An Open-data, Agent-based Model of Alcohol Related Crime

Joseph Redfern, Kirill Sidorov, Paul L. Rosin, Simon C. Moore[†], Padraig Corcoran, David Marshall
School of Computer Science and Informatics, Cardiff University
[†]School of Dentistry, Cardiff University
RedfernJM@cardiff.ac.uk

Abstract

The allocation of resources to challenge city centre violent crime traditionally relies on historical data to identify hot-spots. The usefulness of such data-driven approaches is limited when historical data is scarce or unavailable (e.g. planning of a new city) or insufficiently representative (e.g. does not account for novel events, such as Olympic Games). In some cities, crime data is not systematically accumulated at all.

We present a graph-constrained agent based simulation model of alcohol-related violent crime that is capable of predicting areas of likely violent crime without requiring any historical data. The only inputs to our simulation are publicly available geographical data, which makes our method immediately applicable to a wide range of tasks, such as optimal city planning, police patrol optimisation, devising alcohol licensing policies.

In experiments, we evaluate our model and demonstrate agreement of our model's predictions on where and when violence will occur with real-world violent crime data. Analyses indicate that our agent based model may be able to make a significant contribution to attempts to prevent violence through deterrence or by design.

1. Introduction

The ability to predict the behaviour of drinkers in the night time economy is a valuable resource for many sectors, including city planning, licensing authorities and police forces. Multiple factors influence behaviour in night time environments including novel events, additional premises licensed for the sale an on-site consumption of alcohol opening and the physical design of the environment. Practitioners involved with managing such spaces, such as the police, seek to plan how best to reduce violent crime and city planners will want to ensure any change in the environment, whether that is changes to the way alcohol is sold or changes to the physical environment, are made in a manner that does not cause violence to escalate. Currently there are no reliable tools available to predict where and when violence will occur and planning often involves a best guess based on historical patterns in violent crime data.

Large events such as football matches (for instance World Cup finals) and concerts can bring hundreds of thousands of extra visitors to cities. Many of these attendees will go on to visit pubs and clubs, impacting the street busyness and risk levels for violence. As it stands, much of the policing is planned based on intuition and historic data [9] – this is effective for typical nights, but is unlikely to be optimal for unusual events, or combinations thereof.

It is estimated that the annual cost of alcohol-related harm to society is around £21bn, £3.5bn of which is incurred by the NHS (National Health Service) [3]. A system capable of identifying areas at risk of increased levels of alcohol-related violence would help enable the introduction of measures to reduce these costs, as well as to reduce the physical harm caused to victims.
The main contributions of this paper are:

- A novel agent-based flexible simulated model is proposed which is capable of predicting levels of alcohol related crime from geographical data alone.
- A fully automatic tool chain is described which allows analysts to effortlessly set up their simulation using freely available data (in the form of `OpenStreetMap` (OSM) data files).
- A mechanism for comparing simulated results against real world crime data is presented. Experiments demonstrate agreement between the simulated and ground-truth data.

2. Related Work

`SimDrink` [10] is an agent-based model of alcohol consumption, aimed at simulating the effectiveness of different alcohol licensing policies. The `SimDrink` model aims to predict whether an agent will encounter any type of "harm"

978-1-5386-2939-0/17/$31.00 ©2017 IEEE

during their night out, including over-intoxication, involvement in verbal abuse, or being unable to get home due to transportation issues. Agents' behaviour varies according to several factors, including age, available funds, drinking rate, and planned duration of imbibement activities. Venues' parameters include closing time, drink price, and drinking limit. Although well suited for informing policy decisions, the approach of `SimDrink` [10] does not directly apply to predictive policing (the planning of police patrol routes) as the true structure of the street networks is not considered.

Davies *et al.* [2] present a study that aims to predict the risk of residential burglary from properties of the road network (betweenness and linearity). They confirmed that betweenness was a significant contributor to the likelihood of being burgled, and that street linearity also had a degree of correlation with this risk. While Davies *et al.* [2] applies to residential burglary, the motion of the agents in our study automatically takes into account some similar properties of the road network, with a specific focus on alcohol serving establishments.

Zhu *et al.* [12] conducted a study quantifying the relationship between density of alcohol outlets and levels of violence. The study was focused on two Texan cities, and found that density of outlets explained 71% and 56% of the variability in crime levels for the two cities respectively. Our model, which relies on exact locations of specific outlets does not require the estimation of aggregate quantities such as outlet density, as this is implicitly taken into account.

3. Simulation Methodology

3.1. Rationale

Our system uses an agent-based model approach to identify areas at risk of alcohol related harm. This approach makes it very easy to explore how additional rules or behaviours for individual drinkers impact on the overall dynamics of the system.

We base our approach around the assumption that the levels of alcohol related violence correlate with how many people are in a given area, and how drunk those people are. The findings of Zhu *et al.* [12] studying correlation of alcohol outlet density to levels of crime suggest that this assumption is fair.

3.2. Method

Our simulator takes as an input `OpenStreetMap` (OSM) data files for the area the analyst wishes to model, and generates a graph comprised of nodes and ways (see Section 3.5.1 for details). This is accomplished using the `graph-tool` [8] Python network analysis library, and `imposm-parser` [6] for processing `OpenStreetMap` data files.

In this simulation, agents represent drinkers who traverse the road network. Agents move from their homes into the city and then around the network, either visiting new venues or going home when their current bar/pub closes. The home location of an agent is determined by randomly sampling `OpenStreetMap` streets tagged as residential, and we assume the widely accepted [1] walking pace of $1.4\,\mathrm{m\,s^{-1}}$ when en-route to venues.

Algorithm 1 Summary of an agent's behaviour.

```
Require: Time step $\Delta T$, total agent session duration $T_{\max}$,
    agent's body mass BM, agent's time between drinks
    $\Delta D$, time spent in each venue $\Delta V$.
 1 Initialisation:
 2 $T \leftarrow 0$                                   ▷ total simulation time
 3 $A \leftarrow 0$                                   ▷ grams of ethanol ingested
 4 while $T < T_{\max}$ do
 5     $\mathbb{V} \leftarrow$ FINDOPENVENUES
 6     if $\emptyset \neq \mathbb{V}$ then
 7         $V \leftarrow \mathbb{V}[\mathcal{U}\{1, |\mathbb{V}|\}]$          ▷ select venue
 8     else
 9         go home
10     end if
11     go to venue $V$
12     $T_d \leftarrow 0$                             ▷ time elapsed since last drink
13     $T_a \leftarrow T$                             ▷ time of arrival at venue
14     while $T < T_a + \Delta V$ and $T < T_{\max}$ do
15         if $T - T_d > \Delta D$ then
16             consume a randomly chosen drink:
17             $A \leftarrow A + D_{\text{ethanol}}(\mathcal{U}\{1, 4\})$    ▷ See Table 3
18             $\mathrm{BAC} \leftarrow \frac{A}{(\mathrm{WC} \times \mathrm{BM})} - \mathrm{EC} \times \mathrm{T_{elapsed}}$
19             $T_d \leftarrow 0$
20         end if
21         $T \leftarrow T + \Delta T$          ▷ simulation step, $\Delta T = 60\ \mathrm{s}$
22         $T_d \leftarrow T_d + \Delta T$
23     end while
24 end while
25 go home
```

The behaviour of an agent is summarised in Algorithm 1 (all agents are simulated concurrently and all follow the same algorithm). An agent, starting initially from home, moves from one venue to the next, with each next venue uniformly randomly sampled (lines 5–11) from the pool of currently open venues (if none exist, the agent goes home). In each venue, the agent spends an amount of time ΔV drawn from a normal distribution (see 1), repeatedly (lines 14–23) drinking until this time is elapsed. The distribution parameters $\mu_v = 114.6$ min, $\sigma_v = 64.2$ min for visit duration ΔV were obtained from survey data [5] detailing venue exit/entry times of drinkers in South Wales. The time between consecutive drinks ΔD (amounting to the rate of alcohol consumption) is drawn from a capped normal distri-

Parameter	Value
Sex	Sampled from the Bernoulli distribution ($P = 1/2$)
Mass	Sampled from a normal distribution with $\mu = 62$ kg (for women), $\mu = 78.5$ kg (for men) , $\sigma = 10$ kg
Time between drinks, ΔD	Sampled from a cut-off normal distribution $\max(10 \text{ min}, \mathcal{N}(90, 15))$ (estimated)
Time spent per venue, ΔV	Sampled from a cut-off normal distribution $\max(20 \text{ min}, \mathcal{N}(114.6, 64.2))$
Session duration, $T_{\max}$	Sampled from a cut-off normal distribution $\max(60 \text{ min}, \mathcal{N}(300, 120))$ (estimated)

Table 1. Fixed agent parameters.

Parameter	Value
Route	A timestamped ordered list of OSM nodes visited/to visit
Current Node	ID of current OSM node
Previous Drinks	A time-stamped list of drinks consumed

Table 2. Dynamic agent parameters.

bution (see Table 1) with $\mu_d = 90$ min and $\sigma_d = 15$ min. Each consumed drink, drawn uniformly randomly from the "menu" (line 17, see also Table 3), is accounted for in the resulting blood alcohol content (lines 17–18). When an agent is moving between venues (lines 9, 11, 25), the shortest path route along the road network between the current position and the destination is chosen, and the agent is gradually advanced along this path.

Agents continue drinking for a maximum duration of $T_{\max}$, which is drawn from the distribution listed in Table 1. In addition, there is also a global user-controlled parameter limiting the duration of the entire simulation for all agents (set to 360 min in our experiments).

3.3. Agent Description

When each agent is instantiated, their fixed parameters are sampled from distributions listed in Table 1, and remain unchanged throughout the simulation. As the simulation progresses, for each agent we maintain the dynamic properties listed in Table 2.

Name	Volume	Strength	Units
Single Shot	25ml	40%	1
Double Shot	50ml	40%	2
Premium Lager	568ml	5%	2.84
Standard Wine	175ml	12%	2.1

Table 3. Drinks available for consumption by agents.

3.3.1 Alcohol Levels

Given our assumption that risk of violence is related to drunkenness, it is necessary to be able to approximate alcohol levels for each agent. Using the Widmark formula [11], we are able to estimate the BAC (blood alcohol content) for a given person based on units of alcohol consumed, the agent's mass and gender, and the time elapsed since their first drink:

$$\mathrm{BAC} \approx \frac{A}{(\mathrm{WC} \times \mathrm{BM})} - \mathrm{EC} \times \mathrm{T}_{\mathrm{elapsed}} \tag{1}$$

where A is the quantity of ethanol ingested (in grams), WC is a water constant (0.49 for women, 0.58 for men), EC is the elimination constant (0.017 for women, 0.015 for men), BM is the body mass in kilograms, and T_{elapsed} is time elapsed since the first drink. (Note that the elimination of alcohol starts immediately and the liver always works to the full capacity: hence the elimination rate does not depend on the amount of alcohol consumed [11]).

3.4. Alcoholic Drinks

The simulation includes 4 drinks, represented internally by their alcohol levels and volume, as described in Table 3.

Drinks are currently chosen at random from a uniform distribution, and agents do not display preferences towards any particular choice. It is assumed that these drinks are available at every venue.

3.5. Simulation Data

The program's main input is the `OpenStreetMap` (OSM) data on which the simulation runs. Given an OSM data-file, our system acquires additional venue data from `Google Places` and `Foursquare`. Additional simulation inputs include the number of agents to deploy, start time, end time, and day of week (to account for varying venue opening hours)

The output of the simulation is the position of each agent along the road network over discretised time with a time step of 60 s, along with each agent's BAC.

The simulation is written in Python, and uses the `graph-tool` [8] library. Results are stored in a `PostgreSQL` database, and queried using `PostGIS` spatial extensions.

3.5.1 OpenStreetMap

We constrain the movements of agents by OpenStreetMap [7] road network data. OpenStreetMap is a collaborative, open-data effort to produce user-generated maps of the entire planet.

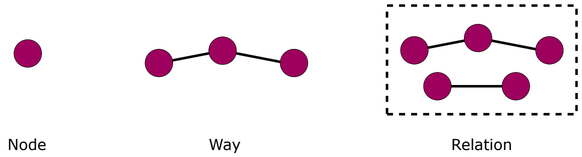

Figure 1. Illustration of OSM data types.

OpenStreetMap data consists of node, way, relation and tag types (see Fig. 1). Nodes represent individual points, which can be stand-alone, or part of a collection making one or more ways. Ways form collections of nodes and typically represent roads, paths or buildings. A relation is a collection of ways and/or nodes — for instance, a hospital comprised of multiple buildings could share a common relation. Tags store attributes of nodes, ways and relations — for example name, address, amenity type (cafe, bar, bank, road). For the purpose for analysing the road network, we consider only nodes and ways (and their associated tags).

3.5.2 OSM Data Augmentation

While OpenStreetMap data has been found to be of good structural quality [4], we found meta-data (such as venue name, venue type and opening times) to often be missing, incomplete or out of date. In order to get around this problem, we augment the OSM data with venues and opening times listed on Foursquare and Google Places using their respective APIs. We found that this significantly increases the number of usable venues in our model, and boosts the accuracy of our results.

4. Evaluation Methodology

We employ both grid-based Spearman's Rho and Pearson's R correlation measures across $250\ \mathrm{m} \times 250\ \mathrm{m}$ cells. Through counting the total number of seconds each agent has resided within the boundaries of each cell (agent-seconds), we are able to determine the relative predicted busyness of each cell. We then weight the busyness of each cell by the average blood alcohol concentration of the agents within the cell.

Having calculated the BAC-weighted busyness for each cell, we then count the number of crimes having occurred per cell. Crime levels are then correlated with the BAC-weighted busyness level on a per-cell basis in order to measure the relationship between our simulation and real data.

	weighted		unweighted		
Day	r	ρ	r	ρ	p-value
Mon	0.558	0.565	0.522	0.552	$\ll 0.0001$
Tue	0.584	0.610	0.550	0.595	$\ll 0.0001$
Wed	0.683	0.700	0.647	0.679	$\ll 0.0001$
Thu	0.651	0.650	0.609	0.624	$\ll 0.0001$
Fri	0.664	0.660	0.632	0.644	$\ll 0.0001$
Sat	0.733	0.685	0.696	0.676	$\ll 0.0001$
Sun	0.565	0.577	0.517	0.559	$\ll 0.0001$

Table 4. Correlation between both BAC-weighted and unweighted simulation results and ground truth historic data for Northampton. Here, $10,000$ agents were used in the simulation covering the period between 18:00–04:00. ρ represents Spearman's rank correlation coefficients, and r represents Pearson's r correlation coefficients. Note that p-value shown is for all correlation coefficients on the given day.

Name	**Description**
Point ID	Unique identifier for data point
Radio ID	Unique identifier for the radio transmitting telemetry
Latitude	WGS84 Datum
Longitude	WGS84 Datum
Incident ID	Identifier for the incident being responded to
Incident Type	The type of crime being responded to
Status	Consisting of: En-route, At Scene, Available, Unavailable

Table 5. Summary of the Northampton police data records.

4.1. Evaluation Data

4.2. Results

Using the data and method described in Section 4, we are able to calculate the Pearson's R (r) and Spearman's Rho (ρ) correlation coefficients between our predicted crime levels and actual crime levels (see Table 4 for full figures). We count the total number of incidents of Violence Against the Person and Rowdy Behaviour (as determined from data describe in Table 5) having taken place in each grid cell between the given times, on the given day of the week. We consider both BAC-weighted agent-second counts (see Fig. 2), and agent-seconds alone. In every case, weighting the agent-second count by average BAC level increases both correlation measures. Even though the increase in correlation value is small, and further investigation is required, results would suggest that risk of violence in the night time economy is correlated with alcohol levels as well as street-busyness.

We believe that variations in correlation coefficients by

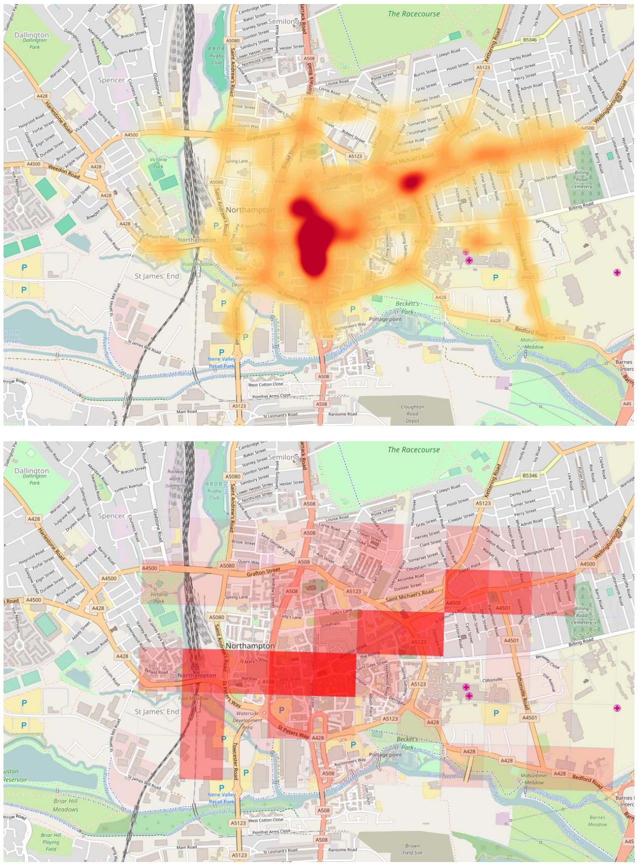

Figure 2. *Top:* heatmap of ground truth rowdy/violent behaviour. (Northampton, Saturday.) *Bottom:* total duration of agent presence weighted by average BAC, for of all the simulated agents within each 250m × 250m cell. (Northampton, Saturday.)

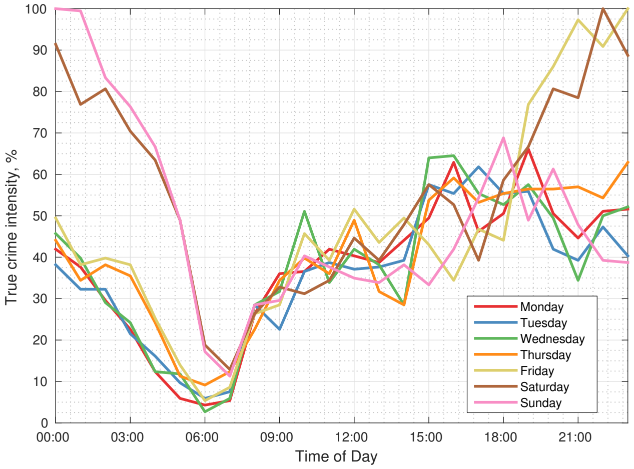

Figure 3. Hourly rates of crime for Northampton (as percentage of maximum rate)

day of week correspond to different real-world levels of activity on each night. For example, Friday and Saturday nights fall on the weekend, and Wednesday is a traditional student drinking night which may explain the increase in correlation relative to that of other nights. This increase in number of drinkers improves the signal to noise ratio in the crime data (where signal corresponds to the number incidents resulting from alcohol consumption, and noise corresponds to the number of incidents which do not, see Fig. 3), and thereby improves our resulting correlation coefficient. As our simulation models agent drinking and movement habits, background crimes (those occurring for reasons other than drunkenness) are not simulated, but are present in the crime data we correlate against.

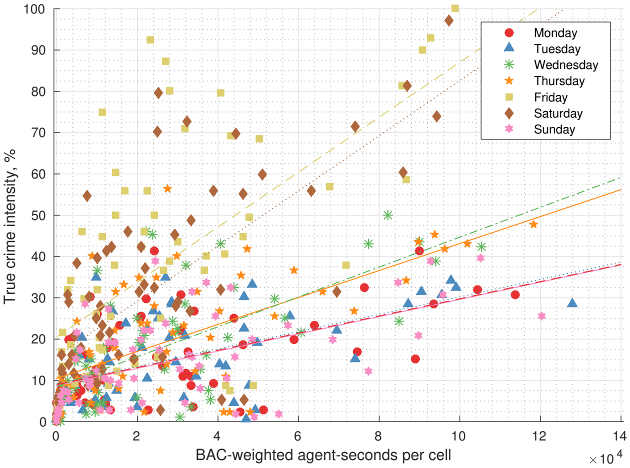

Figure 4. Scatter plot of true incident count (as percentage of maximum) versus alcohol-weighted node-seconds (note that lines for Sunday and Monday are very similar).

We have plotted the incident count versus alcohol-weighted node-seconds for each day of the week in Fig. 4, including lines of best-fit describing the relationship between the two variables. Distinct trends are observed, showing varying relations between incident count and model output by day of week — for example, it appears that on Fridays and Saturdays the number of crime incidents grows faster with BAC-weighted agent-seconds than on other days.

5. Conclusion

We have shown that our approach to predicting areas at-risk of alcohol related violence using a graph-constrained agent-based model shows promise. Our system is able to produce output that is correlated with true levels of alcohol-related violent crime without knowledge of prior incidents, making use only of freely available data. This indicates that sufficient evidence exists to warrant moving towards a phase II feasibility study that would be able to better test causal mechanisms and determine whether the tool can provide information that can be used to deter or better manage violence.

By comparing correlation measures that both do and do not take into account alcohol levels, we show that drinking behaviour is likely to contribute towards the risk of being involved in violent crime.

6. Future Work

There is large scope for expansion of this research. Potential areas for exploration include modelling venue demographic, automatic inclusion of events (from social media), modelling group behaviour, fast-food outlets and alternative transport modes (for instance, including taxi ranks and bus stations).

Routing choices could also be developed — it is currently assumed that agents will take the shortest path along the road network between their current and target destination. A more realistic model of pedestrian routing is likely to improve results.

Investigating additional constraints and logic to agent behaviour may also result in increased accuracy. For instance, drawing on some of the ideas presented in `SimDrink` [10] and allocating each agent a spending limit, or taking into account "pre-drinking" is likely to result in more natural model behaviour.

Additionally, further research into the quality and completeness of the `OpenStreetMap` data should be conducted — if residential areas or venues are missing from the tag data, then simulation results are likely to be skewed.

Acknowledgements

We would like to thank Northampton Police Force for providing us with the crime data that made this study possible.

References

[1] Raymond C. Browning et al. "Effects of obesity and sex on the energetic cost and preferred speed of walking". In: *J. of Applied Physiology* 100.2 (2006), pp. 390–398.

[2] Toby Davies and Shane D. Johnson. "Examining the Relationship Between Road Structure and Burglary Risk Via Quantitative Network Analysis". In: *J. of Quantitative Criminology* 31.3 (2015), pp. 481–507. ISSN: 1573-7799.

[3] Public Health England. *Alcohol treatment in England 2013–14*. URL: `http://www.nta.nhs.uk/uploads/adult-alcohol-statistics-2013-14-commentary.pdf`.

[4] Mordechai Haklay. "How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets". In: *Environment and Planning B: Planning and Design* 37.4 (2010), pp. 682–703.

[5] Simon C Moore, Iain Brennan, and Simon Murphy. "Predicting and measuring premises-level harm in the night-time economy". In: *Alcohol and Alcoholism* 46.3 (2011), pp. 357–363.

[6] Omniscale. *`imposm.parser`*. `https://github.com/omniscale/imposm-parser`. 2017.

[7] OpenStreetMap contributors. *Planet dump retrieved from `https://planet.osm.org`*. 2017.

[8] Tiago P. Peixoto. *The `graph-tool` Python library*. `http://figshare.com/articles/graph_tool/1164194`. 2014. (Visited on 09/10/2014).

[9] College of Policing. "The effects of Hot-Spot Policing on Crime". In: *What works briefing* (2013). URL: `http://library.college.police.uk/docs/what-works/What-works-briefing-hotspot-policing-2013.pdf` (visited on 06/13/2017).

[10] Nick Scott et al. "SimDrink: An Agent-Based NetLogo Model of Young, Heavy Drinkers for Conducting Alcohol Policy Experiments". In: *J. of Artificial Societies and Social Simulation* 19.1 (2016), p. 10.

[11] Erik M. P. Widmark. *Alkoholblodprovet: sambandet mellan alkoholförtäring och promillevärdena i blodet: tabeller för beräkningen*. Lund: Gleerup, 1941.

[12] L. Zhu, D. M. Gorman, and S. Horel. "Alcohol outlet density and violence: a geospatial analysis". In: *Alcohol and Alcoholism* 39.4 (2004), p. 369.