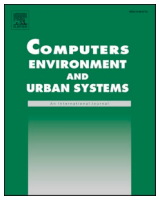
Computers, Environment and Urban Systems 89 (2021) 101660

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus



A data-driven agent-based simulation to predict crime patterns in an urban environment

Raquel Rosés [a,*], Cristina Kadar [a], Nick Malleson [b]

[a] Swiss Federal Institute of Technology, 8092 Zurich, Switzerland
[b] University of Leeds, Leeds LS2 9JT, United Kingdom

ARTICLE INFO

Keywords:
Simulation
Agent-based models
Data-driven models
Open data
Urban data
Location-based social services
Crime prediction
2000 MSC: 91D25
91D10

ABSTRACT

Spatial crime simulations contribute to our understanding of the mechanisms that drive crime and can support decision-makers in developing effective crime reduction strategies. Agent-based models that integrate geographical environments to generate crime patterns have emerged in recent years, although *data-driven* crime simulations are scarce. This article (1) identifies numerous important drivers of crime patterns, (2) collects relevant, openly available data sources to build a GIS-layer with static and dynamic geographical, as well as temporal features relevant to crime, (3) builds a virtual urban environment with these layers, in which individual offender agents navigate, (4) proposes a data-driven decision-making process using machine-learning for the agents to decide whether to engage in criminal activity based on their perception of the environment and, finally, (5) generates fine-grained crime patterns in a simulated urban environment. The novelty of this work lies in the various large-scale data layers, the integration of machine learning at individual agent level to process the data layers, and the high resolution of the resulting predictions. The results show that the spatial, temporal, and interaction layers are all required to predict the top street segments with the highest number of crimes. In addition, the spatial layer is the most informative, which means that spatial data contributes most to predictive performance. Thus, these findings highlight the importance of the inclusion of various open data sources and the potential of theory-informed, data-driven simulations for the purpose of crime prediction. The resulting model is applicable as a predictive tool and as a test platform to support crime reduction.

1. Introduction

Crime is a complex phenomenon with significant social and financial implications. Crimes occur as individuals interact with each other and their environment (Cohen & Felson, 1979; Felson, 2011). Such interactions may be hard to capture using statistical techniques. Thus, simulation techniques have been applied to crime and are being utilized in an effort to generate realistic crime patterns. In the era of big data and predictive analytics, researchers and industry representatives alike are developing advanced models to predict crime (Bowers, Johnson, & Pease, 2004). These models are intended to improve the effectiveness of crime reduction strategies (Wang & Brown, 2012), e.g. by informing police departments about how to allocate their resources more efficiently or by allowing urban planners to test the impact of urban interventions on crime (Groff, Johnson, & Thornton, 2018).

One approach in particular, agent-based modeling, is ideally suited to the simulation of social systems and complex social phenomena, like crime. Agent-based models (ABMs) of crime are predominantly used to explore theory (Groff et al., 2018), which limits their use by practitioners. This is in contrast to crime prediction models with statistical techniques (e.g. machine learning) that rely on using as much information from the real world as possible. However, researchers building ABMs more generally – e.g. in the domains of energy markets (Zhang, Vorobeychik, Letchford, & Lakkaraju, 2016), epidemiology (Hunter, Mac Namee, & Kelleher, 2018), and counter-piracy (Vaněk, Jakob, Hrstka, & Pěchouček, 2013) – are demonstrating the value of agent-based models that are both theoretically and empirically informed.

In light of these successful data-driven statistical models and theory-based ABMs, this article aims at building a data-driven crime prediction model, based entirely on openly available data, that is also broadly informed by relevant environmental criminology theory. The model aims to generate realistic crime patterns and is therefore evaluated by comparing the results to real crime data at the level of the street segment. In particular, this approach is useful in determining which data

* Corresponding author.
E-mail address: raquelroses@outlook.com (R. Rosés).

https://doi.org/10.1016/j.compenvurbsys.2021.101660
Received 3 August 2020; Received in revised form 16 May 2021; Accepted 18 May 2021
Available online 6 July 2021

sources, in combination with ML techniques, contribute most to the predictive performance of the model.

In summary, this article (1) identifies numerous important drivers of crime patterns, (2) collects relevant, openly available data sources to build static and dynamic geographical features as well as temporal features relevant to crime, (3) builds a virtual urban environment with these layers, in which individual offender agents navigate and interact, (4) proposes a data-driven decision-making process for the agents using ML to decide whether to engage in criminal activity based on their perception of the environment and, finally, (5) generates fine-grained crime patterns in a simulated urban environment. The case study area here is New York City (NYC). It is important to note that the goal of the simulation is to generate a realistic crime pattern and while agent-based modeling allows for the instantiation of individual behavior, virtual offender agents in this simulation do not attempt to precisely replicate the behavioural patterns of real individuals. The aim of the simulation is to develop offender agent behavior leading to a realistic robbery pattern in an urban environment.

The model in this study shows the value of building a data-driven simulation, supported by insights from criminology, as well as the advantage of combining a variety of spatial and temporal data sources with ML techniques to increase model performance. Overall, the resulting simulation shows high predictive power for crime counts at the street segment level. Interestingly, compared to temporal data sources and an interaction component, the spatial data sources contribute most to model performance. This points to the possibility that the surrounding spatial environment is more important in crime prediction than the temporal fluctuations caused by weather.

2. Related work

Simulation techniques, such as ABM, have been applied to the study of crime in a wide variety of setups, including theory testing (Birks, Townsley, & Stewart, 2014), testing of prevention strategies (Bosse & Gerritsen, 2010; Devia & Weber, 2013), and predicting crime patterns (Gunderson & Brown, 2000). In line with the previous section, we identify three topics related to crime simulation which are relevant to our work: the importance of realistic spatial environments in order to simulate crime, theories and environmental risk factors for crime, and the structure of agents' internal representation.

2.1. Realistic spatial environments

A systematic review of existing ABMs for crime has been performed by Groff et al. (2018). They identify only 33% of crime simulations as being built on top of a GIS-layer or street centerlines, in contrast to the remainder of the simulation, which do not use spatially explicit environments. This is out of step with the growing empirical evidence showing the relevance of street network configuration for modeling crime (Davies & Johnson, 2015; Rosser, Davies, Bowers, Johnson, & Cheng, 2016; Summers & Johnson, 2017). While some researchers argue against the importance of realistic spatial layers for understanding mechanisms driving crime (Elffers & van Baal, 2008), others (Malleson, Heppenstall, & See, 2010) call for realistic models that can be used to derive actionable insights for police and policymakers.

The first model that integrated a GIS layer with simulated virtual robbers moving along a street network demonstrated the impact of using real-world data (total population, employment rate, potential activities) within a crime simulation (Groff, 2007a; Groff, 2008). In contrast, Furtado, Melo, Coelho, Menezes, and Perrone, (2009) build a bio-inspired crime simulation focused on finding the offender agents' starting locations using AI (genetic algorithms) along with a geographic layer with points of interest (POI) located on the street network. Furthermore, Malleson et al. (2010), Malleson and Birkin (2012) and Malleson, Heppenstall, See, and Evans (2013) build detailed burglary simulations integrating geographical features from real-world data such as the road network, buildings, and community characteristics (from census data). These works incorporate a realistic representation of the environment specifically relevant for burglars. Devia and Weber (2013) also put a focus on a variety of geographic features by building a model to generate crime data integrating businesses, banks, offices, educational establishments, and public transit stations aggregated on a grid, and instantiating unique structures for different urban environments. Likewise, Peng and Kurland (2014) build a simulation to reproduce burglary patterns in Beijing, including the street and subway networks as a layer in the spatial environment. This work shows the potential of simulating crime at street segment level rather than as an aggregated unit (e.g. grid cells). Furthermore, Rosés, Kadar, Gerritsen, and Rouly (2018, 2020) uses a micro-simulation focused on informing offender mobility using large-scale geographical data (e.g. taxi trip data and location-based social networks) to represent a realistic environment relevant to criminals, which is intended to be further developed into an ABM. These latter works generate patterns of individual mobility cumulatively matching real crime patterns, demonstrating the benefit of using large-scale human activity data to inform offender mobility.

There are a number of limitations associated with these previous works that this paper takes steps to overcome. First, they often use known offender home locations for modeling offenders' spatial mobility; thus, their insights might not be generalizable to unknown offenders (Lammers & Bernasco, 2013). Furthermore, the models make a number of assumptions regarding the behavior of offenders that, whilst informed by theory, are not fully quantified due to a lack of detailed empirical data. With regard to contextual data, the geographic features representing the environment are often static and therefore unable to capture changes within the environment which affect the evolution of crime patterns. Finally, previous models do not capture the dynamics of mobile populations, such as visitors to city centres, which will undoubtedly have a significant influence on observed crime patterns. Overall, researchers have yet to build a model that accounts for a large variety of static and dynamic environmental features that have been shown to relate to crime.

2.2. Theories, environmental risk factors for crime and available data sources

Specific human behavior leads to offenders converging with victims in time and space. According to routine activity theory (Cohen & Felson, 1979), crime happens when a motivated offender meets a suitable target in the absence of capable guardians. Such situations arise as part of the repetitive motions (or regular activities) of offenders and victims alike: people go to work, school, recreational activities, or socialize. Furthermore, crime pattern theory (Brantingham & Brantingham, 1993) ascribes a strong geographic component to such situations. Offenders commonly commit crimes in spaces that are part of their regular activities. These theoretical approaches lead to an understanding of offenders as part of a general population with similar movement strategies. Moreover, this means that no clear differentiation between the general movement patterns of victims and offenders can be assumed. Together with the latest advances in studying human movement by means of openly available data, there is a unique opportunity to study offender movement in new ways. Previous work has focused on studying offender mobility strategies using openly available data and insights from human mobility studies (Rosés et al., 2020), evaluating different data sources and rules for virtual offenders navigating an urban environment. The current study builds on this previous simulation model and extends it with further data sources and decision-making models. Indeed, to be able to reproduce crime patterns, virtual offender agents need not only to be able to navigate the space, but also to choose when to commit crimes along the paths they travel. Theories such as rational choice (Cornish & Clarke, 1987) propose that offenders exhibit rational behavior, taking into account aspects of the specific situation to decide whether to commit a crime. This implies that the choice to commit a

crime is dependent on personal preferences, but also on a variety of characteristics of a specific location. As the personal preferences are difficult to model and account for, this work focuses especially on capturing a large variety of situational characteristics that may affect simulated individual behavior at a small spatial unit of analysis: the street segment.

This work focuses on robbery as a case study. There exists a large body of theoretical and empirical literature analyzing the relationship between robbery and related spatio-temporal factors, which can be included in a simulation to account for aspects related to crime. Table 1 outlines some of the factors that influence robbery, as established in the literature, for which data is publicly available. The remainder of this section critiques the literature in detail; subsequent sections then describe the data (see appendices for more details) and explain how the data will be used within the simulation (see Section 4).

Traditionally, crime models have been built either on the basis of statistical data, such as population size or density (Bernasco & Block, 2009), or land use data reflecting information on types of zoning (Browning et al., 2010; Mccord & Ratcliffe, 2009; Stucky & Ottensmann, 2009; Sypion-Dutkowska & Leitner, 2017; Twinam, 2017). Lately there has been a shift towards making additional datasets publicly available,

Table 1
Factors influencing crime from empirical literature, to be used as features for further modeling.

Factor	Source	Feature	Corresponding data source
Urban structure and connectivity of places	(Beavon, Brantingham, & Brantingham, 1994; Davies & Bowers, 2019; Davies & Johnson, 2015)	Street network	Street centerlines
Connectivity and movement of potential victims or offenders	(Liu, Kang, Gao, Xiao, & Tian, 2012a; Tang, Liu, Wang, & Wang, 2015; Wang, Schoenebeck, Zheng, & Zhao, 2016a)	Human mobility patterns	Taxi trips
Presence of potential victims and type of activities at the location	(Chainey & Desyllas, 2008; Kadar & Pletikosa, 2018)	Human activity patterns	LBSN
	(Browning et al., 2010; Mccord & Ratcliffe, 2009; Stucky & Ottensmann, 2009; Sypion-Dutkowska & Leitner, 2017; Twinam, 2017)	Land use	land use
Presence of potential victims	(Tompson, 2015)	Points of interest	points of interest
	(Drucker, 2010)	Schools	schools
	(Drucker, 2010)	Public transportation stops	bus stops and subway entrances
Availability of victims in public spaces	(Sorg & Taylor, 2011; Tompson & Bowers, 2015)	Weather	weather
Visibility of offenders	(Bogar & Beyer, 2016; Donovan & Prestemon, 2012; Kondo, Han, Donovan, & MacDonald, 2017)	Urban trees	tree census
Density of residents as potential victims	(Bernasco & Block, 2009)	Population density	census data: resident population
Indicator for physical disorder of an area	(Wheeler, 2018)	Calls for service	311 calls

such as those pertaining to public buildings and parks (Tompson & Bowers, 2015), public transportation stops like buses and subways (Bernasco & Block, 2011; Drucker, 2010; Hart & Miethe, 2014), schools (Drucker, 2010; Murray & Swatt, 2013; Willits, Broidy, & Denman, 2013), and urban trees (Bogar & Beyer, 2016; Donovan & Prestemon, 2012; Kondo et al., 2017), all of which have been linked to crime.

Street network connectivity is related to crime, as crime concentrations are higher in more connected areas (Beavon et al., 1994; Davies & Bowers, 2019; Davies & Johnson, 2015). Indeed, a newer stream of criminology looking in detail at *crime at places* stresses the importance of studying crime in micro-places in order to be able to correctly capture crime patterns (Lee, Eck, Soohyun, & Martinez, 2017; Weisburd, 2015). A small proportion of places contain most of the crime in an urban environment (Lee et al., 2017); therefore, models should operate on the level of those micro places.

The newest stream of available data is from location-based social networks (LBSNs),[1] whereby location types and popularity can be used as a proxy to model the attractiveness of specific urban locations for victims and offenders alike (Chainey & Desyllas, 2008; Reid, Frank, Iwanski, Dabbaghian, & Brantingham, 2013). In this sense, offender movement patterns are presumed to be similar to those of the general population (Brantingham & Brantingham, 1993; Morselli & Royer, 2008; Song et al., 2019), which is also in accordance with routine activity theory. In this sense, Brantingham and Brantingham (1993) describe how some crimes are highly opportunistic and dependent on offender's and victim's daily activities and the availability of opportunities arising from these activities. More importantly, victims and offenders must necessarily coincide at specific locations for a crime to happen; Morselli and Royer (2008) argue about the increased earnings achieved by mobile offenders (committing crimes in multiple cities) in contrast to"immobile" offenders (committing crimes only in their home city), focusing on the relationship between earnings and travel distance, but without analyzing distance travel behavior of offenders; and Song et al. (2019) studies the use of urban population flows to predict crime locations, finding that higher connectivity between a location and an offender's residence increases risk of crime. This study suggests that criminals flow in similar ways than the general population, while acknowledging that offenders may be more spatially restricted towards their homes (at least for crime commission purpose). In other domains, human activity patterns have been extracted from LBSNs and found to be reliable within urban areas (although not necessarily in rural areas) (Cranshaw, Schwartz, Hong, & Sadeh, 2012; Hecht & Stephens, 2015; Nguyen & Szymanski, 2012; Noulas, Mascolo, & Frias-Martinez, 2013; Noulas, Scellato, Lambiotte, Pontil, & Mascolo, 2012; Noulas, Scellato, Mascolo, & Pontil, 2011). The advantage of LBSN data over other similar data sources (e.g. GPS phone location data), is the possibility to derive information regarding the context of activities that the users engage in; users 'check in' to locations broadly categorized by a type of activity (Cranshaw et al., 2012; Noulas et al., 2013). Once again, this is in line with routine activity theory, whereby popular locations can be viewed as popular activity nodes for offenders and victims alike. Concerns and advantages of using LBSNs in this context have been discussed among others in Rosés et al. (2020). Numerous researchers have shown LBSNs to be useful for predicting crime (Al Boni & Gerber, 2016; Bogomolov et al., 2014; Kadar & Pletikosa, 2018; Rosés et al., 2020; Wang, Kifer, Graif, & Li, 2016b; Yang, Heaney, Tonon, Wang, & Cudré-Mauroux, 2017) for example as a proxy for the *ambient population* (Andresen, 2011).

Another source of data that can be used as a proxy for human mobility is taxi trip data (Liu, Kang, et al., 2012a; Liu, Wang, Xiao, & Gao, 2012b; Tang et al., 2015). Although only a fraction of the population uses taxis, the penetration of this service in New York City (NYC,

[1] LBSNs are social networks offering GPS features tracking a person's location and broadcasting this location and other content from a mobile device.

the case study area) is very high. Additionally, taxi services are most commonly used for social and recreational trips (New York State Department of Transportation, 2018), which makes them a good proxy for human activity in the study of crime. This is especially true in light of crime pattern theory, which identifies a correspondence between locations more prone to crime (crime attractors and generators) and locations attracting higher number of visitors (Kinney, Brantingham, Wuschke, Kirk, & Brantingham, 2008). Taxi trip data has successfully been used to inform crime prediction models in Kadar and Pletikosa (2018), Vomfell, Händle, and Lessmann (2018) and Wang, Kifer et al. (2016b). Indeed, Vomfell et al. (2018) concludes that the addition of dynamic features, such as taxi data, in combination with static data, such as point-of-interest locations, greatly reduces prediction errors in a model of property crimes.

The final source of dynamic data used here is calls for service, or 311 calls (non-emergency calls). Calls for service are citizens' complaints to public servants about a variety of topics such as noise complaints, blocked driveways, illegal parking, animal abuse, or damaged trees. This service does not allow callers to report crimes, though it still includes calls that are redirected to the NYPD. A call for service implies that an issue has arisen and that a resident has decided to report it. On one hand, 311 calls can be understood as a measure of the demand for services placed to the city government by citizens within a certain neighborhood (White & Trump, 2018), while reporting incivilities has been linked to the residents' desire to enforce social norms (O'Brien, 2016). On the other hand, 311 calls are a potential source of information about physical disorder in a neighborhood (Wheeler, 2018). The use of 311 data in criminology research has been theoretically underpinned by its association with "broken windows theory".[2] Physical disorder has been hypothesised to both lead to social disorder (Wilson & Kelling, 1982) and to be a sign of a lack of guardianship (Cohen & Felson, 1979; Newman, 1972; Taylor & Gottfredson, 1986), both of which are factors that have been found to be associated with violent crime (Ahlin & Lobo Antunes, 2017; Schnell, Grossman, & Braga, 2019; Stein, Conley, & Davis, 2016). Moreover, Wheeler (2018) empirically identified 311 calls to be a valid indicator for disorder. In summary, calls for service (or 311 calls) can capture different aspects of street segments, from disorder to the responsiveness of citizens to disorder and other issues within their neighborhood, thereby providing specific and dynamic information about a particular street segment over time.

Apart from geographical data, crime is also known to be influenced by temporal factors related to weather conditions and their subsequent impact on activity patterns (Ceccato, 2005; Horanont, Phithakkitnukoon, Leong, Sekimoto, & Shibasaki, 2013). Empirical findings especially link temperate and more favourable weather in winter to robbery (Tompson & Bowers, 2015), rain on weekends to less robbery (Tompson & Bowers, 2015), and larger variations in climate during the summer months to an increase in property crimes (Linning, Andresen, Ghaseminejad, & Brantingham, 2017). Weather affects the routine activities of victims and offenders, thus impacting the availability of opportunities.

Yet another factor influencing crime is previously committed crime. This is referred to as the phenomenon of near-repeat (i.e. localized repeat victimization) (Johnson, Bowers, & Hirschfield, 1997), which has been shown to exist for robberies (Glasner & Leitner, 2017). Near-repeat theories detail the increased risk of certain locations to be victimized repeatedly within a short period of time. This means that after a robbery has occurred at a certain location, the probability of a subsequent robbery happening within a small amount of time is high. Near-repeats are usually related to characteristics of a specific location that increase their attractiveness over a short period of time, which may be hard to capture using other indicators. Accounting for this phenomenon within a simulation, which allows for individual interaction, may contribute to a more accurate representation of crime.

So far, none of the crime simulations discussed have included dynamic environmental features or temporal features. Thus, our aim is to build a simulation model taking into account all of the abovementioned factors, and to test the contribution of the dynamic features, temporal features, and interaction features for crime prediction simulations in different scenarios.

2.3. Agents' internal representation

In the previous subsections we paid close attention to the importance of a realistic representation of the environment for crime simulations and looked at the risk factors related to crime. Agent-based modeling allows us to represent individual agents (e.g. virtual offenders) performing actions in the environment and affecting the overall status of the simulation (e.g. the resulting crime pattern). This individual representation tends to incorporate functions or rules that contribute to the overall goal of the simulation (i.e. to reproduce a realistic crime pattern). In this section, we look into how crime simulations build the internal representation of the agents, which is how the agents process the information that they receive from the virtual environment.

In the first model of its kind, Groff and colleagues (Groff, 2007a; Groff, 2008) build a spatial crime simulation with a detailed, theoretically informed agent decision-making process. Many subsequent efforts build on this, including: Wang, Liu, and Eck (2008) and Wang, Liu, and Eck (2014), who integrate artificial intelligence reinforcement learning (although their simulation is largely theory-based and contains no real data); Malleson et al. (2010), Malleson et al. (2013), Malleson and Birkin (2012) and Malleson, Evans, and Jenkins (2009), who include a detailed behavioural framework called PECS (Urban & Schmidt, 2001); Devia and Weber (2013), who develop a probability function to represent decision-making; and Peng and Kurland (2014), who use a theory-based statistical approach. All of the above representations of offender decision-making are largely theory-based and conceptual, relying on limited empirical data. In contrast, we propose an approach to infer offender decision-making from real-world data using machine learning techniques, allowing for easy transfer of the simulation results to real-world situations. The real-world data is selected based on ideas from environmental criminology and concrete empirical findings.

2.4. Research gap

Given all of the above, we ask: Will a theory-informed, data-driven crime simulation generate patterns similar to real crime patterns in a urban environment? What is the contribution of different types of data sources that are processed in a data-driven manner? In this article, we explore the effectiveness of modeling the decision-making process of offenders deciding whether to commit crimes while using existing data-driven strategies to inform offender mobility. The process is theoretically informed and data-driven, using large amounts of static and dynamic open data for assessing the contribution of different types of data (spatial vs temporal) to predict crime patterns.

3. Data

This article uses openly available urban data for NYC to instantiate a realistic environment. The data includes NYPD complaint data (i.e. crime data), census tract and street segment data, LBSN data, taxi trip data, weather data, land use information, population density, points of interest, calls for service (i.e. 311 calls), public transportation stops, tree census data, and school locations. Details about all data sources can be found in Appendix A; in the following, we briefly describe the main ones. The crime data consists of robberies from June 2014 through June 2015, whereby data from June 2014 through May 2015 is used for model training and validation, and data from June 2015 is used for testing the

[2] Broken windows theory describes the vicious cycle of how visible signs of crime, anti-social behavior, and civil disorder encourage further crime and disorder

final model. The calls for service are non-emergency calls to public servants (which do **not** include calls to report crimes) and are used as a dynamic proxy for the spatio-temporal state of street segments. This dataset is used to inform the offenders' decision of whether to commit crimes at certain locations. The LBSN data is extracted from Foursquare,[3] used by more than 50 million active users who created over 122 million check-ins in NYC by the collection date. Venues and check-in counts from Foursquare are used as a proxy for locations attracting people. This dataset informs agent movement together with the taxi data, which includes taxi trips from origin to destination and is aggregated to compute flows from one census tract (CT) to another. This dataset is used as a proxy for populations flows towards areas with higher activity levels.

4. Methods

This section will introduce the simulation model, the different simulation scenarios, and the way in which model performance will be assessed.

4.1. Model overview

As one can observe in Fig. 1, the model is composed of layers of the environment and agents interacting with the environment by (1) moving and (2) deciding whether to commit crimes (which is the focus of our work). The following gives a short overview of these components and other relevant information before the remainder of the section outlines the various components in detail.

1. In the simulation, virtual **offender agents move** along the street network while engaging in routine activities. The movement of the agents is strongly based on previous work (Rosés et al., 2020) exploring how to build offender mobility rules using large-scale, real-world data. Agents start each day at their home location and travel between their home location and a number of activity nodes. The agents' movement choices are informed by taxi trip data, used to estimate destination neighborhoods, and LSBN data, used to estimate destination venues within neighborhoods. The previous model determined the use of the aforementioned datasets as well as the optimal number of agents (225) for this type of simulation in NYC.[4] Details regarding the agents' movement strategies are outlined in Section 4.2.
2. While traveling, the **agents decide in a rational manner whether to engage in criminal activities** at each street segment. Agents perceive their immediate environment and process the information in a data-driven manner by means of an (1) internal decision tree, (2) calculating the difference in expected crime counts due to temporal variations on a daily basis (using a negative binomial regression), and (3) counting the simulated crimes at street segment level, accounting for the effect of near-repeat. The agents use these three mechanisms to calculate a probability and make a binary choice as to whether to commit a crime on the street segment. The agents' decision to commit crimes is the focus of this current study and described in greater detail in Section 4.3.

Using Mesa, a Python agent-based model framework (Masad & Kazil, 2015), a simplified version of NYC with the various data layers is instantiated. The simulation runs over 30 epochs, whereby one epoch (i.e. step) corresponds to one day. Moreover, each simulation is run 10 times to diminish the bias of fixed and statistically informed offender starting (i.e. home) locations. The results of multiple simulation runs are aggregate to evaluate the results. Aggregating multiple simulations contributes to capturing the additional uncertainty introduced by the stochasticity in the randomness of the agents' starting locations.

In the following subsection, we engage in a more detailed discussion about the agents' movement process and the agent's decision of whether to commit crimes.

4.2. Agent movement

The movement of the offender agents (see Fig. 2) is data-driven and strongly based on Rosés et al. (2020). The agents perceive three layers of the environment:

- the street network layer allows the agents to perceive the street network and process travel paths using Dijkstra's Shortest Path algorithm.[5] It also includes basic information about the street segments from census data, such as the number of buildings on the street segments and population counts for the specific census tract mapped to the segments.
- the census tract (CT) flows layer, created by aggregating taxi trips over census tracts, is used to decide which CT to travel to when agents conduct their routine activities;
- the activity nodes layer represents a proxy for areas attracting visitors in the context of crime pattern theory and is derived from Foursquare check-ins, allowing agents to decide on a particular venue, within a CT, to actually visit (i.e. activity destination).

Movement decisions are based on the assumption that robbers move in similar ways to regular citizens[6] (Brantingham & Brantingham, 1993) and are as follows. Agents are first assigned a home location, which they always return to by the end of the day. The home location is assigned to any street segment with residential buildings. Virtual offenders start at a street segment with residential buildings on it. Therefore, a street segment within a census tract with a higher residential population (census data) has a higher probability of acting as a starting location. Next, the agents decide on a number of locations (or destinations) to visit during the current day (3.8 locations on average, drawn from a uniform distribution), in representing a form of routine activities. Agents do this in two steps: (1) they choose a census tract (CT) to travel to by weighting CT with higher taxi inflow more strongly (CTs with higher taxi flows originating at the agent's home CT are more likely to be chosen). This provides the agents with a perception of the connectivity and a proxy for flows between CTs; (2) After choosing a CT, the agents then decide on a concrete activity node on a street segment within that CT as a destination by using the activity node layer derived from Foursquare venue popularity (check-ins). Venues with a higher popularity are more likely to be chosen as the concrete destination. The agents travel to locations using Dijkstra's Shortest Path algorithm, taking into account the length of the street segments. This process is repeated for each day of the simulation. To account for path dependency, the simulation is run multiple times and aggregated results are evaluated (see Section 4.5 for further details).

4.3. Agent decision to commit crimes

Fig. 3 provides an overview of the decision-making process. The agents perceive their immediate environment through three different layers: a spatial layer, a temporal layer, and an interaction layer.

The **spatial layer** describes the *environmental backcloth* (Brantingham & Brantingham, 1993): the street network, human activity patterns

[3] http://www.foursquare.com/

[4] This previous simulation was performed for NYC with robbery data for June 2015.

[5] https://networkx.github.io/documentation/stable/reference/algorithms/shortest_paths.html

[6] We depart from the notion that crime is a legal definition and does not necessarily define distinct group behavior (Tappan, 1947)

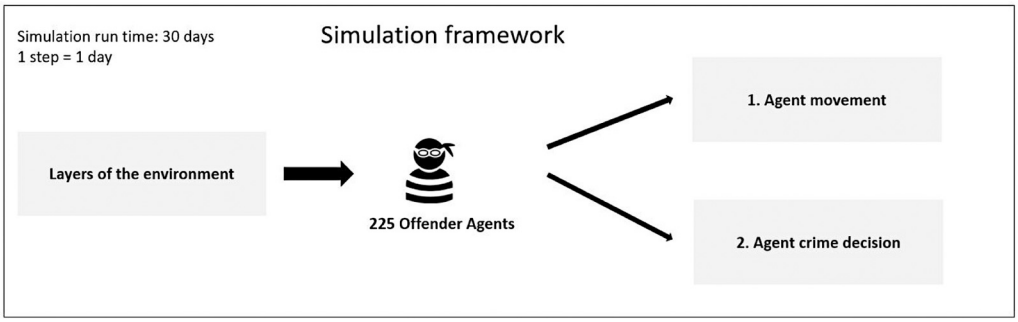


Fig. 1. Model framework.

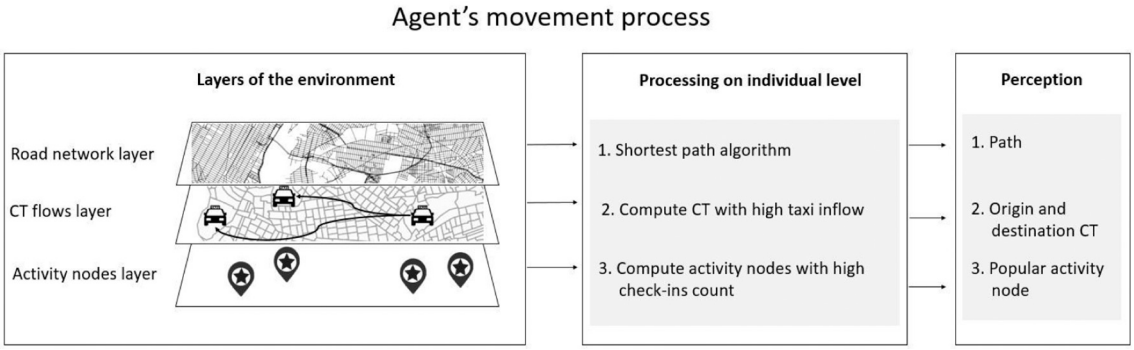


Fig. 2. Process of agent movement with three layers containing spatial data, the processing of the layers at individual level, and the perception of the agents.

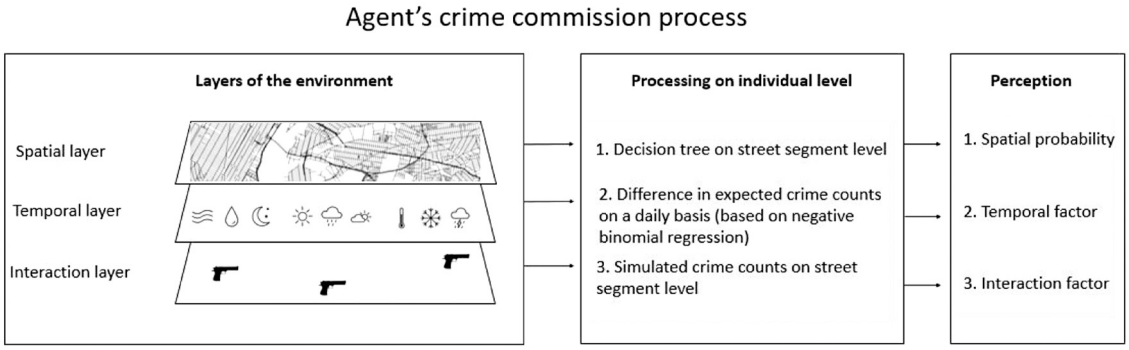


Fig. 3. Process of agent deciding whether to commit crimes in three layers containing the layers of the environment, how individual agents processed these layers, and what individual agents perceive.

from LBSN, land use information, points of interest, schools, public transportation stops (bus and subway), urban trees, population density, and calls for service (including emergency and non-emergency calls to any public service).[7] Since the calls for service are a dynamic feature with daily granularity, we aggregate the call counts over each month for each street segment. Each feature is mapped to a specific street segment. Agents process the information in the spatial layer and decide whether to commit a crime or not by means of a decision tree (full details on the implementation of the decision tree are provided in Appendix B). Ultimately, each agent calculates a *spatial probability*, S.

The **temporal layer** is a proxy for the presence of victims on a given day with specific weather characteristics (e.g. on a rainy winter Saturday there might be fewer potential robbery victims present in NYC). It is built using the day of the week and weather data with information about cloud cover, moon phase, precipitation intensity, precipitation accumulation, precipitation type, pressure, wind speed, humidity, UV index, visibility, average high temperature, and season of the year. As the temporal data is invariant for the whole area of NYC and is on a daily granularity, the layer is processed in a different manner as compared to the spatial data, allowing the agents to obtain information about the availability of opportunities given the temporal conditions.

A regression predicting daily crime counts for the whole of NYC given the temporal conditions is built every day. Then the predicted crime counts are compared to the median crime counts from the previous days. The result gives an idea of the extent to which the current day differs from the median of the previous days and, thus, whether more or fewer potential victims will be present on the given day. The result is used as a multiplier for augmenting or diminishing the spatial crime probability, i.e. the resulting temporal multiplier is combined

[7] In a realistic setup, the calls for service would not be available for the same month (i.e. before the call happens). Therefore, in an ideal scenario, this model would include a prediction of calls for each month within the simulation (Zha & Veloso, 2014). In the interest of simplification, we have used actual 311 calls for each simulated day.

with the spatial crime probability. The idea behind this is that more extreme predicted variations from the mean should have a greater effect on the commission of crimes: the higher the deviance of the predicted crime count from the mean of the previous days, the higher the impact on the decision on whether to commit crimes for the particular day (Tompson & Bowers, 2015).

The temporal model is initialized at the beginning of the simulation using previous data from the temporal layer up to the start date of the model, and is updated for every new simulation day using an additional data point (for the previous day) from the test dataset. Thus, the temporal model is always up to date and uses the maximal amount of information available to predict crime counts for the next day. Full details on the implementation of the regression are provided in Appendix C.

In summary, the regression for each day predicts the counts of the temporal features for the next day (*PC*), and compares the output with the median crime counts for all the previous days (*MC*); see Eq. (1). As a result, the agents obtain a *temporal multiplier* (*T*).

$$T = PC/MC \tag{1}$$

The **interaction layer** takes into account the crimes previously committed by any agent within the simulation, allowing the agents to perceive the actions of other agents and thus affecting their decision regarding whether to commit crimes. This interaction is based on near-repeat theories (Johnson et al., 1997) and empirical research (Glasner & Leitner, 2017). Crimes committed on a street segment create a multiplier effect, increasing the probability of a subsequent crime occurring. We refer to this value as the *interaction value*, *I*, which is set to 0.5 for the first day and subsequently halved for the following two days, after which the effect disappears.

In summary, we have outlined the three elements that the agents need to combine to obtain the crime probability: the spatial probability *S*, the temporal multiplier *T*, and the interaction value *I*. These are combined in the following manner to obtain an overall crime probability *P* (i.e. the probability that an agent will commit a crime at a given street segment):

$$P = S^{*}T + I \tag{2}$$

Given *P*, the optimal threshold for the binary robbery decision is unknown. Therefore, we define a threshold parameter, which we will calibrate to find the optimal value. All calibration is done by training the decision tree with data for 11 months (June 2014–April 2015) and then calibrating the simulation on data for May 2015. After calibration, the model is run for the next month, June 2015, with the optimal parameters resulting from the calibration. For more details, see Section 4.5.

4.4. Scenarios

To study the impact of data-driven decision-making on the outcome of the simulation, we split the crime decision process into three separate steps and progressively build scenarios by adding one data layer at a time.

- **Scenario 0 – Baselines**: The first scenario is meant to set a baseline for comparison to the more comprehensive models. Two baselines are built in order to determine whether including open data as environmental layers: (1) is useful for informing an agent's crime decision, and (2) increases model performance as compared to a simple model based on past crime data. Therefore, *Baseline I* is built using the simulation model with agents deciding randomly whether to commit a crime at each street segment. At each street segment the agents have a 2% chance of committing a crime, which reflects the number of positive crime examples in the crime dataset. *Baseline II* is built solely using past crime data from May 2014 (i.e. data from the same month of the previous year) with no model.
- **Scenario 1– spatial data and machine learning**: The spatial scenario provides the first data-driven internal decision-making process for the agents. In this scenario, the agents decide whether to commit crimes along their path, perceiving the spatial layer only.

  As agents move, the rules from the decision tree are applied to calculate *P* of committing a crime on each street segment, based purely on the spatial layer:

$$P = S \tag{3}$$

- **Scenario 2 – spatio-temporal data and machine learning**: This extends the previous scenario by including the temporal layer, allowing agents to additionally perceive daily weather and weekday variations:

$$P = S^{*}T \tag{4}$$

- **Scenario 3 – interaction with spatio-temporal data and machine learning**: The interaction scenario adds the interaction between the agents, thus taking all layers of the environment into account. Crimes committed affect the crime decision process of other agents when they process the given street segment.

$$P = S^{*}T + I \tag{5}$$

4.4.1. Performance assessment

The performance of the different scenarios is evaluated in terms of the Predictive Accuracy Index (PAI) (Adepeju, Rosser, & Cheng, 2016). PAI is a well-known metric in criminology for assessing the performance of crime prediction models (Chainey, Tompson, & Uhlig, 2008; Rummens, Hardyns, & Pauwels, 2017). See Eq. (6), where *hit rate* (*hr*) is the percentage of predicted crimes within the prediction area and *area percentage* (*a*) is the prediction area in relation to the whole study area. The value of PAI is one if a model predicts all of the crimes in the entire study area, while it can be greater than one if the selected coverage area is small. The area percentage can be understood as the area the police can cover (i.e. coverage area). Typical values for coverage area range from 3% to 20% (Adepeju et al., 2016; Chainey et al., 2008; Rummens et al., 2017). This article aims to achieve the highest PAI index for a coverage area of 3%, while we also report results for coverage areas of 5% and 10%. This is in line with *crime at places* literature, which emphasizes the importance of looking at crimes in micro-places. In this sense, a small number of street segments may account for a large number of crimes within a urban environment (Lee et al., 2017).

$$PAI = \frac{hr}{a} \tag{6}$$

Additionally, we compare the aggregated simulated crime pattern to the real crime pattern on two different levels: street segments and census tracts (CT). We report:

- whether the relative counts of crimes per road coincide using the mean squared error (RMSE_1);
- whether the relative counts of crimes per CT coincide using the mean squared error (RMSE_2);
- the number of roads that have had at least one crime using the area under the ROC curve (AUC_ROC) as well as error rate (ER)

We specifically look at relative crime counts by normalizing crime counts over 1 (using the MinMaxScaler[8] function from scikit-learn Python package), as we are not interested in absolute crime counts but rather in finding the roads or CTs that are more at risk for robbery.

[8] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

4.5. Model calibration

To calibrate the abovementioned parameters within the various scenarios, we run each scenario a minimum of eight times for May 2015 (unseen data) and aggregate the results to find the averaged optimal parameters. Within the calibration step, aggregating eight runs leads to small standard deviation in the performance measures, meaning that the number of runs is sufficient.[9] Then we assess the performance of each parameter in terms of the highest PAI (3%). Thus, we prioritize predicting the 3% of road segments at the highest risk for crime in the next month. Once we have the optimal parameters resulting from the calibration, we proceed to run the model for June 2015. The number of model runs, with the optimal parameters, is set to 10, which is reasonable given the running time of around 4.5 h per model.

5. Results

For the purpose of assessing the performance of the various scenarios described in the previous section, after calibration, we run each simulated scenario 10 times and report the average results. In this manner we account for small variations due to path dependency, especially as related to the agents' starting positions. The standard deviation of the different performance values is small and we therefore conclude that aggregating the results over 10 runs is enough.

5.1. Baseline

Baseline I simulates offender agents deciding randomly with a 2% probability to commit a crime at each traveled street segment. This model achieves a PAI (3%) of 1.45 and a RMSE_1 of 0.065 (see Table 3). In Fig. 4(a) we can see that crimes for Baseline I are rather more randomly distributed over the different areas of the urban environment, while somewhat clustered around paths taken by the agents. This model shows the extent to which agent movement affects the crime pattern; agent movement alone is not able to predict the real crime pattern. Baseline II, in Fig. 4(b), is the robbery pattern for June 2014 (the previous year) with no model and represents an informed baseline. The PAI (3%) value at street segment level is 13.31, which is very high, while RMSE_1 is 0.052, lower than for Baseline I. In contrast to the two baselines, Fig. 4(c) shows the crime pattern which this model attempts to predict (June 2015).

5.2. Scenario 1

The first scenario includes static and dynamic spatial data in the simulation environment. Recall that the output of the agents' decision tree is a probability that needs to be converted into a binary variable to determine whether the agent engages in crime (1) or not (0). To find the optimal cut-off threshold for this conversion, the simulation was run for May 2015 with varying threshold values between 0.1 and 1. The best-performing parameter is for a threshold of 0.9, achieving a PAI (3%) of 10.11 (and RMSE_1 = 0.040).

We run the scenario using a threshold of 0.9 to predict crimes for June 2015 and achieve a PAI (3%) of 10.02 for predicted crime counts at street segment level (see Table 3 and Fig. 6(a)). Additional insight can be gained by exploring the contribution of the different spatial features to the decision-making process by looking at the feature importance for the decision tree (see Fig. 5.). The most important features are the check-ins on the current street followed by the 311 calls and the sum of check-ins within 400 ft. of the current street. Street traffic and POI on the same street do not contribute to the model.

[9] With a PAI [3%] standard deviation between 0.024 and 0.085, which is lowest for Scenario 2.

5.3. Scenario 2

In the second scenario, the agents perceive and process the spatial data as in Scenario 1 and additionally perceive temporal data related to the day of the week and weather conditions. The temporal multiplier ranges from 0.946 to 1.116 for the month of June 2015. To estimate the crime/no crime threshold value, we again run the simulation for May 2015 and test threshold values ranging from 0.1 through 1. The best results are achieved by a threshold of 1: PAI (3%) of 10.25 and a RMSE_1 of 0.040. The model run for June 2015 achieves a PAI (3%) of 10.24 and RMSE_1 of 0.40 (see Table 3 for more results and Fig. 6(b) for a visualization of the output).

To gain more insights into the contribution of single temporal features to the temporal multiplier, we explore the significance of the features for the negative binomial regression. On average, the most significant features across the daily linear regressions are temperature, winter, and temperature combined with autumn (see Table 2). Higher temperatures and the season being winter or autumn are positively related to robbery (i.e. higher crime counts). Additionally, elevated temperature combined with autumn is negatively related to robbery, meaning that a higher temperature in autumn leads to lower robbery counts in the urban area. Furthermore, visibility in combination with spring is positively related to crime, meaning that higher visibility in spring leads to higher crime counts.

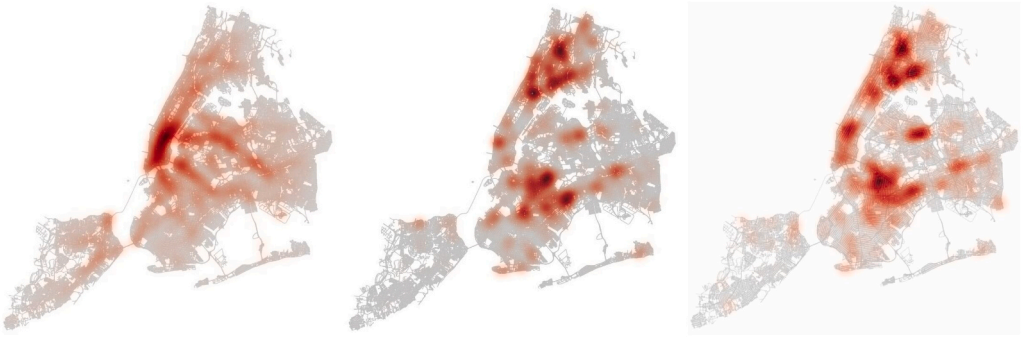
5.4. Scenario 3

The third scenario includes interaction between the agents. This is the ‘full’ version of the model. This scenario achieves a PAI value of 10.25% for a coverage area of 3%, with RMSE_1 of 0.040 (see Table 3 and Fig. 6(c)) with a threshold of 1.

5.5. Scenario performance and comparison

In this subsection we proceed to compare the results of the different scenarios in terms of PAI. In Table 3:, described in the previous section, we have seen that Scenarios 2 & 3 perform best in terms of PAI (3%) when predicting crime counts at street segment level, while Baseline II still performs slightly better in terms of PAI (3%) but not in terms of RMSE_1. Furthermore, Scenario 3 performs best when PAI (3%) is assessed at CT level, almost comparable to Baseline II. To assess model performance, we are interested in the highest PAI for a 3% coverage area. 3% of street segments in NYC represents more than 300 street segments, an appropriate number to be targeted by crime reduction initiatives. The coverage areas (i.e. prediction areas) can be understood as the percentage of the area the police can cover. Thus, PAI returns a measure for whether the top x percentage of street segments coincide in terms of the simulated and real crime patterns. To get a better notion of the value of this simulation, we also look at PAI for various coverage areas between 5 and 20%. Additionally, we look at the hit rate, the percentage of accurately predicted crimes within the prediction area (see Table 4). Thus, we get a wider sense of how well the different scenarios perform.

Across all PAI values, the three scenarios perform significantly better than the random baseline. This shows the value of adding a spatial environment, a temporal multiplier, and agent interaction to the simulation. Scenarios 2 and 3 perform best and very similar to each other in terms of PAI (3%) and PAI (5%). Indeed, Scenario 3 outperforms the baseline at a coverage area of 3% (hit rate = 30.75, PAI = 10.25) and Scenario 1 outperforms at coverage areas of 10%, 15%, and 20% (e.g. hit rate = 90.23, PAI = 4.51). Thus, the temporal multiplier and interaction contribute to predicting the 3% of street segments with highest risk for crime, while the spatial data is more robust in predicting the most prolific street segments up to 20%. In turn, Baseline II achieves a PAI (3%) of 13.31, which is higher than any scenario. Baseline II performs best in terms of PAI overall, while its RMSE_1 is slightly higher,



(a) Baseline I (b) Baseline II (c) Crime occurrences June 2015

Fig. 4. Crime patterns heat map for Baseline I, Baseline II, and real crime occurrences for June 2015.

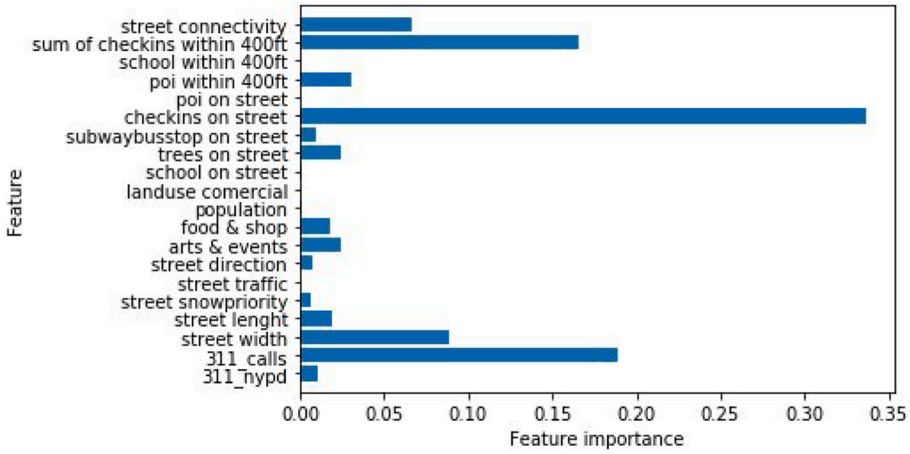
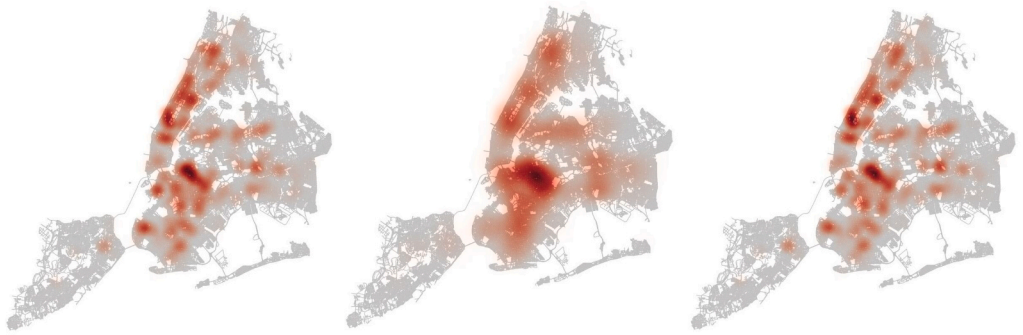


Fig. 5. Scenario 1: decision tree feature importance.



(a) Scenario 1 – spatial data. (b) Scenario 2 – spatial and temporal data. (c) Scenario 3 – interaction with spatial and temporal data.

Fig. 6. Crime pattern heat map for Scenarios 1, 2, and 3.

indicating a larger error in predicting the crime counts at street segment level. Additionally, the AUC_ROC and the recall values are lower compared to the three scenarios, indicating less precision in predicting the street segments that have at least one robbery. In general, AUC_ROC shows very low discriminative power for all scenarios, including the baselines. This is because when categorizing street segments into crime (1) vs no crime (0), the important information about how prolific (multiple crimes) a street segment is, is lost.

6. Discussion

This article has highlighted the value of combining various data sources into an ABM to generate crime patterns. The resulting model can be used by practitioners engaging in crime prevention as a tool to predict street segments at higher risk for future robberies. To our knowledge, this is the first ABM for crime utilizing a data-driven approach integrating ML to inform agent behavior. Moreover, the model has

Table 2
Negative binomial regression with temporal features for 1. June 2015.

Feature	Reg. coeff	*p*-Value
Intercept	4.39	0.04
Snow	−0.10	0.10
Rain	−0.06	0.31
Moonphase	0.04	0.27
Precipintensity	−0.05	0.38
Precipaccum	0.00	0.81
Humidity	0.21	0.13
Pressure	0.00	0.77
Windspeed	−0.01	0.78
Temperaturehigh	0.02	**0.00**
Weekend	0.05	0.12
Winter	0.43	**0.00**
Spring	−0.19	0.44
Summer	0.29	0.36
Autumn	0.63	**0.04**
Temperature*spring	0.00	0.55
Temperature*summer	−0.01	0.20
Temperature*autumn	−0.01	**0.02**
Visibility*spring	0.02	**0.04**
Visibility*summer	0.00	0.93
Visibility*autumn	−0.01	0.57
Windspeed*spring	0.02	0.68
Windspeed*summer	0.00	0.93
Windspeed*autumn	0.00	0.90
Snow*spring	0.08	0.57
Snow*summer	0.00	0.26
Snow*autumn	−0.13	0.37
Rain*spring	0.05	0.50
Rain*summer	0.02	0.72
Rain*autumn	0.00	0.99

integrated static and dynamic features representing different layers (spatial and temporal), which has not been done previously. If such models lead to more accurate simulations generating spatial patterns comparable to real crime patterns (as opposed to theory-driven simulations), this would boost the potential for this technique to be used in practice for crime prevention purposes. The main potential of using simulation techniques is that it is possible to perform experiments that are unfeasible in real life (Groff et al., 2018). For instance, ABMs can be used by police departments to test the effect of patrolling strategies on crime or they can be used by urban planners to assess how changes in urban design may affect the development of crime. For ABMs to be used for these purposes, they need to accurately represent crime patterns in a realistic manner. Thus this research advances the use of simulation for experimental studies of crime.

The model described in this article achieved predictive values on the higher end (PAI 3% = 10.25, PAI 20% = 4.51) compared to other existing research using machine learning techniques: Adepeju et al. (2016) reaching a PAI (20%) of 2.99 for violence and 4.58 for shoplifting; Rummens et al. (2017) reporting a PAI (6%) value reaching 13.80 for street robbery and 3.77 for battery; and Chainey et al. (2008) achieving a PAI (3%) of 6.59 for street crime. Predictions in this article are made on a lower level, i.e. street segments instead of larger areas (e.g. grid cells of 250 km^2) (Chainey et al., 2008). It is highly notable that the results for prediction at street segment level in this article are comparable to prediction models in literature on higher units of analysis. Additionally, this article has implemented simulation techniques integrating ML, as opposed to solely utilizing ML, making the resulting model applicable to a wider range of use cases. Thus, we conclude that the technique developed in this article is valuable for building data-driven ABMs for crime prediction and especially advances the field of computational criminology integrating GIS, real data, and ML. Furthermore, the simulation model shows consistent results with significant improvement over the random baseline and similar results to the informed baseline (built using crime data for the previous year). High crime areas are relatively stable over time, especially the approximately 5% of places (such as street segments) that account for 50% of recurrent crimes (Sherman, Gartin, & Buerger, 1989; Weisburd, Bushway, Lum, & Yang, 2004). These street segments often exhibit characteristics which make them prolific sites of criminal activity. For this reason, the informed baseline performs well compared to the random baseline and even compared to the simulation results. We have investigated the benefit of adding different static and dynamic data sources by means of ML techniques. Compared to a random baseline, adding spatial data (first scenario) to the environment and having agents processing the data by means of a decision tree yields a further benefit. Moreover, within this scenario, the most important features are the check-ins on the current street followed by the 311 calls and the sum of check-ins within 400 ft. This is in line with other research linking 311 calls (Wheeler, 2018), LBSN check-ins (Kadar & Pletikosa, 2018) and street network connectivity (Kadar, Feuerriegel, Noulas, & Mascolo, 2020) with crime. The second scenario, combining spatial and temporal (weather and day of the week) data by means of a binomial regression, does add a slight improvement to the predictive power of the simulation for small coverage areas (3% and 5%). For this scenario, high temperatures, the season of winter, high temperatures in autumn, and good visibility in spring contribute significantly to the temporal model. This is in line with findings by other researchers testing the influence of weather on robbery (Tompson & Bowers, 2015). The third scenario,

Table 3
Aggregated performance results for counts at street segment level, counts at CT level, and binary at street segment level.

Scenario	Segments count		CT count		Segments binary		
	PAI	RMSE_1	PAI	RMSE_2	AUC_ROC	ER	Recall
Baseline I	1.45	0.065	1.67	0.130	0.555	0.210	0.239
Baseline II	13.31	0.052	3.69	0.138	0.500	0.030	0.080
Scenario 1	10.02	0.040	3.45	0.139	**0.512**	0.030	0.035
Scenario 2	10.24	0.040	3.33	0.137	0.510	0.030	0.029
Scenario 3	**10.25**	0.040	**3.59**	0.138	0.510	0.030	0.027

Table 4
Aggregated hit rate and PAI results at street segment level. Higher hit rate and PAI values indicate better model performance.

Scenario	3% coverage		5% coverage		10% coverage		15% coverage		20% coverage	
	hit rate	PAI	hit rate	PAI	hit rate	PAI	hit rate	PAI	hit rate	PAI
Baseline I	12.27	4.42	24.89	4.98	35.79	3.58	41.14	2.74	52.34	2.61
Baseline II	39.93	13.31	63.19	12.64	83.17	8.32	89.42	5.96	92.48	4.62
Scenario 1	30.07	10.02	56.40	11.28	**80.79**	**8.08**	**86.74**	**5.78**	**90.23**	**4.51**
Scenario 2	30.73	10.24	**57.24**	**11.45**	80.49	8.05	86.58	5.77	89.80	4.49
Scenario 3	**30.75**	**10.25**	57.16	11.43	80.50	8.05	86.57	5.77	89.81	4.49

including interaction in the form of near-repeat, performed best for a coverage area of 3%. Thus, interaction contributes most to predicting the most criminally prolific 3% of street segments, while the weather multiplier contributes most to predicting the most prolific 5% of street segments, and spatial data is more robust in informing the most prolific street segments constituting up to 20% coverage area.

The simulation model has limitations, however, especially in relation to the data used. First of all, the study focuses on New York City as a case study, and we acknowledge that this urban environment may exhibit characteristics that make it a special case. Furthermore, crime data is based on police records, and thus contains only robberies known to the police, rather than covering the full spectrum of robberies (i.e. we predict robberies known to the police). More importantly, it is well known that the distribution of reported crimes may vary by type of neighborhood (e.g. less reporting in more deprived areas, where trust in police is fragile), and thus unreported crimes are not evenly distributed across an urban environment. Therefore, researchers assume that the pattern of known crimes may differ from that of crimes unknown to the police. The previous also applies to offender home locations and the use of known offender's home location for modeling crime patterns. In this sense, this work does not take into account any specific information to derive offender starting location other than population density and aggregating multiple simulation runs to diminish the bias of a specific choice, limiting the information which could be gained from more accurate and informed offender "home" locations. Additionally, this simulation uses LBSN data, which shows great potential as a measure for the presence of potential victims (i.e. ambient population), but is inherently biased towards younger users and does not accurately depict activities undertaken by the whole population. Other researchers have already identified the potential and risks of using crowd-sourced data for crime analysis (Malleson & Andresen, 2014; Malleson & Andresen, 2015). Future research could look more carefully into the contribution of LBSNs for simulating crime and add additional information sources to overcome bias in the data. In terms of environmental data, we suggest that future models should include a larger variety of dynamic data capturing the spatial changes in the environment at street segment level. In terms of the data concerning calls for service (311 calls), we also acknowledge a bias in the type of people reporting to the public servants. Additionally, we have used real 311 calls for the same day for which we predict crime. In a realistic scenario, the model should produce a prediction for 311 calls for the same day, as the 311 calls for the same day are not available. Future work could include a more realistic representation of 311 calls.

In terms of modeling, crime data is inherently sparse, especially when building prediction models on a small unit of analysis. Therefore, we have under-sampled the majority class to train the decision tree for the spatial data. We suggest that future work should use informed sampling instead of random under-sampling. In this sense, paths taken by the agents throughout the previous month could inform sampling and lead to individually fitted decision trees. Thus the agents would learn from the street segments they travel and build individual decision-making models in line with RAT. Moreover, this would create heterogeneous agents in the simulation.

Finally, although this research attempts to reflect widely accepted environmental criminology theories, there are inevitable weaknesses in the theoretical foundations. One of these is the absence of guardians (Cohen & Felson, 1979). Guardianship is likely to have an impact on whether or not an individual robbery takes place, as 'capable guardians' may intervene or deter offenders by their very presence before any crime has been committed. Guardianship can be included by modeling individual guardians directly [e.g. (Groff, 2007b; Groff, 2007c)] or through aggregate neighborhood- or community-based measures [e.g. (Malleson et al., 2010)]. Future work can begin to include measures of guardianship as appropriate.

7. Conclusion

This article has explored building a theory-informed, data-driven ABM generating realistic crime patterns. The model has put a special focus on the contribution of different types of data sources (spatial, temporal, and interactive) in combination with ML techniques to process this data at individual agent level. We conclude that the combination of spatial, temporal, and interaction layers contributes to predicting the most criminally prolific 3% of street segments. These findings are relevant because they highlight the significance of spatial data and the potential of data-driven simulations for crime prediction purposes and advance the field of computational criminology by exploring the integration of real data and ML techniques to study crime patterns.

Appendix A. Data

This appendix provides a detailed description of the data used in the article. NYC is the most densely populated city in the United States with around 27,000 people per square mile and over 8.5 million inhabitants altogether. NYC is known for high levels of activity within the city and provides a good opportunity to study urban behavior like crime. Thus, a large variaty of openly available data sets is avalable to instantiate NYC's urban environment with information about human activities. The majority of this data has been downloaded from the NYC Open Data portal[10] or other governmental open data platforms. The datasets including geographical information, which can be downloaded as shapefiles and have been pre-processed in Postgres.

The most important dataset for the model is the **New York Police Department (NYPD) complaint data.**[11] It contains felony crimes reported to the police with information such as type of crime, date, time, and location [lat/long point coordinates at the middle or any end of a street segment]. The article only analyses robberies and within a period of 12 months (Jun 2014 through May 2015, with 16,413 robberies) for model training. Long-term past crime data (e.g. at least 12 months) is a good indicator for future crime (Groff & La Vigne, 2002), and robberies can successfully be modeled using openly available data sources (Kadar & Pletikosa, 2018; Rosés et al., 2018). One month crime data (June 2015, with 1303 robberies) serves form model testing. This dataset contains the target variable, to be predicted by the simulation, and allows us to validate the simulation.

Census tracts (CT) are spatial units subdividing counties [lat/long polygon coordinates] used for statistical data collection and defined by the United States Census Bureau[12] for the New York region. This dataset containing CT shapes is used to assess the performance of the model on an aggregated level, i.e. to see how the performance of the simulation varies by neighborhood. In NYC there are 2168 CTs with populations ranging from 3000 to 4000 and an average land area of 90 acres. The dataset has been cleaned to remove 6 CTs containing only water and shorelines.

Street segments are the spatial unit of analysis in the simulation. The **street segments** dataset[13] used in this study contains street centerlines with information such as street segment length, width, traffic direction, snow priority (level of priority if snow needs to be removed from the street segment), as well as shape and location of the segment line [lat/long line coordinates]. The street segment line shapes are used to build a street network layer utilized by the agents within the simulation to move from one location to another. An additional feature is built for street connectivity, with a count of the number of connecting streets for each

[10] https://opendata.cityofnewyork.us/data/

[11] https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243/data

[12] https://data.census.gov/cedsci/

[13] https://data.cityofnewyork.us/City-Government/NYC-Street-Centerline-CSCL-/exjm-f27b

segment. Thus the street network instantiates the urban structure and the connectivity between places.

LBSNs are social networks that broadcast users' locations and other content from their mobile devices. Data from Foursquare, a popular LSBN, can be used as a proxy for the popularity of locations. Foursquare data was collected from the Foursquare API (Application Programming Interface)[14] in June 2016, as in Kadar, Iria, and Pletikosa Cvijikj (2016), Kadar, Rosés Brüngger, and Pletikosa Cvijikj (2017) and Rosés et al. (2018). The dataset contains information such as venue name, location [lat/long point coordinates], check-in counts (accumulated over time), and categories (e.g. arts, college and university, events, food, nightlife, shops and services, traveling and transportation). This dataset contains 236,294 venues and over 119 million check-ins in the proximity of every incident from the crime dataset (from the venue creation date in the platform until the collection date in June 2016). This data is used to simulate the popularity of locations, including information regarding the context of the activities at these locations. Context is extracted form the associated venue categories (i.e. types).[15] Indeed, since not all venue categories are useful for this simulation, three new categories aggregating single categories have been created: artevent (arts, events, nightlife, outdoors, and recreation) as a proxy for leisure activities, foodshop (food, shops, and services) as a proxy for services, and collegeprofessional (college, professional) as a proxy for professional activities.

The **taxi trips** data combines Yellow Taxi Trip Data[16] and Green Taxi Trip Data[17] into one dataset for a one-year time period (July 2014 to June 2015). Both taxi services pick up passengers hailing from the street and, combined, cover all of NYC. Yellow taxis are concentrated around Manhattan as well as the JFK International Airport and LaGuardia Airport, while green taxis offer their services above 110th Street in Manhattan and in the outer boroughs of NYC. The dataset is composed of over 248 million taxi trips within a span of 12 months. It includes pick-up and drop-off dates/times/locations [lat/long point coordinates]. Pick-ups and drop-offs are aggregated over CTs, creating a new dataset pairing CTs with each other and weighting by total number of pick-ups and drop-offs from and to each census tract. This new dataset reveals information about the connectivity and popularity of transitions from one CT to any other CT in the city and is a proxy to instantiate the movement of offenders (who are known to move similarly to potential victims).

The **weather** dataset has been built using the darksky API,[18] which provides a wide variety of weather information. The data has been downloaded on a daily basis (daily average) for the same period as the existing crime data and includes information on cloud cover, moon phase, precipitation intensity, precipitation accumulation, precipitation type, humidity, pressure, wind speed, UV index, visibility, highest temperature, etc. Weather influences the presence of potential victims and the types of activities in which they engage.

Land use information from the PLUTO[19] database (2016) is used to identify the types of land use. The dataset covers three categories: commercial areas, residential areas, and mixed areas at census block level (this unit is smaller than the CT and contains one block). Land use information can be used as a proxy for the presence of potential victims in the urban environment.

To create a dataset of **population density per census tract** (CT), the NYC American Community Survey Public Use Microdata Sample (ACS 2011–2015 5-year PUMS)[20] is used. The census tract population total for the period from 2011 to 2015 is extracted and population density for each CT is calculated. The ACS PUMS data provides information about the populations within the different census tracts. ACS 2016 PUMS data were not yet available when the data for this study was collected; however, the overall change in population from one year to the next is less then 0.5%. Population density is also an indicator for potential victims.

Points of interest (POIs) are locations considered common places by the different city agencies and are provided by the NYC Open Data platform.[21] Features within 80 ft. of a street segment comprise the dataset for this article, resulting in 10,150 POIs close to 7671 street segments. POIs can be used as a proxy for the presence of potential victims.

The **311 call** (or calls for service) dataset from the 311 Call Center Inquiry is a record of all agent-handled calls to the city's 311 information line, with information on the topic of the call as well as the time and date.[22] These are non-emergency calls to public servants addressing a variety of topics such as noise, street conditions, and blocked sidewalks. This dataset contains calls over time, which allows us to build a dynamic spatial feature. For this article this dataset is split into two parts, one for all 311 calls not directed to the NYPD (referred to as"general 311 calls"), and the 311 calls to the NYPD only (many of them about noise or illegal parking). Furthermore, all the calls that can be assigned to a street segment are kept (i.e. within 200 ft of a street segment), resulting in 1,660,132 general 311 calls and 613,202,311 calls to the NYPD for the period of May 2015–June 2015. These calls can be used as an indicator for responsiveness of residents to issues within their neighborhood, and can therefore capture different dynamic aspects of an area.

The **public transportation** dataset contains sheltered bus stops[23] and subway entrances.[24] Public transportation stops are often linked to higher rates of crime because they attract a large number of potential victims who stay at the location for a short but predictable period of time.

The **trees census**[25] is a dataset with the location of each tree as a geographic point, through which allows to gain information about the number of trees on each street segment. Trees and green spaces have been linked to crime in the sense that they sometimes provide a hiding spot for offenders, thus encouraging crime, but can also be an indicator for higher-status neighborhoods linked to lower crime rates. The relationship is therefore not necessarily unidirectional.

The **schools** dataset[26] contains the locations of all schools within NYC. Schools attract a large number of pupils, and thus representing a pool of potential victims.

Appendix B. Details decision tree (Spatial layer)

This appendix provides a detailed description of the Decision Tree.

[14] http://www.foursquare.com/

[15] 8254 venues were not within 80 ft of a street segment and therefore not included in this study.

[16] https://data.cityofnewyork.us/Transportation/2014-Yellow-Taxi-Trip-Data/gn7m-em8n

[17] https://data.cityofnewyork.us/Transportation/2014-Green-Taxi-Trip-Data/2np7-5jsg

[18] https://darksky.net/dev

[19] https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page

[20] https://www.census.gov/programs-surveys/acs/data/pums.html

[21] https://data.cityofnewyork.us/City-Government/Points-Of-Interest/rxuy-2muj

[22] https://data.cityofnewyork.us/City-Government/311-Call-Center-Inquiry/tdd6-3ysr

[23] https://data.cityofnewyork.us/Transportation/Bus-Stop-Shelters/qafz-7myz

[24] https://data.cityofnewyork.us/Transportation/Subway-Entrances/drex-xx56

[25] https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/pi5s-9p35

[26] https://data.cityofnewyork.us/Education/School-Point-Locations/jfju-ynrr

Table B5
Decision tree features as drawn from environmental criminology literature; see also tab:crimefactors.

Feature	Variable	Discarded	Binary
Checkins	LBSN: number of check-ins	no	no
Checkins400ft	LBSN: number of check-ins within 400 ft	no	no
Venues	LBSN: number of venues	yes	no
Commercial	land use	no	yes
Residential	land use	yes	yes
Mixed	land use	yes	yes
Artevents	LBSN: venue type fraction of arts and events in CT	no	no
Foodshops	LBSN: venue type fraction of food and shops in CT	no	no
Nightlife	LBSN: venue type fraction of nightlife in CT	yes	no
Univprof	LBSN: venue type fraction of university and professional	yes	no
Residential	LBSN: venue type fraction of residential	yes	no
Popdensity	population density	no	no
Poi	number of points of interest	no	no
Poi400ft	sum of poi over 400 ft	no	no
calls	number 311 calls except calls to NYPD	no	no
NYPDcalls	number 311 calls to NYPD	no	no
subwaybusstop	at least on public transportation stops	no	yes
Road width	street width	no	no
Road length	street length	no	no
Road traffic	traffic on the street	no	yes
Road direction	one-way or bidirectional traffic	no	yes
Snowpriority	priority for snow removal	no	yes
Trees	number of trees	no	no
Schools	number of schools	no	no
Schools400ft	schools within 400 ft	no	no

To train the decision tree we use data from the previous 12 months (June 2014 – May 2015), aggregated on each month.[27] The dependent variable is crimes at each street segment aggregated over a month, which are then binarized. As some of the environmental features vary by month, each street segment occurs at least 12 times in the training data, once for each month. Segments with more than one crime are duplicated. In total, the dataset contains 1,360,240 instances, while a large proportion of the examples are negative. As a result, the dataset presents a class imbalance. Hence we apply sampling methods to under-sample the majority class (Kadar, Maculan, & Feuerriegel, 2019).

The decision tree for the agents is created following well-established best practice in ML (Hastie, Friedman, & Tibshirani, 2020). It is implemented using the scikit-learn[28] software library for machine learning in Python. We run a grid search with 5-fold cross validation to find the optimal hyper-parameters for the decision tree, varying maximal tree depth, class weight (none or balanced), and criterion (gini, entropy), using the area under the curve (roc_auc) as the scoring metric. We then train a decision tree on all of the past data with the best hyper-parameters resulting from the 5-fold cross validation. This creates the final decision tree, which determines the rules the agents will apply to assess the probability of committing a crime on a specific street segment. See Table B5 for more details on how the spatial features for the decision tree were operationalized. Features correlating with each other by more than 0.6 (absolute value) were discarded and not used within the decision tree. When an agent is on a street segment, it applies the rules resulting from the decision tree to generate a probability. We refer to this probability as the *spatial probability*.

[27] Daily crime counts at street segment level are too scarce to build a decision tree with daily granularity

[28] https://scikit-learn.org/stable/

Table C6
Negative binomial regression features.

Feature	Variable	Discarded	Binary
Clouds	cloud coverage	yes	no
Moonphase	moon phase	no	no
Precipintensity	precipitation intensity	no	no
Precipaccum	precipitation accumulation	no	no
Rain	precipitation type: rain	no	yes
Snow	precipitation type: snow	no	yes
Pressure	pressure	no	no
Windspeed	wind speed	no	no
Humidity	humidity	no	no
Uvindex	UV index	yes	no
Visibility	visibility	no	no
Temperaturehigh	average high temperature	no	no
Weekend	Friday and Saturday vs. rest of the week	no	no
Autumn	autumn	no	yes
Summer	summer	no	yes
Spring	spring	no	yes
Temperature*autumn	interaction between 2 variables	no	no
Temperature*summer	interaction between 2 variables	no	no
Temperature*spring	interaction between 2 variables	no	no
Humidity*autumn	interaction between 2 variables	no	no
Humidity*summer	interaction between 2 variables	no	no
Humidity*spring	interaction between 2 variables	no	no
Windspeed*autumn	interaction between 2 variables	no	no
Windspeed*summer	interaction between 2 variables	no	no
Windspeed*spring	interaction between 2 variables	no	no
Visibility*autumn	interaction between 2 variables	no	no
Visibility*summer	interaction between 2 variables	no	no
Visibility*spring	interaction between 2 variables	no	no
Snow*autumn	interaction between 2 variables	no	yes
Snow*summer	interaction between 2 variables	no	yes
Snow*spring	interaction between 2 variables	no	yes
Rain*autumn	interaction between 2 variables	no	yes
Rain*summer	interaction between 2 variables	no	yes
Rain*spring	interaction between 2 variables	no	yes

Appendix C. Details negative binomial regression (Temporal Layer)

This appendix provides a detailed description of the Negative Binomial Regression. The temporal model is strongly based on Tompson and Bowers (2015) and uses a negative binomial regression to predict crime counts for the whole area of NYC. A negative binomial regression is adequate to predict crime counts in this context because variance does not equal the mean of crime counts, and thus the criterion for a poisson regression is not fulfilled (Hilbe, 2011). For implementation, we use the generalized linear models[29] from the statsmodels software library in Python.

The features for the regression are depicted in Table C6, while we built interaction features combining multiple single features, as in Tompson and Bowers (2015). This gives more valuable insights into how weather may affect crime counts. For instance, higher temperatures may have a different effect during different seasons of the year.

To validate this temporal model against an existing model (Tompson & Bowers, 2015), we build a negative binomial regression based on previous data (training set for June 2014 to May 2015) and predict on the basis of all of the future data points (testing set for June 2015), achieving a pseudo R2 of 0.21, very similar to the existing example (Tompson & Bowers, 2015), with a pseudo R2 of 0.23.

References

Adepeju, M., Rosser, G., & Cheng, T. (2016). Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions - a crime case study. *International Journal of Geographical Information Science, 30*, 2133–2154. https://doi.org/10.1080/13658816.2016.1159684.

[29] https://www.statsmodels.org/devel/glm.html

Ahlin, E. M., & Lobo Antunes, M. J. (2017). Levels of guardianship in protecting youth against exposure to violence in the community. *Youth Violence and Juvenile Justice, 15*, 62–83. https://doi.org/10.1177/1541204015590000.

Al Boni, M., & Gerber, M. S. (2016). Predicting crime with routine activity patterns inferred from social media. In *Proceedings of the 2016 IEEE international conference on systems, man, and cybernetics (SMC)*.

Andresen, M. A. (2011). The ambient population and crime analysis. *The Professional Geographer, 63*, 193–212. https://doi.org/10.1080/00330124.2010.547151.

Beavon, D. J. K., Brantingham, P. L., & Brantingham, P. J. (1994). The influence of street networks on the patterning of property offences. In R. V. Clark (Ed.), *Crime prevention studies, Crime prevention studies* (pp. 115–148). Monsey, N.Y: Criminal Justice Press.

Bernasco, W., & Block, R. (2009). Where offenders choose to attack: A discrete choice model of robberies in Chicago. *Criminology, 47*, 93–130.

Bernasco, W., & Block, R. (2011). Robberies in Chicago: A block-level analysis of the influence of crime generators, crime attractors, and offender anchor points. *Journal of Research in Crime and Delinquency, 48*, 33–57. https://doi.org/10.1177/0022427810384135.

Birks, D. J., Townsley, M., & Stewart, A. (2014). Emergent regularities of interpersonal victimization: An agent-based investigation. *Journal of Research in Crime and Delinquency, 51*, 119–140. https://doi.org/10.1177/0022427813487353.

Bogar, S., & Beyer, K. M. (2016). Green space, violence, and crime: A systematic review. *Trauma, Violence & Abuse, 17*, 160–171. https://doi.org/10.1177/1524838015576412.

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 ACM international conference on multimodal interaction (ICMI)* (pp. 427–434). https://doi.org/10.1145/2663204.2663254.

Bosse, T., & Gerritsen, C. (2010). A model-based reasoning approach to prevent crime. In J. Kacprzyk, L. Magnani, W. Carnielli, & C. Pizzi (Eds.), *Model-based reasoning in science and technology, volume 314 of Studies in computational intelligence* (pp. 159–177). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-15223-8{\textunderscore}8.

Bowers, K. J., Johnson, S. D., & Pease, K. (2004). Prospective hot-spotting: The future of crime mapping? *British Journal of Criminology, 44*, 641–658. https://doi.org/10.1093/bjc/azh036.

Brantingham, P. L., & Brantingham, P. J. (1993). Environment, routine, and situation: Toward a pattern theory of crime. In R. V. Clarke, & M. Felson (Eds.), *Routine activity and rational choice, advances in criminological theory* (pp. 259–294). New Brunswick: Transaction Publisher.

Browning, C. R., Byron, R. A., Calder, C. A., Krivo, L. J., Kwan, M.-P., Lee, J.-Y., & Peterson, R. D. (2010). Commercial density, residential concentration, and crime: Land use patterns and violence in neighborhood context. *Journal of Research in Crime and Delinquency, 47*, 329–357. https://doi.org/10.1177/0022427810365906.

Ceccato, V. (2005). Homicide in São Paulo, Brazil: Assessing spatial-temporal and weather variations. *Journal of Environmental Psychology, 25*, 307–321. https://doi.org/10.1016/j.jenvp.2005.07.002.

Chainey, S., & Desyllas, J. (2008). Modelling pedestrian movement to measure on-street crime risk. In L. Liu, & J. Eck (Eds.), *Artificial crime analysis systems, Information Science Reference, Hershey, N.Y. and London*.

Chainey, S., Tompson, L., & Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal, 21*, 4–28. https://doi.org/10.1057/palgrave.sj.8350066.

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review, 44*, 588–608.

Cornish, D. B., & Clarke, R. V. (1987). Understanding crime displacement: An application of rational choice theory. *Criminology, 25* (993–947).

Cranshaw, J., Schwartz, R., Hong, J. I., & Sadeh, N. (2012). The livehoods project: Understanding collective activity patterns of a city from social media. In *Proceedings of the 6th international AAAI conference on weblogs and social media (ICWSM)* (pp. 58–65).

Davies, T., & Bowers, K. (2019). Patterns in the supply and demand of urban policing at the street segment level. *Policing and Society, 44*, 1–23. https://doi.org/10.1080/10439463.2019.1598997.

Davies, T., & Johnson, S. D. (2015). Examining the relationship between road structure and burglary risk via quantitative network analysis. *Journal of Quantitative Criminology, 31*, 481–507. https://doi.org/10.1007/s10940-014-9235-4.

Devia, N., & Weber, R. (2013). Generating crime data using agent-based simulation. *Computers, Environment and Urban Systems, 42*, 26–41. https://doi.org/10.1016/j.compenvurbsys.2013.09.001.

Donovan, G. H., & Prestemon, J. P. (2012). The effect of trees on crime in Portland. *Oregon, Environment and Behavior, 44*, 3–30. https://doi.org/10.1177/0013916510383238.

Drucker, J. (2010). Risk factors of street robbery. *RTM Insights, 5*, 1–3.

Elffers, H., & van Baal, P. (2008). Realistic spatial backcloth is not that important in agent based simulation: An illustration from simulating perceptual deterrence. In L. Liu, & J. Eck (Eds.), *Artificial crime analysis systems, information science reference, Hershey, N.Y. and London*.

Felson, M. (2011). Routine activity approach. In R. Wortley, & L. G. Mazerolle (Eds.), *Environmental criminology and crime analysis, crime science series, Routledge, Oxon, England* (pp. 70–93).

Furtado, V., Melo, A., Coelho, A. L., Menezes, R., & Perrone, R. (2009). A bioinspired crime simulation model. *Decision Support Systems, 48*, 282–292. https://doi.org/10.1016/j.dss.2009.08.008.

Glasner, P., & Leitner, M. (2017). Evaluating the impact the weekday has on nearrepeat victimization: A spatio-temporal analysis of street robberies in the city of Vienna, Austria. *ISPRS International Journal of Geo-Information, 6*, 3. https://doi.org/10.3390/ijgi6010003.

Groff, E. R. (2007a). Situating simulation to model human spatio-temporal interactions: An example using crime events. *Transactions in GIS, 11*, 507–530.

Groff, E. (2007b). "Situating" simulation to model human spatio-temporal interactions: An example using crime events. *Transactions in GIS, 11*, 507–530.

Groff, E. R. (2007c). Simulation for theory testing and experimentation: An example using routine activity theory and street robbery. *Journal of Quantitative Criminology, 23*, 75–103. https://doi.org/10.1007/s10940-006-9021-z.

Groff, E. R. (2008). Characterizing the spatio-temporal aspects of routine activities and the geographic distribution of street robbery. In L. Liu, & J. Eck (Eds.), *Artificial crime analysis systems, information science reference, Hershey, N.Y. and London*.

Groff, E. R., & La Vigne, N. G. (2002). Forecasting the future of predictive crime mapping. *Crime Prevention Studies, 13*, 29–57.

Groff, E. R., Johnson, S. D., & Thornton, A. (2018). State of the art in agent-based modeling of urban crime: An overview. *Journal of Quantitative Criminology, 76*, 21. https://doi.org/10.1007/s10940-018-9376-y.

Gunderson, L., & Brown, D. (2000). Using a multi-agent model to predict both physical and cyber criminal activity. In *Proceedings of the IEEE international conference on systems, man and cybernetics, cybernetics evolving to systems, humans, organizations, and their complex interactions, Piscataway, N.J* (pp. 2338–2343). https://doi.org/10.1109/ICSMC.2000.884340.

Hart, T. C., & Miethe, T. D. (2014). Street robbery and public bus stops: A case study of activity nodes and situational risk. *Security Journal, 27*, 180–193. https://doi.org/10.1057/sj.2014.5.

Hastie, T., Friedman, J. H., & Tibshirani, R. (2020). *The elements of statistical learning: Data mining, inference, and prediction*. Cham: Springer International Publishing.

Hecht, B., & Stephens, M. (2015). A tale of cities: Urban biases in volunteered geographic information. In *Proceedings of the 8th international AAAI conference on weblogs and social media (ICWSM)* (pp. 197–205).

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CB09780511973420.

Horanont, T., Phithakkitnukoon, S., Leong, T. W., Sekimoto, Y., & Shibasaki, R. (2013). Weather effects on the patterns of people's everyday activities: A study using GPS traces of mobile phone users. *PLoS One, 8*, Article e81153. https://doi.org/10.1371/journal.pone.0081153.

Hunter, E., Mac Namee, B., & Kelleher, J. (2018). An open-data-driven agent-based model to simulate infectious disease outbreaks. *PLoS One, 13*, Article e0208775. https://doi.org/10.1371/journal.pone.0208775.

Johnson, S. D., Bowers, K. J., & Hirschfield, A. (1997). New insights into the spatial and temporal distribution of repeat victimization. *British Journal of Criminology, 37*, 224–241. https://doi.org/10.1093/oxfordjournals.bjc.a014156.

Kadar, C., & Pletikosa, I. (2018). Mining large-scale human mobility data for long-term crime prediction. *EPJ Data Science, 7*, 344. https://doi.org/10.1140/epjds/s13688-018-0150-z.

Kadar, C., Iria, J., & Pletikosa Cvijikj, I. (2016). Exploring foursquare-derived features for crime prediction in new York City. In *Proceedings of the 5th SIGKDD international workshop on Urban computing (Urb-comp)*.

Kadar, C., Rosés Brüngger, R., & Pletikosa Cvijikj, I. (2017). Measuring ambient population from location-based social networks to describe urban crime. In G. L. Ciampaglia, A. Mashhadi, & T. Yasseri (Eds.), *Social informatics, lecture notes in computer science* (pp. 521–535). Cham: Springer.

Kadar, C., Maculan, R., & Feuerriegel, S. (2019). Public decision support for low population density areas: An imbalance-aware hyper-ensemble for spatio-temporal crime prediction. *Decision Support Systems, 119*, 107–117. https://doi.org/10.1016/j.dss.2019.03.001.

Kadar, C., Feuerriegel, S., Noulas, A., & Mascolo, C. (2020). Leveraging mobility flows from location technology platforms to test crime pattern theory in large cities. In *Proceedings of the fourteenth international AAAI conference on web and social media (ICWSM)* (pp. 339–350).

Kinney, J. B., Brantingham, P. L., Wuschke, K., Kirk, M. G., & Brantingham, P. J. (2008). Crime attractors, generators and detractors: Land use and urban crime opportunities. *Built Environment, 34*, 62–74. https://doi.org/10.2148/benv.34.1.62.

Kondo, M. C., Han, S., Donovan, G. H., & MacDonald, J. M. (2017). The association between urban trees and crime: Evidence from the spread of the emerald ash borer in Cincinnati. *Landscape and Urban Planning, 157*, 193–199. https://doi.org/10.1016/j.landurbplan.2016.07.003.

Lammers, M., & Bernasco, W. (2013). Are mobile offenders less likely to be caught? The influence of the geographical dispersion of serial offenders' crime locations on their probability of arrest. *European Journal of Criminology, 10*, 168–186. https://doi.org/10.1177/1477370812464533.

Lee, Y., Eck, J. E., Soohyun, O., & Martinez, N. N. (2017). How concentrated is crime at places? a systematic review from 1970 to 2015. *Crime Science, 6*, 487. https://doi.org/10.1186/s40163-017-0069-x.

Linning, S. J., Andresen, M. A., Ghaseminejad, A. H., & Brantingham, P. J. (2017). Crime seasonality across multiple jurisdictions in British Columbia, Canada. *Canadian Journal of Criminology and Criminal Justice, 59*, 251–280. https://doi.org/10.3138/cjccj.2015.E31.

Liu, Y., Kang, C., Gao, S., Xiao, Y., & Tian, Y. (2012a). Understanding intraurban trip patterns from taxi trajectory data. *Journal of Geographical Systems, 14*, 463–483. https://doi.org/10.1007/s10109-012-0166-z.

Liu, Y., Wang, F., Xiao, Y., & Gao, S. (2012b). Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. *Landscape and Urban Planning, 106*, 73–87. https://doi.org/10.1016/j.landurbplan.2012.02.012.

Malleson, N., & Andresen, M. A. (2014). The impact of using social media data in crime rate calculations: Shifting hot spots and changing spatial patterns. *Cartography and*

Geographic Information Science, 42, 112–121. https://doi.org/10.1080/15230406.2014.905756.
Malleson, N., & Andresen, M. A. (2015). Spatio-temporal crime hotspots and the ambient population. *Crime Science, 4*, 258. https://doi.org/10.1186/s40163-015-0023-8.
Malleson, N., & Birkin, M. (2012). Analysis of crime patterns through the integration of an agent-based model and a population microsimulation. *Computers, Environment and Urban Systems, 36*, 551–561. https://doi.org/10.1016/j.compenvurbsys.2012.04.003.
Malleson, N., Evans, A., & Jenkins, T. (2009). An agent-based model of burglary. *Environment and Planning. B, Planning & Design, 36*, 1103–1123. https://doi.org/10.1068/b35071.
Malleson, N., Heppenstall, A., & See, L. (2010). Crime reduction through simulation: An agent-based model of burglary. *Computers, Environment and Urban Systems, 34*, 236–250. https://doi.org/10.1016/j.compenvurbsys.2009.10.005.
Malleson, N., Heppenstall, A., See, L., & Evans, A. (2013). Using an agent-based crime simulation to predict the effects of urban regeneration on individual household burglary risk. *Environment and Planning. B, Planning & Design, 40*, 405–426. https://doi.org/10.1068/b38057.
Masad, D., & Kazil, J. (2015). Mesa: An agent-based modeling framework. In *Proceedings of the 14th python in science conference (SciPy)*.
Mccord, E., & Ratcliffe, J. H. (2009). Intensity value analysis and the criminogenic effects of land use geatures on local crime patterns. *Crime Pattern and Analysis, 2*, 17–30.
Morselli, C., & Royer, M.-N. (2008). Criminal mobility and criminal achievement. *Journal of Research in Crime and Delinquency, 45*, 4–21. https://doi.org/10.1177/0022427807309630.
Murray, R. K., & Swatt, M. L. (2013). Disaggregating the relationship between schools and crime. *Crime & Delinquency, 59*, 163–190. https://doi.org/10.1177/0011128709348438.
New York State Department of Transportation. (2018). New York City mobility report. http://www.nyc.gov/html/dot/downloads/pdf/mobility-report-2018-print.pdf.
Newman, O. (1972). *Defensible space*. New York: Macmillan.
Nguyen, T., & Szymanski, B. K. (2012). Using location-based social networks to validate human mobility and relationships models. In *Proceedings of the 2012 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)* (pp. 1215–1221).
Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2011). An empirical study of geographic user activity patterns in foursquare. In *Proceedings of the 5th international AAAI conference on weblogs and social media (ICWSM)*.
Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., & Mascolo, C. (2012). A tale of many cities: Universal patterns in human urban mobility. *PLoS One, 7*. https://doi.org/10.1371/journal.pone.0037027.
Noulas, A., Mascolo, C., & Frias-Martinez, E. (2013). Exploiting Foursquare and cellular data to infer user activity in urban environments. In *Proceedings of the 14th IEEE international conference on mobile data management (MDM)* (pp. 167–176). IEEE. https://doi.org/10.1109/MDM.2013.27.
O'Brien, D. T. (2016). Using small data to interpret big data: 311 reports as individual contributions to informal social control in urban neighborhoods. *Social Science Research, 59*, 83–96. https://doi.org/10.1016/j.ssresearch.2016.04.009.
Peng, C., & Kurland, J. (2014). The agent-based spatial simulation to the burglary in Beijing. In *Computational science and its applications – ICCSA 2014, volume 8582 of lecture notes in computer science* (pp. 31–43). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-09147-1{\textunderscore}3.
Reid, A. A., Frank, R., Iwanski, N., Dabbaghian, V., & Brantingham, P. (2013). Uncovering the spatial patterning of crimes: A criminal movement model (CriMM). *Journal of Research in Crime and Delinquency, 51*, 230–255. https://doi.org/10.1177/0022427813483753.
Rosés, R., Kadar, C., Gerritsen, C., & Rouly, C. (2018). Agent-based simulation of offender mobility: Integrating activity nodes from location-based social networks. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems (AAMAS)*.
Rosés, R., Kadar, C., Gerritsen, C., & Rouly, O. C. (2020). Simulating offender mobility: Modeling activity nodes from large-scale human activity data. *Journal of Artificial Intelligence Research, 68*, 541–570. https://doi.org/10.1613/jair.1.11831.
Rosser, G., Davies, T., Bowers, K. J., Johnson, S. D., & Cheng, T. (2016). Predictive crime mapping: Arbitrary grids or street networks? *Journal of Quantitative Criminology*. https://doi.org/10.1007/s10940-016-9321-x.
Rummens, A., Hardyns, W., & Pauwels, L. (2017). The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context. *Applied Geography*. https://doi.org/10.1016/j.apgeog.2017.06.011.
Schnell, C., Grossman, L., & Braga, A. A. (2019). The routine activities of violent crime places: A retrospective case-control study of crime opportunities on street segments. *Journal of Criminal Justice, 60*, 140–153. https://doi.org/10.1016/j.jcrimjus.2018.10.002.
Sherman, L. W., Gartin, P. R., & Buerger, M. E. (1989). Hot spots of predatory crime: Routine activities and the criminology of place. *Criminology, 27*, 27–56. https://doi.org/10.1111/j.1745-9125.1989.tb00862.x.
Song, G., Bernasco, W., Liu, L., Xiao, L., Zhou, S., & Liao, W. (2019). Crime feeds on legal activities: Daily mobility flows help to explain thieves' target location choices. *Journal of Quantitative Criminology, 35*, 831–854. https://doi.org/10.1007/s10940-019-09406-z.
Sorg, E. T., & Taylor, R. B. (2011). Community-level impacts of temperature on urban street robbery. *Journal of Criminal Justice, 39*, 463–470. https://doi.org/10.1016/j.jcrimjus.2011.08.004.
Stein, R. E., Conley, J. F., & Davis, C. (2016). The differential impact of physical disorder and collective efficacy: A geographically weighted regression on violent crime. *GeoJournal, 81*, 351–365. https://doi.org/10.1007/s10708-015-9626-6.
Stucky, T. D., & Ottensmann, J. R. (2009). Land use and violent crime. *Criminology, 47*, 1223–1264.
Summers, L., & Johnson, S. D. (2017). Does the configuration of the street network influence where outdoor serious violence takes place? Using space syntax to test crime pattern theory. *Journal of Quantitative Criminology, 33*, 397–420. https://doi.org/10.1007/s10940-016-9306-9.
Sypion-Dutkowska, N., & Leitner, M. (2017). Land use influencing the spatial distribution of urban crime: A case study of Szczecin, Poland. *ISPRS International Journal of Geo-Information, 6*, 74. https://doi.org/10.3390/ijgi6030074.
Tang, J., Liu, F., Wang, Y., & Wang, H. (2015). Uncovering urban human mobility from large scale taxi GPS data. *Physica A: Statistical Mechanics and its Applications, 438*, 140–153. https://doi.org/10.1016/j.physa.2015.06.032.
Tappan, P. W. (1947). Who is the criminal? *American Sociological Review, 12*, 96–102.
Taylor, R. B., & Gottfredson, S. (1986). Environmental design, crime, and prevention: An examination of community dynamics. *Crime and Justice, 8*, 387–416.
Tompson, L. (2015). Street robbery: Summary. *JDiBrief - Crime*. London: University College. ISSN 2050-4853.
Tompson, L. A., & Bowers, K. J. (2015). Testing time-sensitive influences of weather on street robbery. *Crime Science, 4*, 1213. https://doi.org/10.1186/s40163-015-0022-9.
Twinam, T. (2017). Danger zone: Land use and the geography of neighborhood crime. *Journal of Urban Economics, 100*, 104–119. https://doi.org/10.1016/j.jue.2017.05.006.
Urban, C., & Schmidt, B. (2001). PECS - agent-based modeling of human behavior. In *In Emotional and Intelligent-The Tangled Knot of Social Cognition*. AAAI Fall Symposoim Series.
Vaněk, O., Jakob, M., Hrstka, O., & Pěchouček, M. (2013). Agent-based model of maritime traffic in piracy-affected waters. *Transportation Research Part C: Emerging Technologies, 36*, 157–176. https://doi.org/10.1016/j.trc.2013.08.009.
Vomfell, L., Händle, W. K., & Lessmann, S. (2018). Improving crime count forecasts using twitter and taxi data. *Decision Support Systems, 113*, 73–85. https://doi.org/10.1016/j.dss.2018.07.003.
Wang, X., & Brown, D. E. (2012). The spatio-temporal modeling for criminal incidents. *Security Informatics, 1*, 2. https://doi.org/10.1186/2190-8532-1-2.
Wang, X., Liu, L., & Eck, J. (2008). Crime simulation using GIS and artificial intelligent agents. In L. Liu, & J. Eck (Eds.), *Artificial crime analysis systems, information science reference, Hershey, N.Y. and London*.
Wang, N., Liu, L., & Eck, J. E. (2014). Analyzing crime displacement with a simulation approach. *Environment and Planning. B, Planning & Design, 41*, 359–374. https://doi.org/10.1068/b37120.
Wang, G., Schoenebeck, S. Y., Zheng, H., & Zhao, B. Y. (2016a). "will check-in for badges": Understanding bias and misbehavior on location-based social networks. In *Proceedings of the 10th international AAAI conference on web and social media (ICWSM)*.
Wang, H., Kifer, D., Graif, C., & Li, Z. (2016b). Crime rate inference with big data. In *Proceedings of the 22nd ACM SIGKDD international conference* (pp. 635–644). https://doi.org/10.1145/2939672.2939736.
Weisburd, D. (2015). The law of crime concentration and the criminology of place. *Criminology, 53*, 133–157. https://doi.org/10.1111/1745-9125.12070.
Weisburd, D. A., Bushway, S., Lum, C., & Yang, S.-M. (2004). Trajectories of crime at places: A longitudinal study of street segments in the city of Seattle. *Criminology, 42*, 283–322.
Wheeler, A. P. (2018). The effect of 311 calls for service on crime in D.C. at microplaces. *Crime & Delinquency, 64*, 1882–1903. https://doi.org/10.1177/0011128717714974.
White, A., & Trump, K.-S. (2018). The promises and pitfalls of 311 data. *Urban Affairs Review, 54*, 794–823. https://doi.org/10.1177/1078087416673202.
Willits, D., Broidy, L., & Denman, K. (2013). Schools, neighborhood risk factors, and crime. *Crime & Delinquency, 59*, 292–315. https://doi.org/10.1177/0011128712470991.
Wilson, J. Q., & Kelling, G. L. (1982). The police and neighborhood safety: Broken windows. march, 29–38. *The Atlantic Monthly, 249*, 29–38.
Yang, D., Heaney, T., Tonon, A., Wang, L., & Cudré-Mauroux, P. (2017). Crime-telescope: Crime hotspot prediction based on urban and social media data fusion. *World Wide Web, 4*, 114. https://doi.org/10.1007/s11280-017-0515-4.
Zha, Y. F., & Veloso, M. (2014). Profiling and prediction of non-emergency calls in nyc. In *Workshops at the twenty-eighth AAAI conference on artificial intelligence*.
Zhang, H., Vorobeychik, Y., Letchford, J., & Lakkaraju, K. (2016). Data-driven agent-based modeling, with application to rooftop solar adoption. *Autonomous Agents and Multi-Agent Systems, 30*, 1023–1049. https://doi.org/10.1007/s10458-016-9326-8.