HHS Public Access

Author manuscript

Comput Environ Urban Syst. Author manuscript; available in PMC 2022 May 01.

Published in final edited form as:

Comput Environ Urban Syst. 2021 May ; 87: . doi:10.1016/j.compenvurbsys.2021.101599.

Incorporating space and time into random forest models for analyzing geospatial patterns of drug-related crime incidents in a major U.S. metropolitan area

Zhiyue Xia[1], **Kathleen Stewart**[1], **Junchuan Fan**[2]

[1]Center for Geospatial Information Science, Department of Geographical Sciences, University of Maryland, College Park 20742, MD, USA

[2]Oak Ridge National Laboratory, Oak Ridge, Tennessee

Abstract

The opioid crisis has hit American cities hard, and research on spatial and temporal patterns of drug-related activities including detecting and predicting clusters of crime incidents involving particular types of drugs is useful for distinguishing hot zones where drugs are present that in turn can further provide a basis for assessing and providing related treatment services. In this study, we investigated spatiotemporal patterns of more than 52,000 reported incidents of drug-related crime at block group granularity in Chicago, IL between 2016 and 2019. We applied a space-time analysis framework and machine learning approaches to build a model using training data that identified whether certain locations and built environment and sociodemographic factors were correlated with drug-related crime incident patterns, and establish the top contributing factors that underlaid the trends. Space and time, together with multiple driving factors, were incorporated into a random forest model to analyze these changing patterns. We accommodated both spatial and temporal autocorrelation in the model learning process to assist with capturing the changes over time and tested the capabilities of the space-time random forest model by predicting drug-related activity hot zones. We focused particularly on crime incidents that involved heroin and synthetic drugs as these have been key drug types that have highly impacted cities during the opioid crisis in the U.S.

Keywords

random forest; machine learning; opioid crisis; heroin; synthetic drugs; spatiotemporal modeling

All correspondence should be addressed to: Zhiyue Xia, Center for Geospatial Information Science, Department of Geographical Sciences, University of Maryland, College Park, MD USA 20742, zyxia@umd.edu.

Author Statement:

N.A.

Competing interests

The authors declare that they have no competing interests.

1. Introduction

The opioid crisis spread across the United States beginning with the west coast, and then expanding to heavily impact the central, mid-Atlantic, and east coast of the U.S. as well as states in the southeast (Ciccarone, 2017; Dasgupta et al., 2017). The estimated number of adolescent and adult illicit drug users increased from 27 million in 2014 to 53.2 million in 2018 (Center for Behavioral Health Statistics and Quality, 2015; Substance Abuse and Mental Health Services Administration, 2019). Opioids are a category of drugs that include heroin as well as synthetic drugs such as fentanyl (National Institute on Drug Abuse, 2019a). The crisis was reflected by the increasing numbers of overdoses involving heroin, that expanded spatially from the west coast of the U.S. in the early 2000s to reach Appalachia, the Great Lakes and Ohio Valley, and New England by 2016 (Hedegaard et al., 2018; Jalal et al., 2018; Stewart et al., 2017). The number of heroin users reached a peak in 2016 with an estimated 948,000 users or 0.4 percent of the U.S. population at that time (Substance Abuse and Mental Health Services Administration, 2019). Fentanyl, a potent synthetic opioid, has also become a threat nationally (Center for Disease Control and Prevention, 2019a; Marshall et al., 2017; Spencer et al., 2019) as synthetic opioids have been increasingly identified as a leading factor in overdose deaths in at least 23 states since 2015 (Center for Disease Control and Prevention, 2019b; National Institute on Drug Abuse, 2019b).

Research on the spatial and temporal patterns of drug activities and detecting clusters of particular drugs is useful for distinguishing hot zones of those drugs that in turn provides a basis for assessing the availability of substance use treatment facilities in these locations (Abraham et al., 2018; Yarbrough et al., 2019). For this study, the research objective was to identify factors related to built environment and sociodemographic characteristics of communities that were related to the changing patterns of drug activities involving particular drugs in a city, and to build a space-time random forest model, that could shed light on the importance of the different factors, and model these changing patterns over space and time so that mitigation steps might be taken. We investigated the spatiotemporal patterns of drug-related crimes using narcotic crime data as a proxy for locations where drugs were likely to present. We analyzed the spatial patterns of more than 52,000 reported incidents of drug-related crime, including over 16,000 heroin and synthetic drug-related crimes at block group granularity in Chicago, IL between 2016 and 2019. Chicago is a major metropolitan city in the U.S. with a population of over 2.7 million in 2018. We designed a space-time random forest model that accounted for spatiotemporal autocorrelation in the patterns of drug-related crimes in order to identify a set of possible underlying drivers for these patterns and track what changes have occurred. These drivers included specific types of locations (e.g., vacant buildings and alleys), additional built environment factors (e.g., road network density and street intersection density), and sociodemographic factors (e.g., level of education, income, and percentage of owner occupancy in a neighborhood). Here, built environment refers to the properties of human-modified places such as streets, buildings, parks and transportation systems, and characterizes city structure (United States Environmental Protection Agency, 2019). Public health problems including mental health and substance use disorder have been shown to be related to built environment factors (Cerdá et al., 2013; Srinivasan et al., 2003). Sociodemographic variables such as education, income and household status have also been

found to be correlated with drug activities (Lipton et al., 2013; Nechuta et al., 2018; Vilalta, 2010). The space-time random forest model was used to identify built environment and sociodemographic factors that were most associated with the drug activity locations. Spatial and temporal autocorrelation were accommodated in the model learning process to capture the changing trend of drug activity patterns over successive time periods.

2. Related works

Crime dataset records including the location and date-time of drug crime incidents for delivery, possession, and the manufacture of drugs, have been intensively used for investigating geographic patterns of drug activities. In a previous study, researchers analyzed crime data as well as recorded calls for police service to examine the relationships between drug activities, social disorder and crime, with results that showed significant spatial links between them (Weisburd and Mazerolle, 2000). Another study investigated the geographical relations between alcohol outlets, drug markets and violence using arrest data on drug possession and trafficking to map estimated drug markets in Boston (Lipton et al., 2013). In Mexico, an analysis of the spatial dynamics of drug arrests related to marijuana and cocaine found sociodemographic factors such as college education, housing conditions, and female-headed households were positively correlated with arrests involving marijuana, but no sociodemographic correlates were significantly established for cocaine patterns (Vilalta, 2010).

Space-time modeling of changing patterns of crime in cities, for example, homicides, burglaries, drug use and gun violations has studied by researchers (Hodgkinson and Andresen, 2019; Mohler, 2014; Piza and Carter, 2018; Shiode et al., 2015; Zhao and Tang, 2017). Existing studies of geographical research on crime patterns have included among other topics, forecasting crime hotspots based on historical crime data, and finding correlations between crime patterns and surrounding environmental characteristics. Recent studies, for example, have used network-based models to detect and forecast street-level crime hotspots by using historical crime data (Shiode and Shiode, 2020; Y. Zhang and Cheng, 2020). While other research investigated correlations between crime patterns and environmental variables extracted from multiple-source data such as social media data, remote sensing imagery and Google Street View (He et al., 2017; Vomfell et al., 2018; Yang et al., 2020). Machine learning methods have been used extensively for predicting crime hot zones, although few studies have combined spatiotemporal patterns of historical crime data together with multiple underlying driving factors in machine learning models.

Hotspot analysis has assisted in identifying where events are densely concentrated, and has been used as a basis for predicting spatial patterns of repeated events in urban environments (Albright et al., 2019; Chainey et al., 2008). For example, a previous study used hotspot analysis to examine the homicide patterns in Chicago from 1960 to 1995 (Ye and Wu, 2011). Prior studies investigated spatial patterns of drug activities and understand any associations with surrounding built environment characteristics (Cerdá et al., 2013; Chaney and Rojas-Guyler, 2015; Darke et al., 2001). Drug-related crime data that records crime incidents including drug transactions, delivery and possession can be analyzed to provide insights on locations where drug use may be present. In this paper, we have incorporated space and time

into a random forest model to analyze drug-related crime incidents and their underlying factors over time. An earlier study described how the Chicago Police Department used a 'heat list' that included approximately 400 individuals who were forecast to be potentially involved in crime (Lum and Isaac, 2016). They pointed out that police-recorded data was biased and could be attributed to the police's determination of where to patrol and search. To examine drug-related crime patterns, and to gain further insights about locations of drug-related activities, we compared the spatial patterns of deaths involving opioids and drug-related crimes between 2016 and 2019, and based on this analysis, investigated the relationships between drug activities in Chicago, a major metropolitan city in the U.S.

For this study, we designed a space-time random forest model that ingested data from multiple geospatial data sources to investigate drug activity patterns and their associated spatiotemporal characteristics in an urban setting. We accommodated both spatial and temporal autocorrelation in the model learning process to assist with capturing the changes over time and tested the model's ability to capture trending changes by predicting future drug incident locations. Random forest, a machine learning model, combines a large number of decision trees, using selected features to split tree nodes and repeated sampling of different subsets of observations to train the models (Breiman, 2001). The output for random forest regression is the mean value of all regression trees, while the majority decision of all classification trees is the output for random forest classification tasks. Random forest model, with the capability of processing massive datasets and handling multicollinear relationships within multi-source datasets, has been used for drug-related public health research (Ancuceanu et al., 2019; Fernández-Delgado et al., 2014; Kamel Boulos et al., 2019). For example, a random forest model was used to detect individuals with substance use disorder based on a set of behavior and health characteristics (Jing et al., 2020). In previous research, random forest models were often implemented without using spatial and temporal relations between model variables. Recent studies have begun to address spatial dependencies into random forest models. For example, a recent study proposed a geographical random forest model, attempting to include spatial heterogeneity in a random forest model by disaggregating a global model into several local sub-models (Georganos et al., 2019) to understand local variations.

3. Data

3.1 Drug-related crime in Chicago

In this study, a public safety dataset provided by the Chicago Data Portal[1] was used for accessing drug-related crime data in Chicago between 2016 and 2019. Chicago Data Portal extracted the public safety dataset from the Chicago Police Department's Citizen Law Enforcement Analysis and Reporting (CLEAR) system. For the reported incidents, the police recorded the type of crime such as possession, delivery and manufacture of drugs and also described the drug types. In order to discover the spatial patterns of heroin and synthetic drug-related crimes, the incidents were categorized by drug types. With the increased

[1]Chicago Data Portal: https://data.cityofchicago.org/

presence of fentanyl in U.S. cities since 2016, we assumed that the category of synthetic drugs would likely include incidents involving fentanyl (Hedegaard et al., 2019).

To capture community-level characteristics in the model, we aggregated all the variables including the frequency of drug-related crime incidents, sociodemographic and built environment factors at U.S. Census block group level. Block group is the smallest geographic unit used by the U.S. Bureau of Census to publish sample data from the American Community Survey (ACS), and we selected this as the unit of analysis for this study. To model communities by block group, 5-year ACS estimates were used to help smooth any uncertainty created by using this fine level of spatial granularity. Using block groups helped us to account for varying community characteristics and reveal patterns at this scale. We defined the drug-related crime rate as the frequency of drug crimes during a time period aggregated by block group units and adjusted by block group population. Chicago has 2210 block groups. Typically, a block group in Chicago has a population of 800–10,000 individuals. It should be noted that there were eight block groups with no associated population counts including, for example, O'Hare International Airport and Chicago Midway International Airport. As our analysis examined built environment and sociodemographic characteristics of locations where populations were living and working, these eight block groups were not included in the model dataset. A complete list of factors we have collected in this study as well as literature that have previously established a link between drug activities and these factors are presented in Table A.1 in Appendix A.

Between 2016 and 2019, drug-related crimes occurred across most of downtown Chicago (Figure 1). For this period, 52,567 drug-related crime incidents occurred in 1,901 block groups out of 2,210 block groups in total. The total number of drug-related crime incidents (across all categories) decreased by 12% in 2017, and then increased by 21% in 2019 (Table 1). The statistics showed that the number of heroin-related incidents was quite consistent from 2016 to 2017 and then increased by 18% over 2018 and 2019. For the same period, incidents involving synthetic drugs consistently increased, with a noticeably high growth rate, 94%, from 2016 to 2019.

3.2 Deaths involving opioids in Chicago

To determine the degree to which the spatial patterns of drug-related crime incidents were associated with locations where drug activities frequently occurred, and illustrate that drug-related crimes were not necessarily being driven by other crime-related aspects (e.g., policing strategies), we analyzed the spatial patterns of opioid-involved mortality and compared these patterns to the reported locations of drug-related crimes. Data on opioid-involved mortality between 2016 and 2019 for Chicago were extracted from the Medical Examiner Case Archive dataset accessed from Cook County Open Data[2]. The Cook County, IL Medical Examiner's Office recorded the time and location of deaths, and determined causes and manners of death cases under its jurisdiction. Deaths involving all categories of opioids in Chicago consistently increased between 2016 and 2019 (Table 2). The number of heroin-involved deaths reached a peak in 2017. During the four-year period, deaths

[2]Cook County Open Data: https://datacatalog.cookcountyil.gov/

involving fentanyl increased in the Chicago area by 69%. A noticeable increase in synthetic drug-related crime incidents was also noted for this same period.

3.3 Built environment and sociodemographic data

We collected built environment data for the study region from the United States Environmental Protection Agency (EPA) Smart Location Database released in 2013[3]. This dataset summarized indicators associated with urban design, location efficiency, transit and demographics, including, for example, street intersections per square mile, road network density and walkability scores. The complete list of variables in the EPA Smart Location Dataset are accessible in their user guide[4]. We also collected supplementary sociodemographic variables from U.S. Census at the granularity of block group including educational factors e.g., percentage of population with college-level education or higher, economic factors e.g., percentage of occupied houses, and social factors, e.g., the percentage of full-time employment. The American Community Survey (ACS) 5-year estimates dataset at block group for 2017 was used for accessing and calculating these sociodemographic variables. In total, 104 candidate factors were collected, including 61 built environment factors and 43 sociodemographic factors (Table A.1 in Appendix A).

3.4 Spatial data

The reported crimes dataset for Chicago included descriptions of locations where each incident occurred. The frequencies of all locations for drug-related crime reports between 2016 and 2019 were analyzed and the most frequent locations for reported heroin and synthetic drug-involved crimes were computed. Based on this analysis, seven geospatial locations were identified as *key locations*. These included gas stations, vacant lots, abandoned buildings, parking lots, alleys, parks, and high schools. Data layers for sites relevant to drug-related crimes including data on vacant lots and abandoned buildings, and a boundary map of park property were collected from the Chicago Data Portal. Locations of gas stations were obtained from the ESRI Business Analyst[5] dataset. Street alley networks were sourced from Chicago WBEZ[6]; public high school locations were obtained from the Chicago Data Portal, and private high school locations from the Cook County Open Data[7]. We calculated spatial variables from these layers and aggregated the values by U.S. census block group.

4. Geospatial patterns of drug-related crime incidents and opioid-involved deaths

Clustering analysis was implemented to determine the spatial patterns of all drug-related crimes, crimes by individual drug type (i.e., heroin and synthetic drugs), and for opioid-involved deaths. Global Moran's I statistic was used to measure the spatial autocorrelations

[3]Smart Location Mapping: https://www.epa.gov/smartgrowth/smart-location-mapping
[4]EPA Smart Location user guide: https://www.epa.gov/smartgrowth/smart-location-mapping
[5]ESRI Business Analyst https://www.esri.com/en-us/arcgis/products/arcgis-business-analyst/
[6]Chicago WBEZ https://www.wbez.org/
[7]Cook County Open Data https://datacatalog.cookcountyil.gov/

for both the drug-related crimes and the opioid-involved deaths (Li et al., 2007; Moran, 1950).

The spatial patterns of both the reported drug-related crime incidents and opioid-involved-deaths were significantly clustered, as illustrated by positive Global Moran's I values of 0.424 and 0.309, respectively, and p-values smaller than 0.001. We examined the correlation between the drug-related crime rates and the number of opioid-involved deaths (adjusted by block group population). The Pearson's correlation coefficient between these two values was 0.61 with the p-value smaller than 0.001. Based on these results, during this period of study, the spatial distribution of drug-related crime incidents and opioid-involved deaths showed a significant positive relation suggesting that the locations of drug-related crimes were indicative of where drug use was also happening in the city.

Anselin's Local Moran's I statistic was developed by Anselin (1995) and was commonly used to implement clustering analysis (Liu and Wang, 2017; Neutens et al., 2013). We used Anselin's Local Moran's I statistic to identify changing patterns of heroin and synthetic drug-related crime incidents over time Block groups that were in a significant high-high value cluster were then identified as hotspot block groups. In 2016 and 2017, two major clusters were found for reported crimes involving heroin (Figure 3a). Hotspot 1 (upper hotspot) was consistent throughout 2016 and 2019 (Figure 3a). Heroin-related crime incidents that occurred in hotspot 1 increased by 22% during the four-year period. By contrast, hotspot 2 (lower hotspot), gradually diminished during the four years and in 2019, impacted only two block groups (Figure 3a). During the four-year period, the heroin hotspots diminished over time and the synthetic drug hotspots took over in those block groups. The clusters of synthetic drug-related crimes changed from a scattered pattern in 2016 to being concentrated in two distinct hotspots by the end of 2019 (Figure 3b). These two distinct synthetic drug hotspots were in similar locations to the heroin hotspots of 2016 and 2017, but included a higher number of block groups.

5. Building a space-time random forest model

5.1 Incorporating space and time into a random forest model

To build a random forest model that was capable of capturing the changing patterns of heroin and synthetic-drug related crime incidents over time, spatial and temporal lag variables were added to the model in order to detect the spatial dependencies on neighboring block groups, and the relationships between successive time periods.

Temporal patterns of the reported crimes were used to guide the selection of time periods over which to detect spatial change. To select the temporal granularity for analyzing heroin-involved crime incidents, we aggregated monthly incidents by census block group and used clustering to determine the hotspot block groups. A time series of heroin-related incidents was created based on the monthly count of hotspot block groups. To smooth out short-term fluctuations and distinguish the overall trend, a moving average technique was applied to the heroin crime hotspot block group time series data (Cryer and Chan, 2008). We selected the largest time interval between highest and lowest values in the time series – five months – as the temporal length to process the spatial pattern of heroin-related crime. Using this interval

helped us to reliably detect the changing pattern of incidents, even as the pattern fluctuated. For synthetic drug incidents, the number of incidents by block group was much smaller, as there were less than 50 monthly synthetic drug-related incidents in all 2202 block groups during the study period. This number was too small to implement clustering analysis and therefore, we created a time series based on the raw count of synthetic drug-related incidents instead of by hotspot block groups. In this time series, the longest time interval between a highest and lowest number of incidents was nine months, and that was used as the temporal granularity to identify the changing patterns for synthetic drug activities.

We trained the random forest model using different combinations of factors including sociodemographic factors and built environment factors, as well as three variables that related to change over space and time. This included two lag variables (a time-lagged variable and a spatiotemporally lagged variable) and a trend variable. The time-lagged variable was a binary variable identified based on whether a given block group was in a hot spot of drug-related crime during the previous time period. The spatiotemporally lagged variable referred to the count of block groups in queen adjacent neighborhoods that belonged to a hotspot during the previous time period. The trend variable was computed based on the time intervals calculated to capture the changing patterns of clusters (Chi and Zhu, 2008), and tracked any increases or decreases in the numbers of neighboring block groups that belonged to a hotspot during the previous time period. As a final step, model prediction accuracy (i.e., percentage of block groups correctly predicted) was computed.

5.2 Including key locations of drug-related crime incidents in the model

To analyze locations for drug-related crime incidents and develop a foundation for building a machine-learning classifier, we summarized the most frequent locations for both heroin and synthetic drug-related crime incidents. According to the crime data provided by the Chicago Data Portal, there were 180 different location descriptions in total. While numerous types of locations were recorded, not all of these locations were useful for our analysis. Sidewalks and streets, for example, were the two most frequent locations for drug-related crime however, these are ubiquitous features in a city and so we did not utilize either of these features in our model.

Similarly, specific locations at a residence (e.g., porch or yard), and vehicles were also recorded as frequent locations for drug-related crimes, however, these were also not used for this research. Instead, we used neighborhood features that were more uniquely identifiable including alleys, vacant buildings, vacant lots, parking lots, gas stations, parks, and high schools. To incorporate these key locations in the model, we calculated seven variables from key location layers, including the counts of gas stations, vacant lots, vacant buildings, and high schools, as well as alley density and area of park properties in each block group.

5.3 Model training, validating and out-of-sample testing

A random forest model was used to gain insights into the contributions of built environment and sociodemographic factors in classifying hotspots of both heroin and synthetic drug-related crimes and to capture the changing patterns of drug-related activities over space and time. The modeling was performed in R, an open-source ecosystem that provides multiple

packages for machine learning, and used R packages ‘ranger’, ‘RRF’, ‘pdp’ and ‘rgdal’ (Bivand et al., 2008; Deng and Runger, 2013; Diggle, 2013; Wright and Ziegler, 2017).

Built environment and sociodemographic factors were used as input variables in the random forest classifier to identify whether a block group belonged to a hotspot or not. Bagging, also called bootstrap aggregating, is an ensemble algorithm that was used for selecting samples for each tree in the random forest. Typically, each bootstrap sample subsets approximately 63% of the training data while leaving out about 37% of the data (Breiman, 1996). The out-of-bag (OOB) error rate is the misclassification error rate of the estimates fitting the trees whose bootstrap samples do not have this observation. In this study, the OOB error rate was used to evaluate the fitting accuracy of the random forest model.

To test the model’s ability of predicting future spatial locations of heroin and synthetic drugs, an out-of-sample test using an independent test dataset was implemented to evaluate the model forecasting performance. We trained the space-time random forest model to associate the spatiotemporal patterns of drug-related crime with potential predictor variables using the data for the first three years (2016 ~ 2018). The drug-related crime data was aggregated by the computed temporal intervals (5 months for heroin-related crime and 9 months for synthetic drug-related crime) to generate spatial patterns. The training model used these aggregated patterns as output, and used one month as an offset during the training process.

Heroin data for the last five months of 2019 and synthetic drug-related crime data for the last nine months of 2019 were used as the independent test datasets in the out-of-sample test. Predictive accuracy (ACC) was used to evaluate the model’s ability to forecast future drug hotspots. ACC has been widely used to assess the predictive performance of machine learning classifiers (Chen et al., 2017; He and Garcia, 2009) and is calculated as:

$$ACC = \frac{TP + TN}{TP + FP + TN + PN} \quad \#(1)$$

where TP (true positive) is the number of actual hotspot block groups that are correctly predicted; TN (true negative) is the number of actual non-hotspot block groups that are correctly predicted; FP (false positive) and FN (false negative) are the numbers of block groups that are incorrectly predicted.

Following the approach by Chainey et al. (2008), we computed a prediction accuracy index (PAI) that returned the density of drug-related crimes in the predicted block groups as compared to the density of drug-related crimes over the whole study area. We also computed a prediction efficiency index (PEI) that measured the ratio of PAI and maximum possible PAI that a model can achieve (Hunt, 2016). PAI and PEI were used to evaluate both the effectiveness and efficiency of the hotspot prediction models.

5.4 Feature selection: guided regularized random forest (GRRF)

For this research, a guided regularized random forest (GRRF) was used for feature selection (Deng and Runger, 2013). GRRF models use a variable importance score calculated from a preliminary random forest model to guide feature selection in the regularized random forest.

Regularization aims at selecting high-quality feature subsets to avoid overfitting problems by penalizing inputting new features into the model (Deng and Runger, 2012). A feature with a higher importance score in the preliminary random forest is penalized less in GRRF. GRRF selects a compact feature subset and reduces redundancy among the selected features.

For this research, we collected 104 candidate factors in total, including 61 built environment factors and 43 sociodemographic factors. GRRF was used to select two compact variable subsets for heroin model and synthetic drug model separately. For heroin-related crime hotspot classification, input variables selected by GRRF included 20 built environment factors and 26 sociodemographic factors, as well as three spatial and temporal lag variables (Table A.2 in Appendix A). For synthetic drug-related crime hotspot classification, selected factors included 17 built environment factors and 23 sociodemographic factors. The three spatiotemporal lag variables were also included (Table A.3 in Appendix A).

5.5 Model optimization: handling class imbalances, grid search for model hyperparameters and weighting key location features at splitting nodes

To handle any class imbalance problems that arose due to the presence of spatial clusters or hotspots, two techniques, oversampling minority classes and underdamping majority classes (i.e. using bagging to repeatedly generate random subsets for each decision tree) have been recognized by researchers as efficient resampling methods (Japkowicz and Stephen, 2002; Kotsiantis et al., 2006; Liu et al., 2009). An imbalanced class problem existed in our synthetic drug training dataset, for example, for the first time period, January through September 2016, 2164 block groups belonged to the majority class (i.e., block groups that were not in a hotspot) while only 39 block groups belonged to the minority class (i.e., block groups belonging to a hotspot). We tested oversampling the minority class at multiple ratios of (α = 100%, 200%, 300%, 400% and 500%) and repeated underdamping the majority class at ratios of β (β = 50%, 25%, 12.5%, 6.25%, and 3.13%) during the bootstrapping process. We compared the classification accuracy of these 25 (5 × 5) combinations and found that when α=200% and β = 25%, the model had the best prediction performance. For the heroin training dataset, the class imbalance problem was present but not quite as strong as with synthetic drug case. Resampling the heroin training dataset did not help to improve model performance in terms of prediction accuracy, so neither underdamping nor oversampling was implemented for the heroin training dataset.

The random forest model uses hyperparameters that need to be set, including numbers of drawn candidate variables at each split (*mtry*), the number of trees (*ntree*), the sample size of observations for each tree, and the minimum number of samples for each node (Probst et al., 2019). Two hyperparameters, *mtry* and *ntree* have been shown to influence prediction performance (Bernard et al., 2009; Biau and Scornet, 2015). Tuning hyperparameters can improve the accuracy of the random forest model to some extent. A model tuning method, grid search, was used for searching for the best combination of two hyperparameters, i.e., the number of trees (*ntree*) and numbers of drawn candidate variables at each split (*mtry*) in our random forest model (Probst et al., 2019). The typical default setting in a random forest classifier for these two hyperparameters is *ntree* = 500 or 1000, and $mtry = \sqrt{nv}$ where nv is the number of input variables. In our model, 49 variables were used for predicting heroin

hotspots and 43 variables were used for predicting synthetic drug hotspots. The typical default setting for our classifier was *ntree* = 500 and *mtry* = 7, however this default setting might not always be the best setting. In our models, with a given *mtry*, the error rate of the misclassification became stable when *ntree* reached approximately 200. Therefore, we used the tuning grid for *ntree* from 200 to 1000 with step = 100 and *mtry* ranging from 3 to 14 with step = 1, then, set the grid search process with 5-fold cross-validation and 3 repeats per fold. We found that the best settings of hyperparameters for classifying the heroin-related crime hotspots was when using *mtry* = 8 and *ntree* = 500, and for synthetic drug hotspots when *mtry* = 14 and *ntree* = 400.

Weighting the features that were known to be more informative for detecting drug-related crime hotspots (i.e., predicting dependent variables) more than other variables can increase the contribution of these features and is likely to improve the classification accuracy (Amaratunga et al., 2008; Ma et al., 2011; Malley et al., 2012). In this study, the key locations that were identified from the drug crime records were considered to be more informative than other built environment or sociodemographic variables. Given this, we set higher weights for the key location variables to have a higher probability of these variables being selected in the splitting nodes.

5.6 Contributions of variables in the model

In order to understand the relationships between the different factors and the reported drug incident patterns, we analyzed variable importance that revealed the contributions of different factors to identify heroin and synthetic drug-related crime hotspots. To interpret each factor's contribution for classification, we used corrected impurity importance, also namely actual impurity reduction (AIR), to measure variable importance (Nembrini et al., 2018). Corrected impurity importance measures the improvement of the classification rate at splitting nodes without bias (Calle and Urrea, 2011; Nembrini et al., 2018; Strobl et al., 2007). We also constructed partial dependence plots (PDP) for the spatial and temporal lag variables (Friedman, 2001; Greenwell, 2017). PDP used a partial dependence function to measure the marginal effect of the variables on the classification outcome (Greenwell, 2017; Molnar, 2019). This process provided insights into the relationships of drug-related activity patterns between successive time periods.

6. Applying the space-time random forest model to drug-related crimes

6.1 Model training and contribution of variables

To train the random forest classifier to learn the changing trends of drug-related crime patterns and reinforce the role of spatiotemporal autocorrelation underlying any of the changes, we fitted and validated the model using the data between 2016 and 2018, and evaluated the model performance in terms of the misclassification rate based on out-of-bag (OOB) error. The final time periods of 2019 ((i.e., August-December 2019 for heroin, April-December 2019 for synthetic drugs) were used to perform an independent test of the model's prediction performance. For classifying heroin crime hotspot clusters, the OOB error 2.41% showing that it correctly classified 97.59 % of block groups during the training

process. For synthetic drug hotspots, the OOB error was 4.52% indicating that the classifier correctly classified 95.48% block groups.

In order to measure the contribution of each variable for classifying the hotspots, we used actual impurity reduction (AIR) to evaluate the variable importance. Variables used for classifying heroin crime hotspots (Figure 4a) and synthetic hotspots (Figure 4b) were ranked by their importance scores. For classifying heroin hotspots, the top two important variables were key location variables, vacant lots (*VacLot*) and vacant buildings (*VacBldg*). The next two high ranking variables were sociodemographic variables, percentage of low-wage workers (*R_PctLowWage*) and percentage of the population with Bachelor's degree or higher (*PBachelorHigher*). For classifying synthetic drug-related hotspots, high ranking variables were similar to those for heroin but the rank varied. The top two variables were *VacLot* and *VacBldg* (Figure 4b). A different set of sociodemographic factors (e.g. race and ethnicity, median household income, percentage of employment) were also correlated with synthetic drug-related crime locations.

We selected four built environment and sociodemographic factors that were in the top five both for classifying heroin hotspots and synthetic drug hotspots (Figure 5). These four variables were visualized by the boxplots for heroin-related crime hotspots, synthetic drug-related crime hotspots and other block groups that belonged to neither a heroin nor a synthetic drug hot spot. There were more vacant buildings and vacant lots in heroin and synthetic drug crime hotspots than in non-hotspot block groups (Figure 5a and 5b). The percentage of low-wage workers was higher in heroin and synthetic drug hotspots than in other areas (Figure 5c). Median *PBachelorHigher* in the other block groups was 19.4% that was much higher than either heroin (5.6%) or synthetic drug hotspots (6.6%) (Figure 5d).

To understand the relationships of drug crime patterns between successive periods $t-1$ and t, partial dependence plots (PDP) were used to analyze the effect of one or two input features (i.e., spatial and temporal lag variables at *t-1*) on the prediction (i.e., the probability of a block group being classified as a hotspot block group at *t*). The length of a time period was calculated from the previous time series analysis. For example, in the heroin model, when time period *t-1* corresponded to January-May 2016, time period *t* corresponded to Feb-June 2016. For predicting heroin hotspots, the partial dependence plot of *nb_t_1* (number of neighboring hotspot block groups during *t-1* period) showed that being surrounded by more hotpot block groups resulted in a higher probability of a given block being classified as a hotspot block group in a successive time period *t* (Figure 6 a). When a given block group at *t-1* was surrounded by a low number of hotspot block groups, this block group was more likely to maintain its status in the successive time period (Figure 6 a, b). For predicting synthetic drug hotspots, a given block group tended to become the same status as its surrounding block groups (Figure 6 c, d).

6.2 Predicting heroin and synthetic drug hotspots

To analyze the predictive power of built environment and sociodemographic factors, and also the spatiotemporal variables in the space-time random forest model, we built seven models using different combinations of categories of variables as model input. We trained the model using the data between 2016 and 2018. As stated above, heroin data for the last five months

and synthetic drug-related crime data for the last nine months of 2019 were used as independent test datasets to evaluate out-of-sample forecasting accuracy. For this out-of-sample testing, we compared the predicted hotspots with the actual hotspots calculated from the 2019 drug crime data.

To test the model's ability to predict heroin hotspots, built environment and sociodemographic factors assisted in locating the potential hotspots, while spatial and temporal lag variables enabled the classifiers to learn the changing spatial pattern of hotspots (Figure 7). Predicting heroin hotspots using only sociodemographic variables (Figure 7a) or built environment variables (Figure 7b) generated overprediction in the lower hotspot because the model learned information only from previous hotspots but did not capture the disappearing pattern, while only using spatial and temporal lag variables in the model (Figure 7 f) resulted in underestimating the size of the upper hotspot. Inputting spatiotemporal autocorrelation variables into the model improved the prediction effectiveness of the model as illustrated by the increase in PAI (Table 3), but the variations on prediction efficiency (PEI) of seven models were minimal. The model using the three types of variables (Figure 7g) was able to correctly predict 81.4% heroin hotspot block groups and 99.3% non-hotspot block groups, as the overall prediction accuracy (ACC) was 98.3% (Table A.4 in Appendix A).

For predicting synthetic drug hotspots, incorporating spatial and temporal variables enabled the random forest classifier to capture the expanding hotspots (Figure 8), while only using spatiotemporal autocorrelation in the model resulted in an overprediction of the size of the hotspot (Figure 8f). The built environment (Figure 8a) and sociodemographic factors (Figure 8b) both played important roles for predicting synthetic drug hotspots but neither of them enabled the model to learn the missing pattern of the small hotspot in the southeastern side. The best model (Figure 8g) for predicting synthetic drug-related hotspots successfully identified which block groups were not likely to become a hotspot block group (93.0 % correctly identified), but might underestimate the size of actual hotspots (60.0% correctly identified). The best model with the highest ACC, PAI, and PEI for predicting synthetic drug hotspots, 90.7%, used a combination of all three types of variables (Table 3 and Table A.4 in Appendix A).

7. Discussion

In this research, a space-time random forest model was designed to investigate changing patterns of reported crime incidents involving different drugs in a major metropolitan city (Chicago) over time. One of the outcomes of this study was that we identified a set of factors useful for identifying heroin and synthetic drug hotspots in a city. We found that both heroin and synthetic drug hotspots were more likely to appear in the neighborhoods where there were more vacant buildings and vacant lots comparing to other non-hotspot block groups, while in synthetic drug-related hotspots, there were relatively more vacant buildings and less vacant lots than in heroin-related hotspots. While the characteristics of heroin and synthetic drug-related crime hotspots shared some similarities, we also investigated the possibility of an incident involving both heroin and synthetic drugs, but we found only a few (less than 0.1%) incidents were reported to involve both drugs. We also found that some

sociodemographic factors were strong indicators for two types of drug hot spots but the rank of the factors varied. Overall, low income and education were the top two sociodemographic factors for predicting heroin hotspots, while low employment was also an important correlate with synthetic drug-related hotspots. Some of these top factors have also been found to be significantly related to drug activities in previous research. For example, research has established significant sociodemographic and built environment correlates for drug activities, typically including low education levels (McCord and Ratcliffe, 2007), low income (Vilalta, 2010) and high unemployment rate (Cerdá et al., 2013).

Incorporating spatial and temporal relationships into a random forest model enabled the model to learn the changing trends of the spatial patterns. In our research, incorporating space and time allowed the model to capture the concentration of block groups over time (heroin) or the increase in block groups (synthetic drugs) in Chicago during the study time period. We also compared our model results with a previous study that used a random forest model approach to predict crime hotspots (Borges et al., 2017). They used urban features and time-varying features such as timestamps of crime occurrences in a random forest model. We found that our approach based on incorporating spatiotemporal autocorrelation, improved the random forest model's performance in terms of prediction accuracy and efficiency by enabling the model to capturing the changing patterns of drug activities.

For this research, the drug-related crime data accessed from Chicago Data Portal were extracted from the Chicago Police Department's Citizen Law Enforcement Analysis and Reporting (CLEAR) system. We do not have information on policing practices during the study time period nor do we know to what degree the drug crime locations may be influenced by these practices. However, our analysis of opioid-involved mortality between 2016 and 2019 showed a positive association between the locations of opioid-involved deaths and drug-related crimes. Understanding, modeling and predicting the changing patterns of drug activities will further assist in locating substance use treatments and counseling services. It should also be acknowledged that drug-related crimes might be underreported (Mosher et al., 2010). For example, Mosher et al. (2010) indicated that high school administrators might be pressured to underreport or not report school crimes and minority groups had a greater tendency to underreport crime behaviors. Using crime-related data may miss additional locations where drugs were present (but no crime incidents were reported).

8. Conclusions

Our approach highlights a promising direction of using machine learning together with spatiotemporal autocorrelation to analyze changing patterns of repeated events – in this case, drug-related incidents – in cities. Incorporating space and time into a machine learning model assists in making more accurate forecasts of changing patterns. Future work could consider combining spatiotemporal heterogeneity with machine learning models by adding spatiotemporal bandwidths during the modeling process to investigate the patterns at varying spatial and temporal scales. In this study, a random forest model was used for a space-time analysis framework. Future research could examine other machine learning approaches such as gradient boosting model to test spatiotemporal approaches using other tools. This study

also applied the space-time analysis framework and machine learning technologies to investigate the spatiotemporal patterns at block group level of more than 52,000 drug-related crime incidents, including more than 16,000 incidents involving heroin and synthetic drugs, in Chicago between 2016 and 2019. Our research found that while heroin-related crime incidents dropped slightly in 2017, they rose again in 2018, and continually increased in 2019, staying mostly in the same locations. The pattern of crime incidents involving synthetic drugs evolved from a scattered pattern in 2016 to two distinct hotspot areas in 2019. These hotspots were in the locations of 2016 and 2017 heroin-related hotspots. Understanding where drug-related crimes have been occurring is important as it can be indicative of where drug use is happening. Analyzing the geospatial patterns of different drugs including successfully being able to predict how these patterns have changed over space and time can provide a basis for applying further treatment services and mitigation efforts, and also be useful for assessing current related services and efforts. Identifying built environment drivers for drug hot zones such as key locations including vacant buildings and vacant lots where drug-related crime frequently occurred can help public safety stakeholders with effective decision making relating to particular drugs. Future work could investigate additional key locations in more detail.

Acknowledgments

These analyses were funded through the National Drug Early Warning System (NDEWS) Coordinating Center at the University of Maryland's Center for Substance Abuse Research (CESAR). Authors are grateful for the support and encouragement received from NDEWS Coordinating Center staff, especially Dr. Eric Wish and Eleanor Artigiani. NDEWS is supported by the National Institute on Drug Abuse of the National Institutes of Health (NIH NIDA) under award number U01DA038360. This content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH NIDA. The authors thank the anonymous reviewers helped improve and clarify this manuscript.

Appendix A.: Supplementary materials for modeling

Table 1.

Summary of factors established to be linked to drug activities in literature and related candidate variables collected in this study

Factors	Related measures	Related variables collected in this study	References
Income inequality	Low income, medium income, and high income	Percentage of low-wage workers, percentage of individuals in poverty status, percentage of medium -wage workers, percentage of high-wage workers	Dasgupta et al., 2017; Lipton et al., 2013; Yarbrough et al., 2019
Low education level	College education	Percentage of population with college-level degree or higher	Nechuta et al., 2018
High unemployment rate	Employment status	Total number of workers, population at working age, employment, unemployment, gross employment density, full-time employment, part-time employment	Cooper et al., 2016; Lipton et al., 2013; McCord and Ratcliffe, 2007; Yarbrough et al., 2019
	Employment by category	Service jobs, industrial jobs, entertainment jobs, retail jobs, office jobs, education jobs, health care jobs, public administration jobs	

Factors	Related measures	Related variables collected in this study	References
Demographics	Population	Total population, residential density, population density	Chaney and Rojas-Guyler, 2015; Iyiewuare et al., 2017; Lipton et al., 2013; Marotta et al., 2019; Marshall et al., 2017; Reisner et al., 2015; Vilalta, 2010
	Household status	Total households, median household income, jobs per household, household workers per job percentage of household in poverty status, number of households own two or more automobiles, number of households own one automobile, number of households own zero automobile	
	Gender	Percentage of female population, percentage of male population	
	Race and ethnicity	Percentage of white Americans, percentage of African Americans, percentage of Asian Americans, Percentage of Hispanic/Latino Americans	
Community environment	Road network	Total road network density, street intersection density, network density of auto-oriented links, network density of pedestrian-oriented links, network density of multi-model links, density of auto-oriented intersections, density of pedestrian-oriented intersections, density of multi-model intersections	Cerdá et al., 2013; Marshall et al., 2017; Srinivasan et al., 2003; Sutter et al., 2019; Weisburd and Mazerolle, 2000
	Land use factors	Walkability index, total land area, total water area, total area, protected conservation area, area of unprotected land	
Community characteristics	Occupancy	Percentage of house occupancy, number of vacant buildings, number of vacant lots	Cerdá et al., 2017, 2013; Darke et al., 2001; Lipton et al., 2013; Martins et al., 2015; McCord and Ratcliffe, 2007; Moore et al., 2018; Sutter et al., 2019; Visconti et al., 2015
	House ownership	Percentage of house ownership, percentage of rental properties	
	Activity density and diversity	Gross house unites and employment density, gross retail employment density, gross office employment density, gross industrial employment density, gross service employment density, gross entertainment employment density, gross education employment density, gross health care employment density, gross public administration employment density, employment entropy, employment and household entropy, commute trip productions and trip attractions equilibrium index, regional diversity	
Key locations	Frequent locations for drug-related crimes	Alley density, number of vacant lots, number of vacant buildings, number of gas stations, number of parking lots, total area of park properties, number of high schools	Key locations were discussed in the Data section

Note: some variable names listed in the table refer to a group of variables (e.g., intersection density referring to intersection density of all intersections, three-leg intersection density and intersection density of intersections having four or more legs).

Table 2.

List of variables in the random forest model for predicting heroin-related activity hot zones

Variable name	Category	Description
GasStation	Built environment	Number of gas stations in each U.S. census block group
VacLot	Built environment	Number of vacant lots in each U.S. census block group
VacBldg	Built environment	Number of vacant buildings in each U.S. census block group
ParkingLot	Built environment	Number of parking lots in each U.S. census block group
AlleyDen	Built environment	Alley density
ParkArea	Built environment	Area of park propertied in each U.S. census block group
HighSch	Built environment	Number of high schools in each U.S. census block group
POccupyHouse	Built environment	% of house occupancy
POwner	Built environment	% of house ownership
PRenter	Built environment	% of house rent
D3a	Built environment	Total road network density
D3amm	Built environment	Network density of auto and pedestrian links
D3apo	Built environment	Network density of pedestrian links
D3b	Built environment	Street intersection density
D3bpo3	Built environment	Pedestrian-oriented intersection (three legs) density
D1A	Built environment	Gross residential density
D1C	Built environment	Gross employment density
D1C8_Off10	Built environment	Gross office employment density
D1D	Built environment	Gross activity density (employment and house units)
D2R_JOBPOP	Built environment	Regional Diversity (ratio of jobs/population)
TotPop	Sociodemographic	Total population in each U.S. census block group
TotHH	Sociodemographic	Total households in each U.S. census block group
PMale	Sociodemographic	Percentage of male population
PFemale	Sociodemographic	Percentage of female population
PWhite	Sociodemographic	% of population selecting race as white American alone
PBlack	Sociodemographic	% of population selecting race as black/African American alone
PAsian	Sociodemographic	% of population selecting race as Asian American alone
Phispanic	Sociodemographic	% of population selecting race as Hispanic/Latino American alone
PBachelorHigher	Sociodemographic	% of population with college level degree
MedianHHIncome	Sociodemographic	Median household income
PPovertyHouse	Sociodemographic	% of household in poverty status
PPovertyIndv	Sociodemographic	% of population in poverty status
PEmploy	Sociodemographic	% of full time and part time employees
Punemploy	Sociodemographic	% of unemployment
PFulltimeEmploy	Sociodemographic	% of full-time employment among all employment
PParttimeEmploy	Sociodemographic	% of part-time employment among all employment

Variable name	Category	Description
P_WRKAGE	Sociodemographic	% of population that is working aged
PCT_AO0	Sociodemographic	% of zero-car households
PCT_AO1	Sociodemographic	% of one-car households
AUTOOWN2P	Sociodemographic	Number of households that own two or more automobiles
PCT_AO2P	Sociodemographic	% of two-plus-car households
WORKERS	Sociodemographic	Number of workers
R_HIWAGEWK	Sociodemographic	Number of high wage workers
R_PCTLOWWA	Sociodemographic	% low wage workers
E5_IND10	Sociodemographic	Industrial jobs
E8_SVC10	Sociodemographic	Service jobs
HS_t_1	Spatial and temporal lag variables	Time-lagged variable (whether this BG belonged to a hotspot at t-1 period)
nb_t_1	Spatial and temporal lag variables	Spatiotemporally lagged variable (number of neighboring BGs belonged to a hotspot at t-1 period)
trend	Spatial and temporal lag variables	Trend variable (increase or decrease in the number of neighboring block groups that belonged to a hotspot at t-1 period)

Table 3.

List of variables in the random forest model for predicting synthetic drug-related activity hot zones

Variable name	Category	Description
GasStation	Built environment	Number of gas stations in each U.S. census block group
VacLot	Built environment	Number of vacant lots in each U.S. census block group
VacBldg	Built environment	Number of vacant buildings in each U.S. census block group
ParkingLot	Built environment	Number of parking lots in each U.S. census block group
AlleyDen	Built environment	Alley density
ParkArea	Built environment	Area of park propertied in each U.S. census block group
HighSch	Built environment	Number of high schools in each U.S. census block group
POccupyHouse	Built environment	% of house occupancy
POwner	Built environment	% of house ownership
PRenter	Built environment	% of house rent
AC_UNPR	Built environment	Total land area (not include park or conservation area)
D3b	Built environment	Street intersection density
D3bmm3	Built environment	Auto- and pedestrian-oriented intersection (three legs) density
D1A	Built environment	Gross residential density
D1C5_Ret10	Built environment	Gross retail employment density
D1D	Built environment	Gross activity density (employment and house units)
D2A_EPHHM	Built environment	Employment and household entropy
TotPop	Sociodemographic	Total population in each U.S. census block group

Variable name	Category	Description
TotHH	Sociodemographic	Total households in each U.S. census block group
PMale	Sociodemographic	Percentage of male population
PFemale	Sociodemographic	Percentage of female
PWhite	Sociodemographic	% of population selecting race as white American alone
PBlack	Sociodemographic	% of population selecting race as black/African American alone
Pasian	Sociodemographic	% of population selecting race as Asian American alone
Phispanic	Sociodemographic	% of population selecting race as Hispanic/Latino American alone
PBachelorHigher	Sociodemographic	% of population with college level degree
MedianHHIncome	Sociodemographic	Median household income
PPovertyHouse	Sociodemographic	% of household in poverty status
PPovertyIndv	Sociodemographic	% of population in poverty status
PEmploy	Sociodemographic	% of full time and part time employees
Punemploy	Sociodemographic	% of unemployment
PFulltimeEmploy	Sociodemographic	% of full-time employment among all employment
PParttimeEmploy	Sociodemographic	% of part-time employment among all employment
P_WRKAGE	Sociodemographic	% of population that is working aged
PCT_AO0	Sociodemographic	% of zero-car households
PCT_AO2P	Sociodemographic	% of two-plus-car households
R_HIWAGEWK	Sociodemographic	Number of high wage workers
R_PCTLOWWA	Sociodemographic	% low wage workers
E8_RET10	Sociodemographic	Retail jobs
E8_OFF10	Sociodemographic	Office jobs
HS_t_1	Spatial and temporal lag variables	Time-lagged variable (whether this BG belonged to a hotspot at t-1 period)
nb_t_1	Spatial and temporal lag variables	Spatiotemporally lagged variable (number of neighboring BGs belonged to a hotspot at t-1 period)
trend	Spatial and temporal lag variables	Trend variable (increase or decrease in the number of neighboring block groups that belonged to a hotspot at t-1 period)

Table 4.

Classification confusion matrixes, accuracy for predicting hotspot and non-hotspot classes and overall prediction accuracy (ACC)

Drug type		Heroin					Synthetic drug				
Model 1 (sociodemographic)	pred obs	hotspot	other	class accuracy	ACC		pred obs	hotspot	other	class accuracy	ACC
	hotspot	106	12	89.8%	97.7%		hotspot	96	56	63.2%	90.5%
	other	38	2046	98.2%			other	154	1896	92.5%	
Model 2 (built env)		hotspot	other					hotspot	other		
	hotspot	106	12	89.8%	97.7%		hotspot	96	56	63.2%	90.5%

Drug type		Heroin					Synthetic drug				
	other	38	2046	98.2%			other	154	1896	92.5%	
Model 3 (sociodemographic + built env)		hotspot	other					hotspot	other		
	hotspot	106	12	89.8%	97.7%		hotspot	96	56	63.2%	90.5%
	other	38	2046	98.2%			other	154	1896	92.5%	
Model 4 (sociodemographic + spatiotemporal autocorrelation)		hotspot	other					hotspot	other		
	hotspot	92	26	78.0%	98.4%		hotspot	93	59	61.2%	90.5%
	other	10	2074	99.5%			other	151	1899	92.6%	
Model 5 (built env + spatiotemporal autocorrelation)		hotspot	other					hotspot	other		
	hotspot	88	30	74.6%	98.2%		hotspot	97	55	63.8%	89.1%
	other	10	2074	99.5%			other	186	1864	90.9%	
Model 6 (spatiotemporal autocorrelation)		hotspot	other					hotspot	other		
	hotspot	87	31	73.7%	98.3%		hotspot	111	41	73.0%	87.5%
	other	7	2077	99.7%			other	234	1816	88.6%	
Model 7 (sociodemographic + built env + spatiotemporal autocorrelation)		hotspot	other					hotspot	other		
	hotspot	96	22	81.4%	98.3%		hotspot	91	61	60.0%	90.7%
	other	15	2069	99.3%			other	144	1906	93.0%	

Reference

Abraham AJ, Andrews CM, Yingling ME, Shannon J, 2018. Geographic Disparities in Availability of Opioid Use Disorder Treatment for Medicaid Enrollees. Health Services Research 53, 389–404. 10.1111/1475-6773.12686 [PubMed: 28345210]

Albright DL, McDaniel J, Kertesz S, Seal D, Prather K, English T, Laha-Walsh K, 2019. Small area estimation and hotspot identification of opioid use disorder among military veterans living in the Southern United States. Substance Abuse 0, 1–7. 10.1080/08897077.2019.1703066

Amaratunga D, Cabrera J, Lee Y-S, 2008. Enriched random forests. Bioinformatics 24, 2010–2014. 10.1093/bioinformatics/btn356 [PubMed: 18650208]

Ancuceanu R, Dinu M, Neaga I, Laszlo FG, Boda D, 2019. Development of QSAR machine learning-based models to forecast the effect of substances on malignant melanoma cells. Oncology Letters 17, 4188–4196. 10.3892/ol.2019.10068 [PubMed: 31007759]

Anselin L, 1995. Local Indicators of Spatial Association—LISA. Geographical Analysis 27, 93–115. 10.1111/j.1538-4632.1995.tb00338.x

Bernard S, Heutte L, Adam S, 2009. Influence of Hyperparameters on Random Forest Accuracy, in: Benediktsson JA, Kittler J, Roli F (Eds.), Multiple Classifier Systems, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 171–180. 10.1007/978-3-642-02326-2_18

Biau G, Scornet E, 2016. A random forest guided tour. TEST 25, 197–227. 10.1007/s11749-016-0481-7

Bivand RS, Pebesma E, Gómez-Rubio V, 2013. Spatio-Temporal Data, in: Bivand RS, Pebesma E, Gómez-Rubio V (Eds.), Applied Spatial Data Analysis with R, Use R! Springer, New York, NY, pp. 151–166. 10.1007/978-1-4614-7618-4_6

Borges J, Ziehr D, Beigl M, Cacho N, Martins A, Sudrich S, Abt S, Frey P, Knapp T, Etter M, Popp J, 2017. Feature engineering for crime hotspot detection, in: 2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation, pp. 1–8. 10.1109/UIC-ATC.2017.8397586

Breiman L, 2001. Random Forests. Machine Learning 45, 5–32. 10.1023/A:1010933404324

Breiman L, 1996. Out-of-bag estimation. Technical report, Department of Statistics: University of California, Berkeley

Calle ML, Urrea V, 2011. Letter to the Editor: Stability of Random Forest importance measures. Brief Bioinform 12, 86–89. 10.1093/bib/bbq011 [PubMed: 20360022]

Center for Behavioral Health Statistics and Quality, 2015. Behavioral health trends in the United States: Results from the 2014 National Survey on Drug Use and Health (HHS Publication No. SMA 15–4927, NSDUH Series H-50). Retrieved from http://www.samhsa.gov/data/

Center for Disease Control and Prevention, 2019a. Fentanyl | Drug Overdose | CDC Injury Center. Retrieved from https://www.cdc.gov/drugoverdose/opioids/fentanyl.html (accessed 8.30.19).

Center for Disease Control and Prevention, 2019b. Synthetic Opioid Overdose Data | Drug Overdose | CDC Injury Center. Retrieved from https://www.cdc.gov/drugoverdose/data/fentanyl.html (accessed 8.30.19).

Cerdá M, Gaidus A, Keyes KM, Ponicki W, Martins S, Galea S, Gruenewald P, 2017. Prescription opioid poisoning across urban and rural areas: identifying vulnerable groups and geographic areas. Addiction 112, 103–112. 10.1111/add.13543 [PubMed: 27470224]

Cerdá M, Ransome Y, Keyes KM, Koenen KC, Tardiff K, Vlahov D, Galea S, 2013. Revisiting the Role of the Urban Environment in Substance Use: The Case of Analgesic Overdose Fatalities. Am J Public Health 103, 2252–2260. 10.2105/AJPH.2013.301347 [PubMed: 24134362]

Chainey S, Tompson L, Uhlig S, 2008. The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. Secur J 21, 4–28. 10.1057/palgrave.sj.8350066

Chaney RA, Rojas-Guyler L, 2015. Spatial patterns of adolescent drug use. Applied Geography 56, 71–82. 10.1016/j.apgeog.2014.11.002

Chen W, Xie X, Wang J, Pradhan B, Hong H, Bui DT, Duan Z, Ma J, 2017. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. CATENA 151, 147–160. 10.1016/j.catena.2016.11.032

Chi G, Zhu J, 2008. Spatial Regression Models for Demographic Analysis. Popul Res Policy Rev 27, 17–42. 10.1007/s11113-007-9051-8

Ciccarone D, 2017. Fentanyl in the US heroin supply: A rapidly changing risk environment. Int. J. Drug Policy 46, 107–111. 10.1016/j.drugpo.2017.06.010 [PubMed: 28735776]

Cooper HLF, West B, Linton S, Hunter-Jones J, Zlotorzynska M, Stall R, Wolfe ME, Williams L, Hall HI, Cleland C, Tempalski B, Friedman SR, 2016. Contextual Predictors of Injection Drug Use Among Black Adolescents and Adults in US Metropolitan Areas, 1993–2007. Am J Public Health 106, 517–526. 10.2105/AJPH.2015.302911 [PubMed: 26691126]

Cryer JD, Chan K-S, 2008. Time Series Analysis: With Applications in R, 2nd ed, Springer Texts in Statistics. Springer-Verlag, New York. 10.1007/978-0-387-75959-3

Darke S, Kaye S, Ross J, 2001. Geographical injecting locations among injecting drug users in Sydney, Australia. Addiction 96, 241–246. 10.1046/j.1360-0443.2001.9622416.x [PubMed: 11182868]

Dasgupta N, Beletsky L, Ciccarone D, 2017. Opioid Crisis: No Easy Fix to Its Social and Economic Determinants. Am J Public Health 108, 182–186. 10.2105/AJPH.2017.304187 [PubMed: 29267060]

Deng H, Runger G, 2013. Gene selection with guided regularized random forest. Pattern Recognition 46, 3483–3489. 10.1016/j.patcog.2013.05.018

Deng H and Runger G, 2012, 6. Feature selection via regularized trees. In The 2012 International Joint Conference on Neural Networks (IJCNN) (pp. 1–8). IEEE.

Diggle PJ, 2013. Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, Third Edition, 3rd ed. Chapman and Hall/CRC. 10.1201/b15326

Fernández-Delgado M, Cernadas E, Barro S, Amorim D, 2014. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? Journal of Machine Learning Research 15, 3133–3181.

Friedman JH, 2001. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics 29, 1189–1232.

Georganos S, Grippa T, Gadiaga AN, Linard C, Lennert M, Vanhuysse S, Mboga N, Wolff E, Kalogirou S, 2019. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto International 0, 1–16. 10.1080/10106049.2019.1595177

Greenwell BM, 2017. pdp: An R Package for Constructing Partial Dependence Plots. 10.32614/rj-2017-016

He H, Garcia EA, 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering 21, 1263–1284. 10.1109/TKDE.2008.239

He L, Páez A, Liu D, 2017. Built environment and violent crime: An environmental audit approach using Google Street View. Computers, Environment and Urban Systems 66, 83–95. 10.1016/j.compenvurbsys.2017.08.001

Hedegaard H, Bastian B, Trinidad J, 2018. Drugs most frequently involved in drug overdose deaths: United States, 2011–2016. National Vital Statistics Reports; vol 67 no 9. Hyattsville, MD: National Center for Health Statistics. 2018. 14.

Hedegaard H, Bastian BA, Trinidad JP, Spencer MR, Warner M. 2019. Regional differences in the drugs most frequently involved in drug overdose deaths: United States, 2017. National Vital Statistics Reports; vol 68 no 12. Hyattsville, MD: National Center for Health Statistics.

Hodgkinson T, Andresen MA, 2019. Changing spatial patterns of residential burglary and the crime drop: The need for spatial data signatures. Journal of Criminal Justice 61, 90–100. 10.1016/j.jcrimjus.2019.04.003

Hunt JM, 2016. Do crime hot spots move? Exploring the effects of the modifiable areal unit problem and modifiable temporal unit problem on crime hot spot stability (Ph.D.). American University, United States -- District of Columbia.

Iyiewuare PO, McCullough C, Ober A, Becker K, Osilla K, Watkins KE, 2017. Demographic and Mental Health Characteristics of Individuals Who Present to Community Health Clinics With Substance Misuse. Health Serv Res Manag Epidemiol 4. 10.1177/2333392817734523

Jalal H, Buchanich JM, Roberts MS, Balmert LC, Zhang K, Burke DS, 2018. Changing dynamics of the drug overdose epidemic in the United States from 1979 through 2016. Science 361. 10.1126/science.aau1184

Japkowicz N, Stephen S, 2002. The class imbalance problem: A systematic study. Intelligent Data Analysis 6, 429. 10.3233/IDA-2002-6504

Jing Y, Hu Z, Fan P, Xue Y, Wang L, Tarter RE, Kirisci L, Wang J, Vanyukov M, Xie X-Q, 2020. Analysis of substance use and its outcomes by machine learning I. Childhood evaluation of liability to substance use disorder. Drug and Alcohol Dependence 206, 107605. 10.1016/j.drugalcdep.2019.107605 [PubMed: 31839402]

Kamel Boulos MN, Peng G, VoPham T, 2019. An overview of GeoAI applications in health and healthcare. International Journal of Health Geographics 18, 7. 10.1186/s12942-019-0171-2 [PubMed: 31043176]

Kotsiantis S, Kanellopoulos D and Pintelas P, 2006. Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering, 30(1), pp.25–36.

Li H, Calder CA, Cressie N, 2007. Beyond Moran's I: Testing for Spatial Dependence Based on the Spatial Autoregressive Model. Geographical Analysis 39, 357–375. 10.1111/j.1538-4632.2007.00708.x

Lipton R, Yang X, Braga A, Goldstick J, Newton M, Rura M, 2013. The Geographyof Violence, Alcohol Outlets, and Drug Arrests in Boston. American Journal of Public Health 103, 657–664. 10.2105/AJPH.2012.300927 [PubMed: 23409885]

Liu C, Wang T, 2017. Identifying and mapping local contributions of carbon emissions from urban motor and metro transports: A weighted multiproxy allocating approach. Computers, Environment and Urban Systems 64, 132–143. 10.1016/j.compenvurbsys.2017.01.010

Liu X, Wu J, Zhou Z, 2009. Exploratory Undersampling for Class-Imbalance Learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39, 539–550. 10.1109/TSMCB.2008.2007853

Lum K, Isaac W, 2016. To predict and serve? Significance 13, 14–19. 10.1111/j.1740-9713.2016.00960.x

Ma X, Guo J, Wu J, Liu H, Yu J, Xie J, Sun X, 2011. Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature. Proteins: Structure, Function, and Bioinformatics 79, 1230–1239. 10.1002/prot.22958

Malley JD, Kruppa J, Dasgupta A, Malley KG, Ziegler A, 2012. Probability machines: consistent probability estimation using nonparametric learning machines. Methods Inf Med 51, 74–81. 10.3414/ME00-01-0052 [PubMed: 21915433]

Marotta PL, Hunt T, Gilbert L, Wu E, Goddard-Eckrich D, El-Bassel N, 2019. Assessing Spatial Relationships between Prescription Drugs, Race, and Overdose in New York State from 2013 to 2015. Journal of Psychoactive Drugs 51, 360–370. 10.1080/02791072.2019.1599472 [PubMed: 31056042]

Marshall BDL, Krieger MS, Yedinak JL, Ogera P, Banerjee P, Alexander-Scott NE, Rich JD, Green TC, 2017. Epidemiology of fentanyl-involved drug overdose deaths: A geospatial retrospective study in Rhode Island, USA. International Journal of Drug Policy 46, 130–135. 10.1016/j.drugpo.2017.05.029 [PubMed: 28601512]

Martins SS, Sampson L, Cerdá M, Galea S, 2015. Worldwide Prevalence and Trends in Unintentional Drug Overdose: A Systematic Review of the Literature. Am J Public Health 105, e29–e49. 10.2105/AJPH.2015.302843

McCord ES, Ratcliffe JH, 2007. A Micro-Spatial Analysis of the Demographic and Criminogenic Environment of Drug Markets in Philadelphia. Australian & New Zealand Journal of Criminology 40, 43–63. 10.1375/acri.40.1.43

Mohler G, 2014. Marked point process hotspot maps for homicide and gun crime prediction in Chicago. International Journal of Forecasting 30, 491–497. 10.1016/j.ijforecast.2014.01.004

Molnar C, 2018. Interpretable machine learning. https://christophm.github.io/interpretable-ml-book/

Moore THM, Kesten JM, López-López JA, Ijaz S, McAleenan A, Richards A, Gray S, Savović J, Audrey S, 2018. The effects of changes to the built environment on the mental health and well-being of adults: Systematic review. Health & Place 53, 237–257. 10.1016/j.healthplace.2018.07.012 [PubMed: 30196042]

Moran PAP, 1950. Notes on Continuous Stochastic Phenomena. Biometrika 37, 17–23. 10.2307/2332142 [PubMed: 15420245]

Mosher CJ, Miethe TD, Hart TC, 2010. The Mismeasure of Crime. SAGE Publications.

National Institute on Drug Abuse, 2019. Opioids. Retrieved from: https://www.drugabuse.gov/drugs-abuse/opioids (accessed 11.19.19).

National Institute on Drug Abuse, 2019b. Overdose Death Rates. Retrieved from: https://www.drugabuse.gov/related-topics/trends-statistics/overdose-death-rates (accessed 8.30.19).

Nechuta SJ, Tyndall BD, Mukhopadhyay S, McPheeters ML, 2018. Sociodemographic factors, prescription history and opioid overdose deaths: a statewide analysis using linked PDMP and mortality data. Drug and Alcohol Dependence 190, 62–71. 10.1016/j.drugalcdep.2018.05.004 [PubMed: 29981943]

Nembrini S, König IR, Wright MN, 2018. The revival of the Gini importance? Bioinformatics 34, 3711–3718. 10.1093/bioinformatics/bty373 [PubMed: 29757357]

Neutens T, Farber S, Delafontaine M, Boussauw K, 2013. Spatial variation in the potential for social interaction: A case study in Flanders (Belgium). Computers, Environment and Urban Systems 41, 318–331. 10.1016/j.compenvurbsys.2012.06.007

Piza EL, Carter JG, 2018. Predicting Initiator and Near Repeat Events in Spatiotemporal Crime Patterns: An Analysis of Residential Burglary and Motor Vehicle Theft. Justice Quarterly 35, 842–870. 10.1080/07418825.2017.1342854

Probst P, Wright MN, Boulesteix A-L, 2019. Hyperparameters and tuning strategies for random forest. WIREs Data Mining and Knowledge Discovery 9, e1301. 10.1002/widm.1301

Reisner SL, Greytak EA, Parsons JT, Ybarra ML, 2015. Gender Minority Social Stress in Adolescence: Disparities in Adolescent Bullying and Substance Use by Gender Identity. The Journal of Sex Research 52, 243–256. 10.1080/00224499.2014.886321 [PubMed: 24742006]

Shiode S, Shiode N, 2020. A network-based scan statistic for detecting the exact location and extent of hotspots along urban streets. Computers, Environment and Urban Systems 83, 101500. 10.1016/j.compenvurbsys.2020.101500

Shiode S, Shiode N, Block R, Block CR, 2015. Space-time characteristics of micro-scale crime occurrences: an application of a network-based space-time search window technique for crime

incidents in Chicago. International Journal of Geographical Information Science 29, 697–719. 10.1080/13658816.2014.968782

Spencer MR, Warner M, Bastian BA, Trinidad JP, Hedegaard H, 2019. Drug Overdose Deaths Involving Fentanyl, 2011–2016. Natl Vital Stat Rep 68, 1–19.

Srinivasan S, O'Fallon LR, Dearry A, 2003. Creating Healthy Communities, Healthy Homes, Healthy People: Initiating a Research Agenda on the Built Environment and Public Health. Am J Public Health 93, 1446–1450. 10.2105/AJPH.93.9.1446 [PubMed: 12948961]

Stewart K, Cao Y, Hsu MH, Artigiani E, Wish E, 2017. Geospatial Analysis of Drug Poisoning Deaths Involving Heroin in the USA, 2000–2014. J Urban Health 94, 572–586. 10.1007/s11524-017-0177-7 [PubMed: 28639058]

Strobl C, Boulesteix A-L, Zeileis A, Hothorn T, 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8, 25. 10.1186/1471-2105-8-25 [PubMed: 17254353]

Substance Abuse and Mental Health Services Administration, 2019. Key substance use and mental health indicators in the United States: Results from the 2018 National Survey on Drug Use and Health (HHS Publication No. PEP19–5068, NSDUH Series H-54). Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. Retrieved from https://www.samhsa.gov/data/

Sutter A, Curtis M, Frost T, 2019. Public drug use in eight U.S. cities: Health risks and other factors associated with place of drug use. International Journal of Drug Policy 64, 62–69. 10.1016/j.drugpo.2018.11.007 [PubMed: 30580132]

United States Environmental Protection Agency, 2019. Basic Information about the Built Environment. US EPA. Retrieved from: https://www.epa.gov/smm/basic-information-about-built-environment (accessed 4.29.20).

Vilalta CJ, 2010. The spatial dynamics and socioeconomic correlates of drug arrests in Mexico city. Applied Geography 30, 263–270. 10.1016/j.apgeog.2009.06.001

Visconti AJ, Santos G-M, Lemos NP, Burke C, Coffin PO, 2015. Opioid Overdose Deaths in the City and County of San Francisco: Prevalence, Distribution, and Disparities. J Urban Health 92, 758–772. 10.1007/s11524-015-9967-y [PubMed: 26077643]

Vomfell L, Härdle WK, Lessmann S, 2018. Improving crime count forecasts using Twitter and taxi data. Decision Support Systems 113, 73–85. 10.1016/j.dss.2018.07.003

Weisburd D, Mazerolle LG, 2000. Crime and Disorder in Drug Hot Spots: Implications for Theory and Practice in Policing. Police Quarterly 3, 331–349. 10.1177/1098611100003003006

Wright MN, Ziegler A, 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software 77, 1–17. 10.18637/jss.v077.i01

Yang B, Liu L, Lan M, Wang Z, Zhou H, Yu H, 2020. A spatio-temporal method for crime prediction using historical crime data and transitional zones identified from nightlight imagery. International Journal of Geographical Information Science 0, 1–25. 10.1080/13658816.2020.1737701

Yarbrough CR, Abraham AJ, Adams GB, 2019. Relationship of County Opioid Epidemic Severity to Changes in Access to Substance Use Disorder Treatment, 2009–2017. PS 71, 12–20. 10.1176/appi.ps.201900150

Ye X, Wu L, 2011. Analyzing the dynamics of homicide patterns in Chicago: ESDA and spatial panel approaches. Applied Geography 31, 800–807. 10.1016/j.apgeog.2010.08.006

Zhang Y, Cheng T, 2020. Graph deep learning model for network-based predictive hotspot mapping of sparse spatio-temporal events. Computers, Environment and Urban Systems 79, 101403. 10.1016/j.compenvurbsys.2019.101403

Zhao X, Tang J, 2017. Modeling Temporal-Spatial Correlations for Crime Prediction, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17. Association for Computing Machinery, Singapore, Singapore, pp. 497–506. 10.1145/3132847.3133024

Highlights:

- Underlying factors for patterns of drug activities involving heroin and synthetic drugs were identified
- Integrating space-time analysis framework and machine learning to analyze patterns of repeated events in an urban context
- Accommodating both spatial and temporal autocorrelation in the model learning process

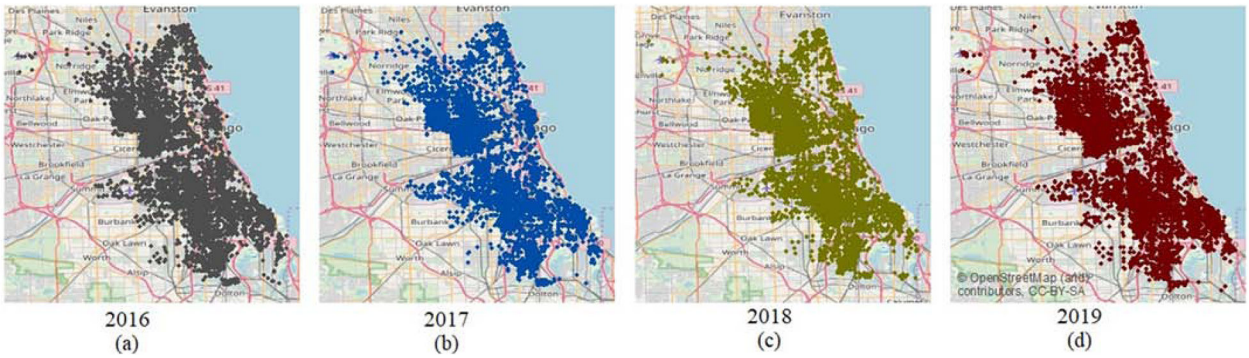

Figure 1.
Drug-related crimes for all categories of drugs in Chicago for 2016, 2017, 2018 and 2019.

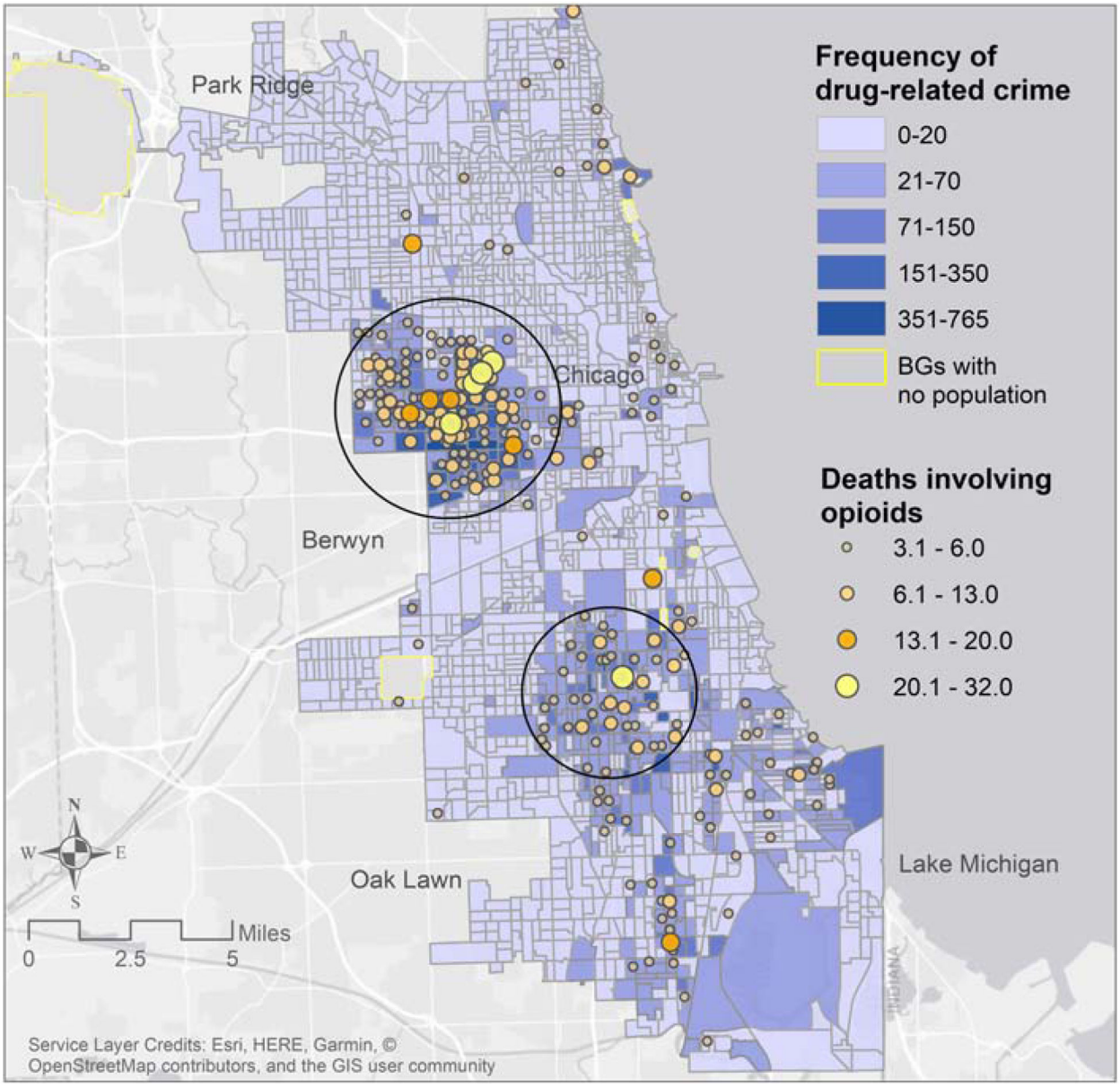

Figure 2.

Frequency of drug-related crime incidents adjusted by population (count/population*1000), and number of deaths involving opioids adjusted by population (count/population*1000) by block group between 2016 and 2019

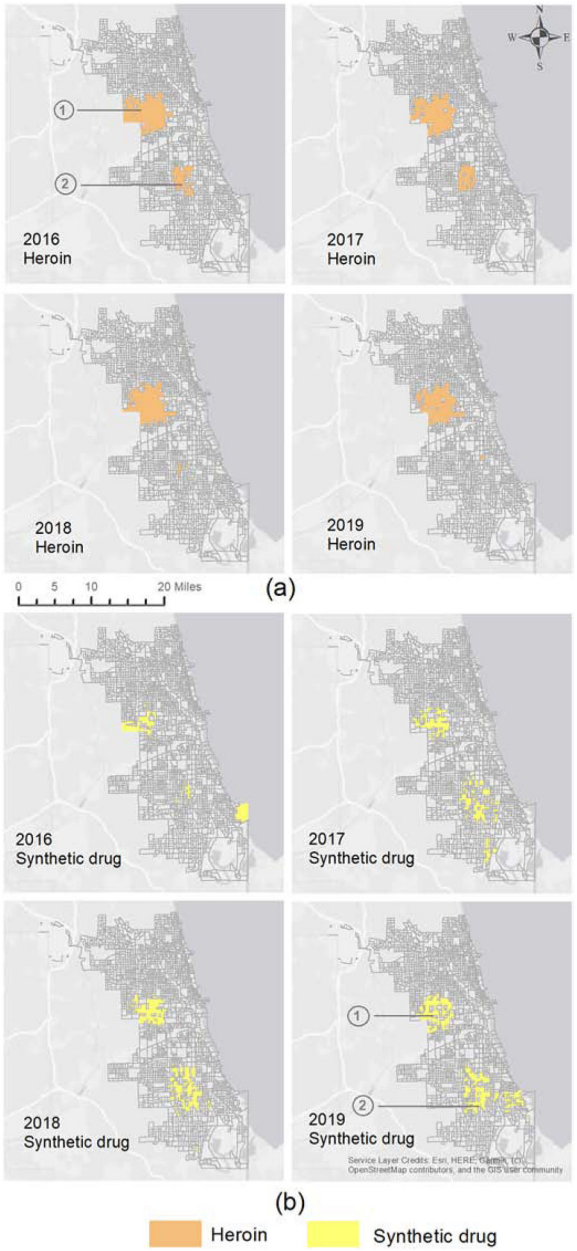

Figure 3.
Spatial clustering of drug-related crime incidents involving (a) heroin and (b) synthetic drugs in Chicago for 2016, 2017, 2018 and 2019

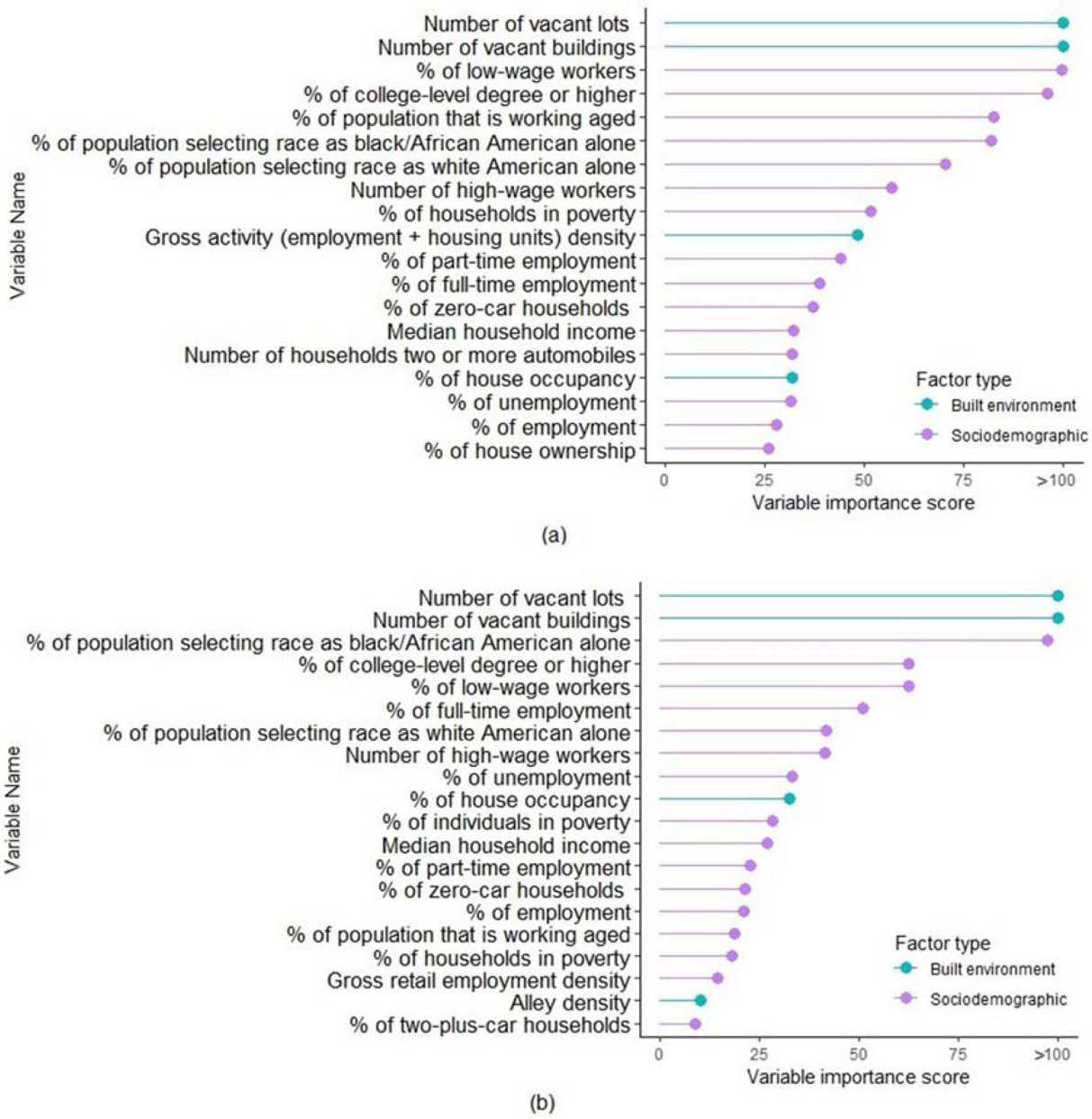

Figure 4.

(a) top 20 variables with the highest importance score for classifying heroin hotspots; (b) top 20 variables with the highest importance score for classifying synthetic drug hotspots.

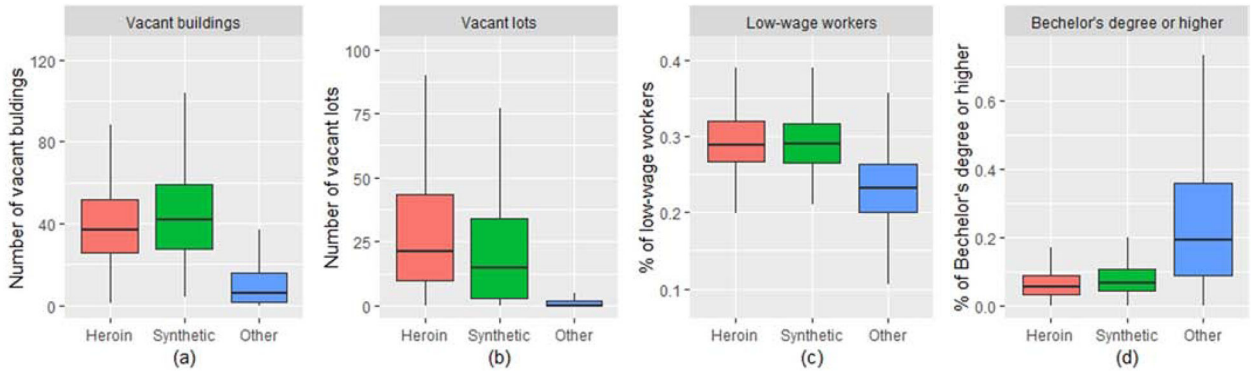

Figure 5.

Boxplots of (a) number of vacant buildings; (b) number of vacant lots; (c) percentage of low-wage workers and (d) percentage of population with Bachelor's or higher degree for heroin-related crime hotspots, synthetic drug-related crime hotspots and other block groups

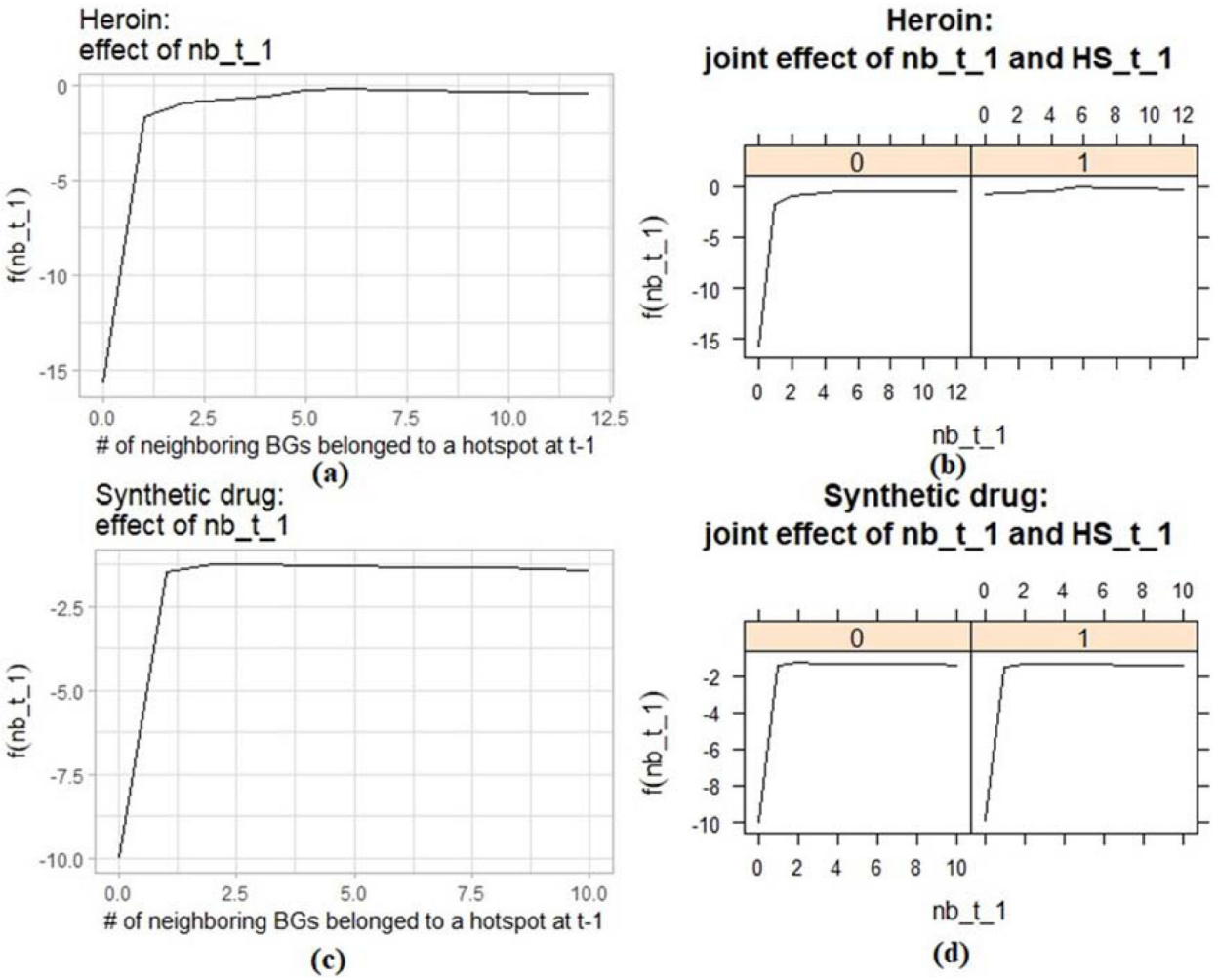

Figure 6.

Partial dependence plots of classification of heroin-related crime hotspots on selected variables (a) *nb_t_1* (number of neighboring block groups belonged to a heroin hotspot at *t-1*); (b) joint effect of *nb_t_1* and *HS_t_1* (whether this block group belonged to a heroin hotspot at *t-1*, class *0* represented non-hotspot block group and class *1* represented hotspot block group); partial dependence plots of classification of synthetic drug-related crime hotspots on selected variables (c) *nb_t_1*; (d) joint effect of *nb_t_1* and *HS_t_1*

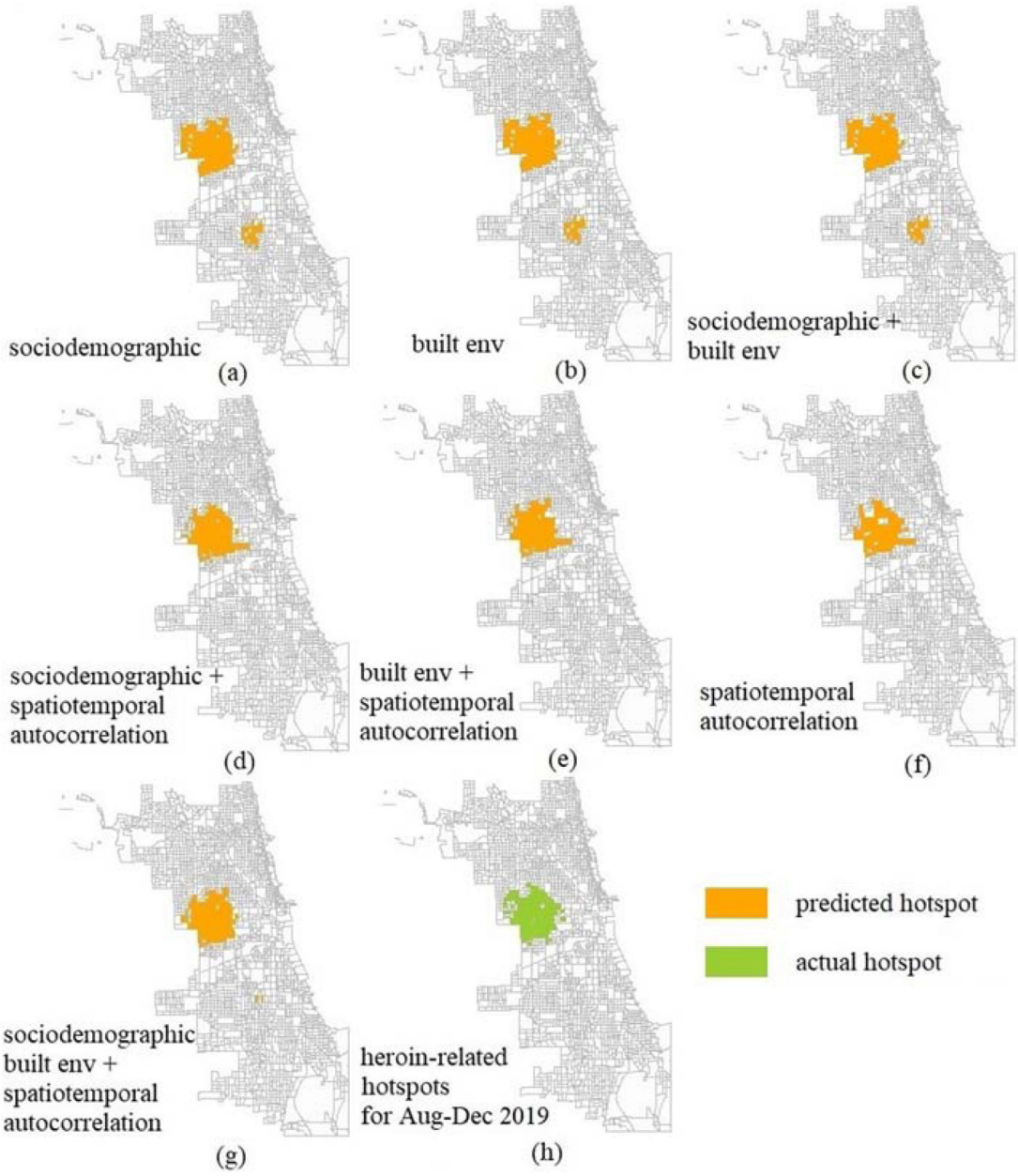

Figure 7.

Space-time random forest model predicted heroin-related hotspot map using input variables: (a) sociodemographic factors (b) built environment factors (c) sociodemographic and built environment factors (d) sociodemographic factors and spatiotemporal autocorrelation (e) built environment factors and spatiotemporal autocorrelation (f) spatiotemporal autocorrelation (g) sociodemographic, built environment factors and spatiotemporal autocorrelation and (h) actual heroin hotspots calculated from the drug crime data, for August-December 2019 in Chicago

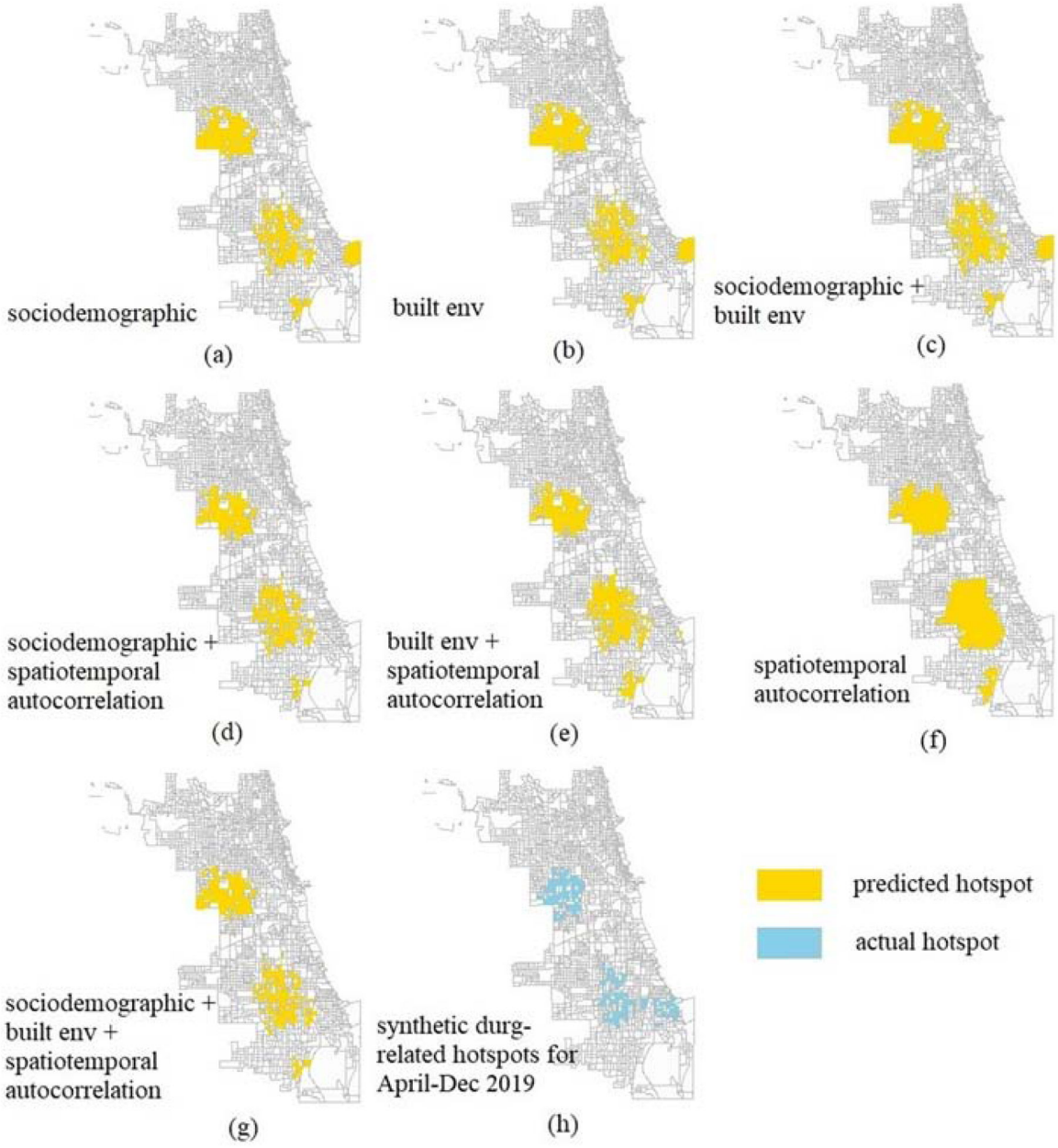

Figure 8.

Space-time random forest model predicted synthetic drug-related hotspot map using input variables: (a) sociodemographic factors (b) built environment factors (c) sociodemographic and built environment factors (d) sociodemographic factors and spatiotemporal autocorrelation (e) built environment factors and spatiotemporal autocorrelation (f) spatiotemporal autocorrelation (g) sociodemographic, built environment factors and spatiotemporal autocorrelation and (h) actual synthetic drug hotspots calculated from drug-related crime data, for April-December 2019 in Chicago

Table 1.

Summary of drug-related crimes in Chicago for 2016, 2017, 2018 and 2019

Year Number of incidents	2016	2017	2018	2019
Total drug-related crimes	13318	11677	13495	14077
Heroin-related incidents	3500	3449	3807	4065
Synthetic drug-related incidents	269	309	396	521

Table 2.

Deaths involving opioid in Chicago for 2016, 2017, 2018 and 2019

Year Number of deaths	2016	2017	2018	2019
Total deaths involving opioids	743	794	801	879
Deaths involving heroin	418	542	489	473
Deaths involving fentanyl	405	465	624	685

Table 3.

Model evaluation: classification accuracy (ACC), prediction accuracy index (PAI) and prediction efficiency index (PEI)

Drug Type Model	Heroin			Synthetic drug		
	ACC	PAI	PEI	ACC	PAI	PEI
Model 1 (sociodemographic)	97.7%	11.02	0.877	90.5%	4.37	0.548
Model 2 (built env)	97.7%	11.02	0.877	90.5%	4.37	0.548
Model 3 (sociodemographic + built env)	97.7%	11.02	0.877	90.5%	4.24	0.542
Model 4 (sociodemographic + spatiotemporal autocorrelation)	98.4%	14.34	0.877	90.5%	3.90	0.540
Model 5 (built env + spatiotemporal autocorrelation)	98.2%	14.75	0.876	89.1%	4.34	0.508
Model 6 (spatiotemporal autocorrelation)	98.3%	15.29	0.878	87.5%	3.47	0.544
Model 7 (sociodemographic + built env + Spatiotemporal autocorrelation)	98.3%	13.44	0.876	90.7%	4.47	0.549