Published in final edited form as:
Int J Inj Contr Saf Promot. 2018 December ; 25(4): 443–448. doi:10.1080/17457300.2018.1467461.

Ability of crime, demographic and business data to forecast areas of increased violence

Daniel A. Bowen[a], **Laura M. Mercer Kollar**[a], **Daniel T. Wu**[b,c], **David A. Fraser**[d], **Charles E. Flood**[d], **Jasmine C. Moore**[b], **Elizabeth W. Mays**[b], and **Steven A. Sumner**[a]

[a]Division of Violence Prevention, National Center For Injury Prevention and Control, U.S. Centers for Disease Control and Prevention (CDC), Atlanta, GA, USA

[b]Grady Health System, Atlanta, GA, USA

[c]Department of Emergency Medicine, Emory University School of Medicine, Atlanta, GA, USA

[d]DeKalb County Police Department, DeKalb County, GA, USA

Abstract

Identifying geographic areas and time periods of increased violence is of considerable importance in prevention planning. This study compared the performance of multiple data sources to prospectively forecast areas of increased interpersonal violence. We used 2011–2014 data from a large metropolitan county on interpersonal violence (homicide, assault, rape and robbery) and forecasted violence at the level of census block-groups and over a one-month moving time window. Inputs to a Random Forest model included historical crime records from the police department, demographic data from the US Census Bureau, and administrative data on licensed businesses. Among 279 block groups, a model utilizing all data sources was found to prospectively improve the identification of the top 5% most violent block-group months (positive predictive value = 52.1%; negative predictive value = 97.5%; sensitivity = 43.4%; specificity = 98.2%). Predictive modelling with simple inputs can help communities more efficiently focus violence prevention resources geographically.

Keywords

Violence; violent injuries; crime forecasting; machine learning

Introduction

Many community- or neighbourhood-level violence-related public health prevention strategies seek to focus public health resources and interventions on geographic locations

Contact: Steven A. Sumner hvo5@cdc.gov.

Authors' contributions
D. A. Bowen, L. M. Mercer Kollar, D. T. Wu, D. A. Fraser, C. E. Flood, J. C. Moore, E. W. Mays, and S. A. Sumner helped conceptualize the study. L. M. Mercer Kollar, D. A. Fraser, and C. E. Flood collected data. D. A. Bowen and S. A. Sumner analyzed data and completed the initial draft of the manuscript. All authors interpreted results, edited the manuscript, and approved the manuscript.

Disclosure statement
We declare no competing interests.

and times when the likelihood of violence is greatest. Such interventions potentially include the deployment of street outreach and conflict mediation workers, social norms campaigns and policies governing the local environment (Butts, Roman, Bostwick, & Porter, 2015; Sampson, Raudenbush, & Earls, 1997; Sumner et al., 2015; The Guide to Community Preventive Services, 2015; Webster, Whitehill, Vernick, & Curriero, 2013). However, there has been limited work by public health researchers to evaluate modern forecasting methods to prospectively predict violence occurrence over brief time periods, which could be used to guide such prevention strategies.

Indeed, to date, we are aware of only one published article by public health researchers in this nascent area of predictive epidemiology (Henry et al., 2014). To help guide the allocation of community health outreach workers, researchers in Chicago used historical crime data from the Chicago Police Department and two neighbourhood demographic characteristics to develop a model to predict future increases in violent crime at the level of city census tracts over a three-month time window (Henry et al., 2014).

Opportunity exists to advance such work and further inform the public health practice of violence prevention by exploring and incorporating additional factors demonstrated by researchers to be associated with violence at the community level. Such factors may include the density of high-risk businesses (Amandus et al., 1996; Scribner, MacKinnon, & Dwyer, 1995; Zhu, Gorman, & Horel, 2004), additional violent and non-violent crimes, and other sociodemographic characteristics of neighbourhoods (Coulton, Korbin, Su, & Chow, 1995; Cunradi, Caetano, Clark, & Schafer, 2000; Pinchevsky & Wright, 2012). Furthermore, comparing the individual and composite predictive ability of these categories of variables would help inform practitioners about the most useful data sources for forecasting activities.

Although work in crime forecasting work has also been undertaken by researchers in the field of criminology, such modelling work is often focused on the goal of improving policing strategies and the presentation of model results in such research is often unique to this goal. Consequently, in this work we aim to display the full set of epidemiologic test characteristics (positive predictive value (PPV), negative predictive value (NPV), sensitivity and specificity) that public health practitioners rely on to guide strategy selection and fully examine all aspects of model performance.

Specifically, our paper seeks to advance upon the previous pilot work in Chicago by validating a model to predict the future occurrence of violent crimes at a smaller population-level geographic unit (census block groups) and in a narrow time window (monthly). We also seek to inform public health practitioners in detail about the performance of three types of data – historical crime, demographic and business information.

Materials and methods

Study area and time period

The present study focuses on DeKalb County, Georgia, a large urban and suburban county in the Atlanta metropolitan area. A single police agency, the DeKalb County Police Department, provides services to the majority of DeKalb County residents, and 279 census block groups contained within the DeKalb County Police Department’s patrol area are included in this study. Because police precinct boundaries do not perfectly align with census block groups, we clipped block groups on police precinct boundaries and excluded all block groups that had <0.25 square kilometres of area within the police precincts. Census block groups in DeKalb are often defined by man-made or natural features (such as large roads/ highways, forested areas and parkways) that help delineate neighbourhoods. The mean number of residents in each block group is approximately 1800 persons. We chose to focus on census block groups as it was a smaller unit of analysis than that used in the violence hotspot public health research in Chicago and a smaller unit of analysis than the DeKalb County police currently uses for allocation of police patrols (approximately 30 ‘territories’), which may ultimately lead to more efficient deployment of resources. Lastly, for statistical modelling, this study uses data over a four-year period from January 2011 through December 2014.

Outcome, predictors and data sources

Similar to the work in Chicago (Henry et al., 2014), our model predicts the occurrence of violent crimes – defined as a composite of homicide, aggravated assault, robbery and rape. All violence data in this study are from official police reports and these violence types are identified through Uniform Crime Report (UCR) codes assigned to each record. UCR codes are used to standardize reporting of crimes by police departments to the Federal Bureau of Investigation. Each crime record is coded to the block-level (centre of block), the police standard for public reporting.

The unit of analysis for this study is a block group-month and for each year of data there are 3348 observations (279 block groups × 12 months). Our main outcome of interest is whether a given block group experiences a particularly high level of violence in a given month (defined as an amount of violence that places it in the top fifth percentile of all block group-months). This equates to the presence of four or more violent crimes in a block group during a month-long period.

To predict the future occurrence of violence at the block group level we assembled three main classes of potential predictors to test their relative contributions to the model: (1) historical crime data, (2) neighbourhood demographic characteristics and (3) local business information. First, for historical crime information, we created monthly lagged variables going back 24 months for each block group that includes the total number of violent crimes. We also created 24 months of lagged variables for property crimes (burglary, larceny and motor vehicle theft) to use as a predictor in our model. In addition to lags at the specified number of months, lagged variables were created that included the cumulative number of violence- and property-related crimes at each time point. All crime predictors were continuous variables.

Second, to assemble block group demographic variables, the U.S. Census Bureau’s American Community Survey (ACS) was used (United States Census Bureau. American Community Survey (ACS), 2016). The ACS includes the most detailed measures of citizen demographics available; however, results for such variables are present as five-year averages (2010– 2014) due to the need to produce stable estimates within small geographic areas.

Thus, the demographic variables included help to inform the baseline, chronic level of violence risk and could contribute to short-term violence risk through their interactions with other variables; however, the values of such predictors are stable through the study period. For this study, the 20 ACS variables with the highest correlation coefficient with the number of violent crimes in a block group were selected for further exploration.

Third, predictor variables were created which specified the presence of various business types within each block group by year. These data were obtained from a commercial data provider, referenceUSA, a commonly used source for business and residential information (Grimm, Moore, & Scanlon, 2013). Seven business types were incorporated into the data-set, including liquor stores, bars, gas stations, check cashing stores, escort services, massage parlours and pawn shops. Business types were chosen based on prior associations in the literature (Scribner et al., 1995) as well as unpublished local investigations.

Statistical model

The above data-set preparation yielded hundreds of predictor variables on historic crime, demographic and business information. It should be noted that the goal of this work was not to produce an explanatory model but rather to maximize prospective predictive accuracy for the deployment of public health interventions. Hence, the focus of predictive modelling is often not on exploring the association between individual predictors and the outcome, but rather on the composite performance of a large number of signals.

Consistent with standard practice in machine learning, we first trained our statistical model on one portion of data and then tested the predictive ability of the stored model parameters on a new data-set not used for training of the model (Chen, Liaw, & Breiman, 2004). Specifically, we trained our model on violent crime in block group-months in 2013 (using lagged historical crime information going back to 2011 as predictors) and then tested our model on 2014 data not used in the original model training. This process helps ensure the generalizability of the model when applied prospectively; the test characteristics reported in this paper are those prospectively validated on the 2014 testing data-set. Our modelling was performed stepwise by first adding historic violent crime predictors lagged over a 12-month period followed by models that included historic violent crime predictors lagged over a 24-month period, historic non-violent property crime, demographic data, and finally, business density data.

For our prediction model, we used Random Forests, a leading machine learning technique for criminology applications (Berk & Bleich, 2013). In brief, the Random Forest algorithm iterates through a data-set a specified number of times, each time calculating a decision tree from a random sample of the entire data-set and using a random sample of predictor variables to consider for the branch points of the tree. At the conclusion of the iterations through the data-set, these multiple decision trees are then collectively used to produce a final classification scheme whose performance has been demonstrated to be superior to regression in many applications (Berk & Bleich, 2013); this is also confirmed by our own sensitivity analyses with the data-set. For the task of predicting the relatively rare outcome of having four or more violent crimes in a block group-month we used down-sampling of event to non-event months in a 1:7 ratio to train the model; this allows the algorithm to better learn to recognize rare events by presenting them to the algorithm more frequently (Chen et al., 2004). For examination of the business and demographic data alone, a 1:3 down-sampling ratio was found to be ideal. Models were implemented using R statistical software (v 3.2.2) with the RandomForest package's implementation of the algorithm.

Results

In 2014, the year used for prospective validation of the forecasting algorithm, 2508 violent, interpersonal crimes were reported across the 279 block groups. Of the 3348 block group-months, 1305 (39.0%) experienced at least one violent crime, while 144 (4.3%) experienced four or more violent crimes. Throughout 2014, 52 out of the 279 block groups had at least one month where four or more violent crimes were reported.

Figure 1 provides a map of the geographic distribution and frequency of four or more violent crimes having been reported in the 279 block groups over a two-year period. This figure shows that areas of high violent crime are distributed across DeKalb County. Figure 2 presents a tabular heatmap showing each of the 279 block groups and all 24 months from 2013 to 2014. While some block groups appear to be more heavily affected by violent crime than others, there is often heterogeneity from month to month.

Table 1 presents the test characteristics for three models that exclusively consider historical crime, business or demographic data alone. Demographic data, which was static throughout the study time period, had low PPV (27.7%) but high sensitivity (71.5%), when used as the sole predictive input. Business data, which changed yearly, had marginally higher PPV (35.6%) but lower sensitivity (32.6%). Historical crime data had a relatively high PPV (50.4%) and a sensitivity (43.1%) that was intermediate. Examination of the F1 score, which is a measure of a test's accuracy which uses the PPV and sensitivity as inputs, indicated that historic crime data had the greatest performance of any single data source.

Table 2 presents the results of several random forest models, each one adding a successive class of variables to the model and attempting to predict the occurrence of four or more violent crimes in a given block group-month. Randomly allocating the same number of block-group months with four or more violent crimes that occurred in 2013–2014 yields only a PPV of 5.3% and an equally low sensitivity of 5.6%. Models 1 through 3 include historic violent crime variables as predictors. We see incremental increase in PPV and sensitivity as these variables are lagged. A full model (Model 6) with historic crime, business and demographic data increased PPV to above 52% and sensitivity to above 40%, providing meaningful improvements in forecasting ability beyond simple prediction models. However, we note relatively little improvement in model performance when adding business and demographic data to historical crime information.

As a sensitivity analysis we compared random forests to a conventional logistic regression model that used backwards stepwise elimination of variables with $p > .20$. As with the random forest model, we used 2013 data as our prediction data-set and 2014 as our testing data-set. For predicting the main outcome of four or more violent crimes in a given month, the logistic regression model yielded a PPV of 49.3% and sensitivity of 25.7%. Thus, to maintain a PPV nearly equal to the random forest model, the logistic regression model suffers an approximately 20% lower sensitivity, missing a large fraction of total violent crime volume.

Discussion

The ability to successfully focus violence prevention efforts to both communities and periods of time of highest risk is an important goal of many violence prevention programmes. However, little published research is available in the public health literature to guide practitioners in forecasting violence in both space and time. Consequently, this study investigates the performance and trade-offs of various data sources for the prediction of geographic hotspots of interpersonal violence. Performing predictive modelling with even simple inputs can help communities focus violence prevention resources geographically and temporally.

In our study, we found a model that used historic crime data, demographic data, and business data (Model 6) yielded the best performance. However, examining each data source individually, we noted that historic crime data was the predictor that best maximized both PPV and sensitivity. Demographic data alone yielded a high sensitivity but low PPV while business data yielded only marginally better PPV but with a lower sensitivity compared to the demographic data. The strongest predictors among the demographic data were variables that were proxies for poverty, such as the number of individuals on food stamps in a given block group. Since relatively large areas of a city may have notable poverty but lack other factors which are associated with crime, this helps explain the high sensitivity but low PPV produced. The relatively low PPV and sensitivity noted when exclusively using the business data could be a product of using a single commercial data source to acquire business data. Although commonly used, business registries are often incomplete. This might have negatively affected the PPV and sensitivity of the business data model.

Current forecasting systems aimed at deploying police resources often use only historical crime data (Mohler et al., 2015); our findings help confirm the rationale for the focus on this data source. The historical crime data likely demonstrates the best balance between PPV and sensitivity as it is the only variable that exhibits significant temporal variation. Consequently, to further improve forecasting efforts, there exists a need for additional predictors that both fluctuate significantly over time and can be located to a small geographic area. New research in this area, for example, is examining the potential contributions of variables such as measures of population movement from cell phone data or social media messages transmitted in an area (Bogomolov et al., 2014; Gerber, 2014; Wang & Gerber, 2015).

Some limitations of this work and directions for future research should be mentioned. First, while this study focused on examining the utility of various data sources for violence forecasting and used a leading machine learning model, additional work is needed in comparing other modern machine learning approaches and head-to-head testing of such models (National Institute of Justice. Real-Time Crime Forecasting Challenge, 2016). However, we chose to primarily examine random forests in this paper due to the increased likelihood of their application by public health practitioners due to being a robust algorithm that is also relatively simple to implement. It is also important to note limitations arising from an examination of irregularly shaped geographic entities, such as census block groups. The geographic boundaries employed can have an arbitrary nature which introduces some statistical biases. For example, even when a hotspot lies on the boundary between two areas, one spatial unit may show a disproportionate number of incidents relative to the neighbouring space simply because of the arbitrary boundary, even though violent crimes are located essentially between these two spaces. Similarly, if the hotspot is within two or more areas, the number of violent incidents may be distributed among all of the nearby spaces, masking the truly elevated incidence. Some solutions to these problems involve the application of strategies to produce continuous rate maps (Ratcliffe & McCullagh, 1999). Future work may also involve attempts at forecasting violence for an even smaller time window than utilized in this study. We selected the goal of creating predictions at the month-level as the occurrence of violent crimes was still a relatively rare outcome and further class imbalance would have affected model performance. However, the parameters of each forecasting project is ultimately dependent on the type of intervention being delivered and examinations of more narrow time windows should be explored. Lastly, it is important to note that successful forecasting of violence is dependent on the underlying data. Some data sources, such as violent crime, are known to be underreported (Truman & Morgan, 2016) and any models based on this data are affected by any ascertainment bias that exists.

Conclusion

Nonetheless, this work makes several important advances to the public health literature. The results provide health agencies or violence prevention organizations insight about the different types of data used in geographic forecasting systems, their relative merits and the potential improvement in focusing of resources that could be obtained. This work can also be applied to other data sources health agencies and violence prevention organization may use. Real-world violence prevention efforts often have significant resource constraints and using a data-driven approach to focus prevention activities may lead to improved violence prevention and community well.

Acknowledgments

Funding

This research was partially supported by a grant from the Robert Wood Johnson Foundation: “Piloting the Cardiff Model for Violence Prevention.” The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

References

Amandus HE, Zahm D, & Friedmann R, Ruback RB, Block C, Weiss J, … Kessler D (1996). Employee injuries and convenience store robberies in selected metropolitan areas. Journal of Occupational and Environmental Medicine, 38(7), 714–720. [PubMed: 8823663]

Berk RA, & Bleich J (2013). Statistical procedures for forecasting criminal behavior. Criminology & Public Policy, 12(3), 513–544.

Bogomolov A, Lepri B, Staiano J, Oliver N, Pianesi F, & Pentland A (2014). Once upon a crime: Towards crime prediction from demographics and mobile data. In Proceedings of the 16th international conference on multimodal interaction (pp. 427–434). Istanbul, Turkey: ACM.

Butts JA, Roman CG, Bostwick L, & Porter JR (2015). Cure violence: A public health model to reduce gun violence. Annual Review of Public Health, 36, 39–53.

Chen C, Liaw A, & Breiman L (2004). Using random forest to learn imbalanced data (pp. 1–12). Berkeley: University of California.

Coulton CJ, Korbin JE, Su M, & Chow J (1995). Community level factors and child maltreatment rates. Child Development, 66(5), 1262–1276. [PubMed: 7555215]

Cunradi CB, Caetano R, Clark C, & Schafer J (2000). Neighborhood poverty as a predictor of intimate partner violence among White, Black, and Hispanic couples in the United States: A multilevel analysis. Annals of Epidemiology, 10(5), 297–308. [PubMed: 10942878]

Gerber M (2014). Predicting crime using Twitter and kernel density estimation. Decision Support Systems, 61, 115–125.

Grimm KA, Moore LV, & Scanlon KS (2013). Access to healthier food retailers—United States, 2011. CDC Health Disparities and Inequalities Report—United States, 2013, 62(3), 20–26.

Henry DB, Dymnicki A, Kane C, Quintana E, Cartland J, Bromann K, … Wisnieski E (2014). Community monitoring for youth violence surveillance: Testing a prediction model. Prevention Science, 15(4), 437–447. [PubMed: 23494404]

Mohler GO, Short MB, Malinowski S, Johnson M, Tita GE, Bertozzi AL, & Brantingham PJ (2015). Randomized controlled field trials of predictive policing. Journal of the American Statistical Association, 110(512), 1399–1411.

National Institute of Justice. Real-Time Crime Forecasting Challenge (2016). Retrieved October 14, 2016, from http://www.nij.gov/funding/Pages/fy16-crime-forecasting-challenge.aspx.

Pinchevsky GM, & Wright EM (2012). The impact of neighborhoods on intimate partner violence and victimization. Trauma, Violence, & Abuse, 13(2), 112–132.

Ratcliffe JH, & McCullagh MJ (1999). Hotbeds of crime and the search for spatial accuracy. Journal of Geographical Systems, 1(4), 385–398.

Sampson RJ, Raudenbush SW, & Earls F (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. Science, 277 (5328), 918–924. [PubMed: 9252316]

Scribner RA, MacKinnon DP, & Dwyer JH (1995). The risk of assaultive violence and alcohol availability in Los Angeles County. American Journal of Public Health, 85(3), 335–340. [PubMed: 7892915]

Sumner SA, Mercy JA, Dahlberg LL, Hillis SD, Klevens J, & Houry D (2015). Violence in the United States: Status, challenges, and opportunities. JAMA, 314(5), 478–488. [PubMed: 26241599]

The Guide to Community Preventive Services. (2015). Retrieved July 27, 2016, from http://www.thecommunityguide.org/.

Truman JL, & Morgan R (2016). Criminal victimization, 2015 Washington, DC: Bureau of Justice Statistics.

United States Census Bureau. American Community Survey (ACS). (2016). Retrieved July 27, 2016, from https://www.census.gov/programs-surveys/acs/

Wang M, & Gerber MS (2015). Using Twitter for next-place prediction, with an application to crime prediction. In Proceedings of the Computational Intelligence IEEE Symposium Series (pp. 941–948). Cape Town, South Africa: IEEE.

Webster DW, Whitehill JM, Vernick JS, & Curriero FC (2013). Effects of Baltimore's safe streets program on gun violence: A replication of Chicago's CeaseFire program. Journal of Urban Health: Bulletin of the New York Academy of Medicine, 90(1), 27–40. [PubMed: 22696175]

Zhu L, Gorman DM, & Horel S (2004). Alcohol outlet density and violence: A geospatial analysis. Alcohol and Alcoholism, 39(4), 369–375. [PubMed: 15208173]

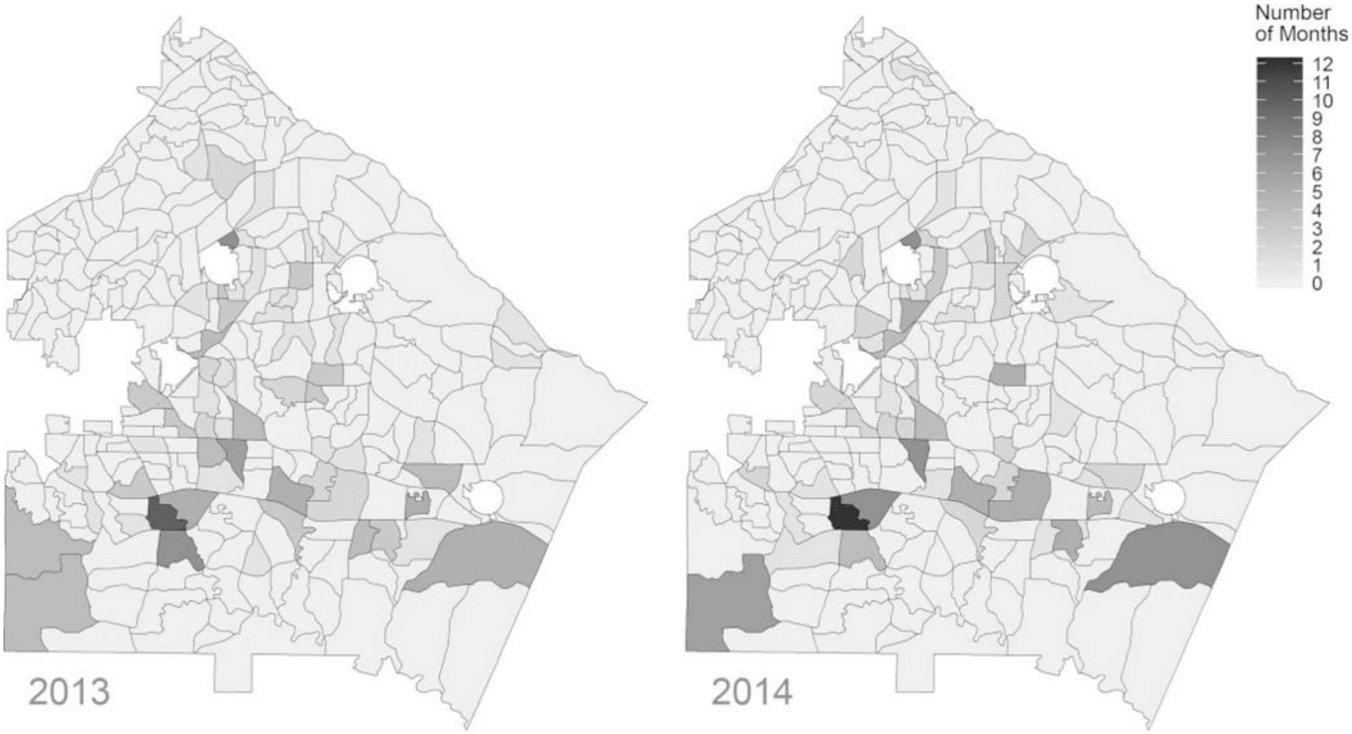

Figure 1.

Number of months per year with four or more violent crime incidents in each block group, 2013–2014, DeKalb County, Georgia.

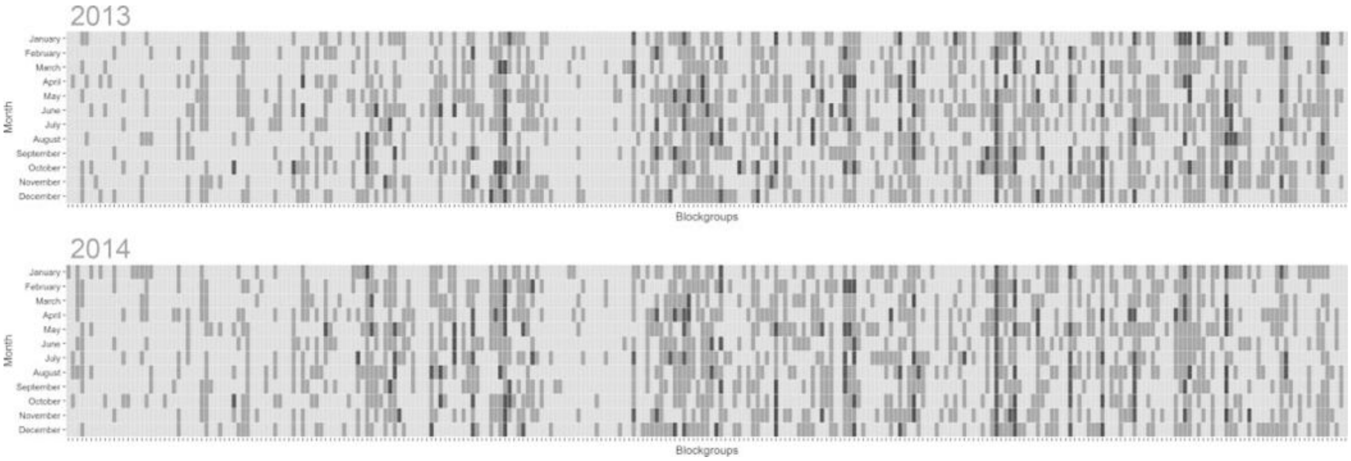

Figure 2.

Occurrence of violent crime in each block group-month by year, 2013–2014, DeKalb County, Georgia.

Note: Legend: light grey (0 violent crimes); medium grey (1–3 violent crimes); black (4 or more violent crimes).

The 279 Census block groups are displayed along the x-axis. Each unit on the y-axis represents a one-month time period.

Table 1.

Comparing the performance of models using only demographic, business or historical crime data to predict the occurrence of four or more violent crimes in a block group-month.

	Positive predictive value (%)	Negative predictive value (%)	Sensitivity (%)	Specificity (%)	F1 score[a]
Demographic data alone	27.7	98.6	71.5	91.6	33.2
Business data alone	35.6	96.9	32.6	97.3	34.0
Historical crime data alone	50.4	97.5	43.1	98.1	46.5

Note: This table is designed to compare the performance of each data source alone. As in Table 2, the demographic data includes 20 variables from the American Community Survey. The business data includes the yearly counts of seven different business types. The historical crime data includes 24 months of lagged variables on both violent crimes and non-violent property crimes

[a] The F1 score is a measure of a test’s accuracy which uses the positive predictive value and sensitivity as inputs. The F1 score is more appropriate for certain applications of predictive modelling than the area under the receiver operating characteristic curve (which utilizes sensitivity and specificity as inputs).

Table 2.

Performance of models predicting the occurrence of four or more violent crimes in a block group-month.

	Positive predictive value (%)	Negative predictive value (%)	Sensitivity (%)	Specificity (%)
Random	5.3	95.7	5.6	95.6
Model 1	35.3	97.0	34.0	97.2
Model 2	37.0	97.1	34.7	97.3
Model 3	45.6	97.2	36.1	98.1
Model 4	50.4	97.5	43.1	98.1
Model 5	50.8	97.5	43.8	98.1
Model 6	52.1	97.5	43.4	98.2

Note: Model descriptions: Random – random allocation of the same number of block-group months with violent crime as in the previous year; Model 1 – uses the presence of four or more violent crimes (coded as a binary variable) in the block group in the preceding month as the sole predictor; Model 2 – uses 12 months of lagged variables on the total number of all violent crimes in each block group; Model 3 – uses 24 months of lagged variables on the total number of all violence crimes in each block group; Model 4 – adds 24 months of lagged variables on non-violent property crimes to Model 3 (burglary, theft from vehicle and vehicle theft); Model 5 – adds 20 block group demographics variables from the American Community Survey to Model 4; Model 6 – adds the count of seven different business types to Model 5.