. *American Journal of Sociology*. 2011; 116(4):1092–1153.
38. Krieger N, Chen JT, Waterman PD, et al. Geocoding and monitoring of us socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? the public health disparities geocoding project. *American journal of epidemiology*. 2002; 156(5):471–482. [PubMed: 12196317]

39. Swanson SA, Colman I. Association between exposure to suicide and suicidality outcomes in youth. *Canadian Medical Association Journal*. 2013; 185(10):870–877. [PubMed: 23695600]
40. Fowler PJ, Tompsett CJ, Braciszewski JM, et al. Community violence: A metaanalysis on the effect of exposure and mental health outcomes of children and adolescents. *Development and psychopathology*. 2009; 21(01):227–259. [PubMed: 19144232]

Figure 1.
A visualization of a regression tree as a plane with two covariates.

Figure 2.
A regression tree with two covariates.

Figure 3.
Map of interpersonal firearm violence rates across urban ZCTAs in California from 2007–2011.

Figure 4.

Map of interpersonal firearm violence rates across urban ZCTAs in the Bay Area from 2007–2011.

Figure 5.

Map of interpersonal firearm violence rates across urban ZCTAs in the Los Angeles Area from 2007–2011.

Figure 6. MSE by number of predictors included. Each line represents one of the 10 cross-validation folds. The black line is the average across the 10 folds.

Figure 7.
 R^2 by number of predictors included. Each line represents one of the 10 cross-validation folds. The black line is the average across the 10 folds.

Figure 8.
Ranked importance of top predictors of community interpersonal firearm violence.