After the Twitter X-pocalypse: Approaches to Characterising Human Behaviour in Agent-based Models and Beyond

Kirsty Watkinson[*1] and Jonny Huck[†1]

[1]Department of Geography, University of Manchester, UK

GISRUK 2024

Summary

Characterising human behaviour is challenging, and datasets about people often suffer from issues of misrepresentation. To account for misrepresentation, researchers have turned to data synthesis. Here, we implement a straightforward data synthesis approach that does not rely upon knowledge of dataset uncertainty and use it to parametrise predictors used in an agent-based model (ABM) to estimate visits by people to greenspaces in Glasgow. The predicted visits follow expected patterns, with more visits on weekends, during daylight, and to popular tourist destinations. The approach is easy to implement and allows researchers to combine datasets of varying veracity to predict human behaviour.

KEYWORDS: data synthesis, human behaviour, agent-based model, social media

1. Introduction

In October 2022, Elon Musk took over Twitter and implemented sweeping changes, including a rebranding to 'X'. Before the takeover, Twitter was a popular and convenient source of data for research, useful for characterising human behaviour, providing data about people's location and activities. However, access to its data is now restricted and costly. Like other social media (e.g., Instagram, Strava) and data about human behaviour (e.g., questionnaires, volunteered geographic information (VGI)), Twitter suffers from long-recognised issues relating to socio-demographic misrepresentation caused by the nature of the user base (Li et al., 2013; Sinclair et al., 2023) and spatial misrepresentation arising from the geocoding process (Huck et al., 2015). These misrepresentations can limit the use of such datasets and must be considered when approaching problems like characterising human behaviour.

Acknowledging these limitations, researchers have turned to data synthesis, using diverse techniques to combine heterogenous datasets to answer questions about society (Janowicz et al., 2015). For example, Gao et al. (2017) used grid-based and point-clustering approaches to classify locations as either Northern or Southern California from geotagged social media. Similarly, Huck et al. (2023) developed a fuzzy Bayesian inference approach to estimate the class membership of a location from different geographical datasets (GPS, VGI). However, such methods often require knowledge of dataset uncertainty or geolocation so are inappropriate where such information is unknown (e.g., online reviews, Google Popular Times). In this paper, we implement a straightforward data synthesis approach, where data about human behaviour are combined to predict the probability of visits to greenspaces across Glasgow, specifically;

1) When are people most likely to visit greenspaces?
2) Which greenspaces are people most likely to visit?

[*] kirsty.watkinson@manchester.ac.uk
[†] jonathan.huck@manchester.ac.uk

2. Methods

2.1. Example application

Visit probability was predicted for 15 greenspaces in the Glasgow City Region included in an ABM exploring Lyme Disease transmission (Figure 1). The model simulates interactions between people, deer, and ticks, to explore how changes to greenspace change transmission risk. Deer and ticks are easily parameterised using existing models (Jepsen and Topping, 2004; Li et al., 2016). The goal here is to parameterise the ABM with human visitor numbers and timings for each greenspace.

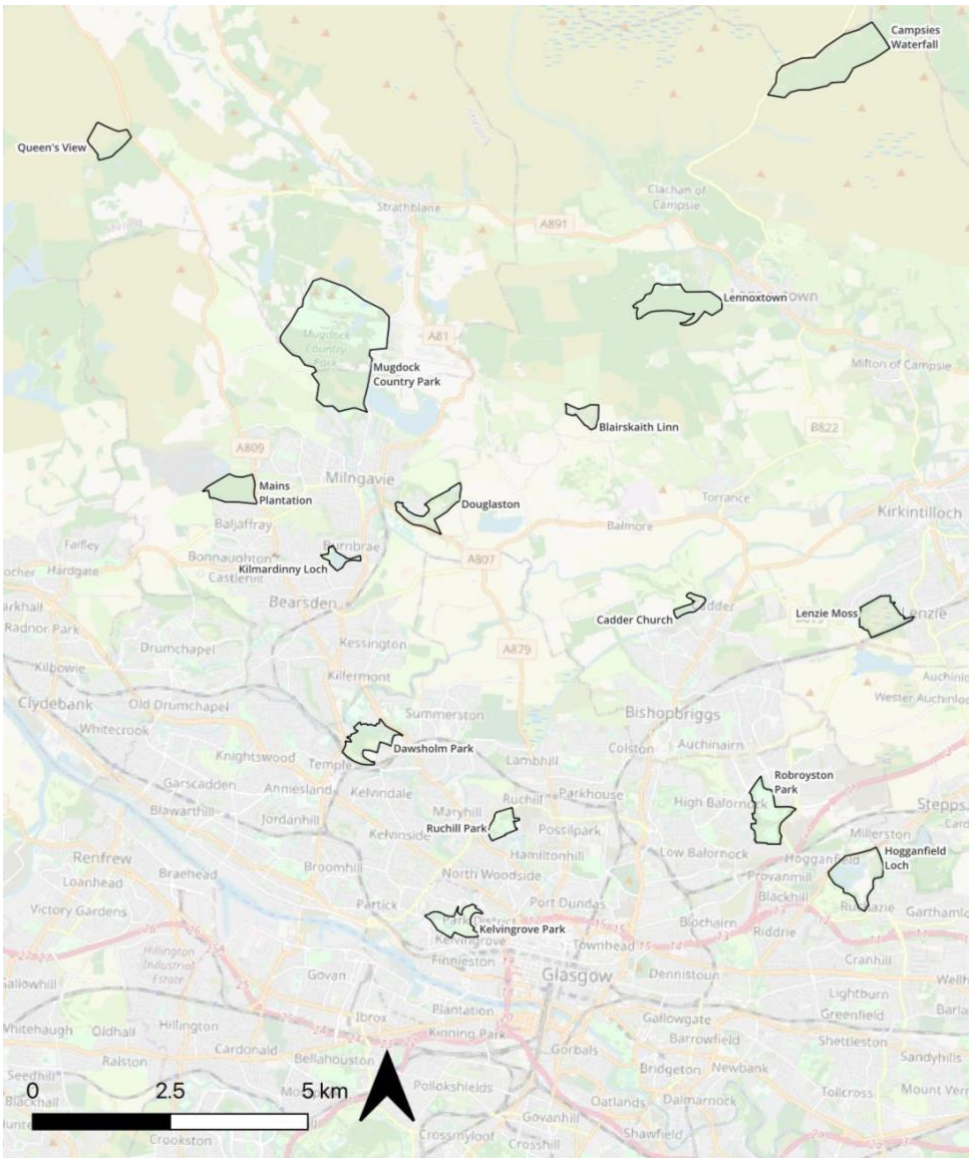

Figure 1 Greenspaces within Glasgow City where human behaviour was characterised.

2.2. Example application

Primary data about greenspace use was collected using an online survey. The survey was accessed by members of the public visiting the greenspace, via a QR code located at greenspace entrances and recorded the time the code was scanned (Figure 2). In total 711 respondents completed the survey.

The name of each greenspace was searched on Google, Tripadvisor, Facebook and Flickr and the number of Google Reviews, Tripadvisor Reviews, Facebook Check-ins, and Flickr images per search result was recorded. The Flickr API was used to record the number of images geolocated within the extent of each greenspace and the image timestamp. Data from Google Popular Times and the BestTime App was gathered for each greenspace to characterise timing of greenspace use.

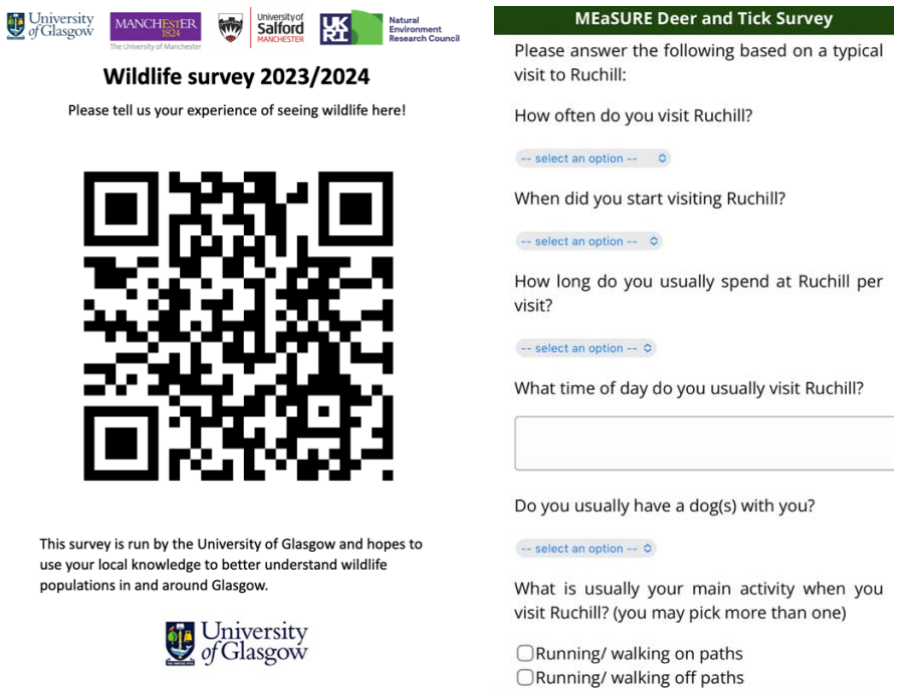

Figure 2 Survey used to characterise human behaviour, accessed via QR codes (left).

2.3. Model parametrisation

In our model, we needed to define parameters for phenomena occurring over fixed time periods: the day people visit greenspaces, which greenspaces they visit, and what hour they visit. The day of visit and greenspace are modelled as a Poisson distribution and hour of visit modelled as a multinomial distribution.

To parameterise the models, we combined count data of 'evidence' of greenspace visitation from the survey data and secondary online data, to determine the estimated visits to greenspaces per day of week, greenspace, and hour of day. For hour of day, data were combined using equal weights of one for each evidence (survey and Flickr timestamps, BestTime, Google Popular Times). For greenspace selection and day of week we use the principle of central tendency (in absence of information on error or bias) to determine weights of each evidence (Google Reviews, Tripadvisor Reviews, Facebook Check-ins, Flickr photos, survey responses, BestTime, Google Popular Times), with the weight for each evidence a function of distance to mean value of all evidence (**Equation 1**).

$$w_e = \frac{\sum_{g=0}^{|G|} |e_g - \hat{e}_g|}{|G|} \quad (1)$$

Where w_e is the weight for a piece of evidence (across all greenspaces), |G| is the length of the set of greenspaces, e_g is the value of the current evidence and greenspace, and $\hat{e}_g$ is the mean value for all evidence for the current greenspace.

The count data of 'evidence' and weights were then combined with the number of people in Glasgow who visit an urban greenspace at least once a week (96,064), estimated from the Scotland People and Nature Survey, and used to estimate visits per day, per greenspace and per hour (**Equation 2**).

$$n = t \sum_{e=0}^{|E|} w_e \; p_e \tag{2}$$

Where n is the estimated number of visits, |E| is the length of the set of evidence (i.e., number of data sources), $w_e$is the weight for the current evidence, $p_e$is the proportion of total visitors attending the greenspace according to the current evidence and t is the total visits to all greenspaces (96,064).

We then used the estimated visits as parameters for the Poisson and Multinomial distributions, from which we draw random values for visitor numbers per hour, day, and greenspace. To demonstrate our approach, visit numbers were generated 100 times. The approach is implemented in Python and is available in *Park Predictors*, accessible via https://gitlab.com/kirstywatkinson/park-predictors .

3. Results

Figure 3 shows the number of visits per day for 96,064 people. Weekend days have the highest number of visitors and is highest on Saturday (35,654). During weekdays, Thursday has the highest predicted number of visits (27,343).

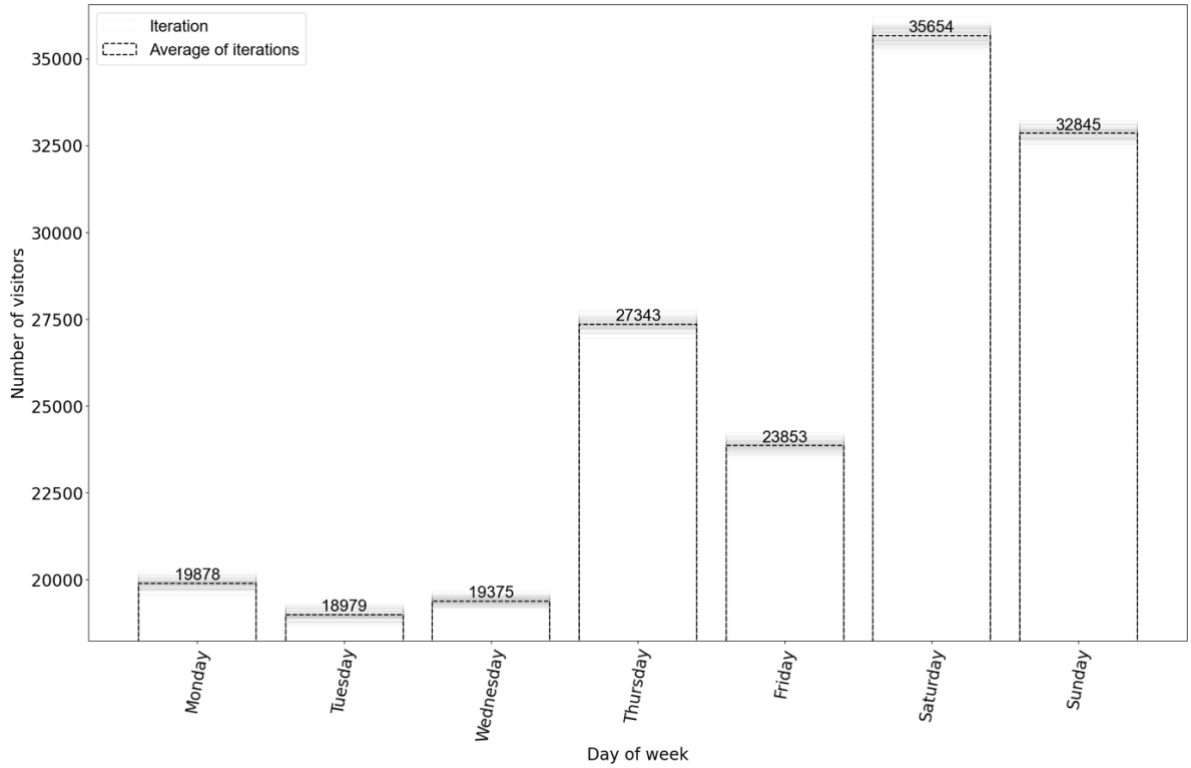

Figure 3 Daily predicted visits. For example, on a Monday, 19,878 human agents would visit a greenspace.

For Saturday, the number of visits per park is shown in Figure 4. Kelvingrove Park has the highest number of predicted visitors (15,249 out of 35,654). Greenspaces further from the city centre have the lowest predicted number of visits (e.g., Campsies), meaning relatively few agents would visit on a Saturday.

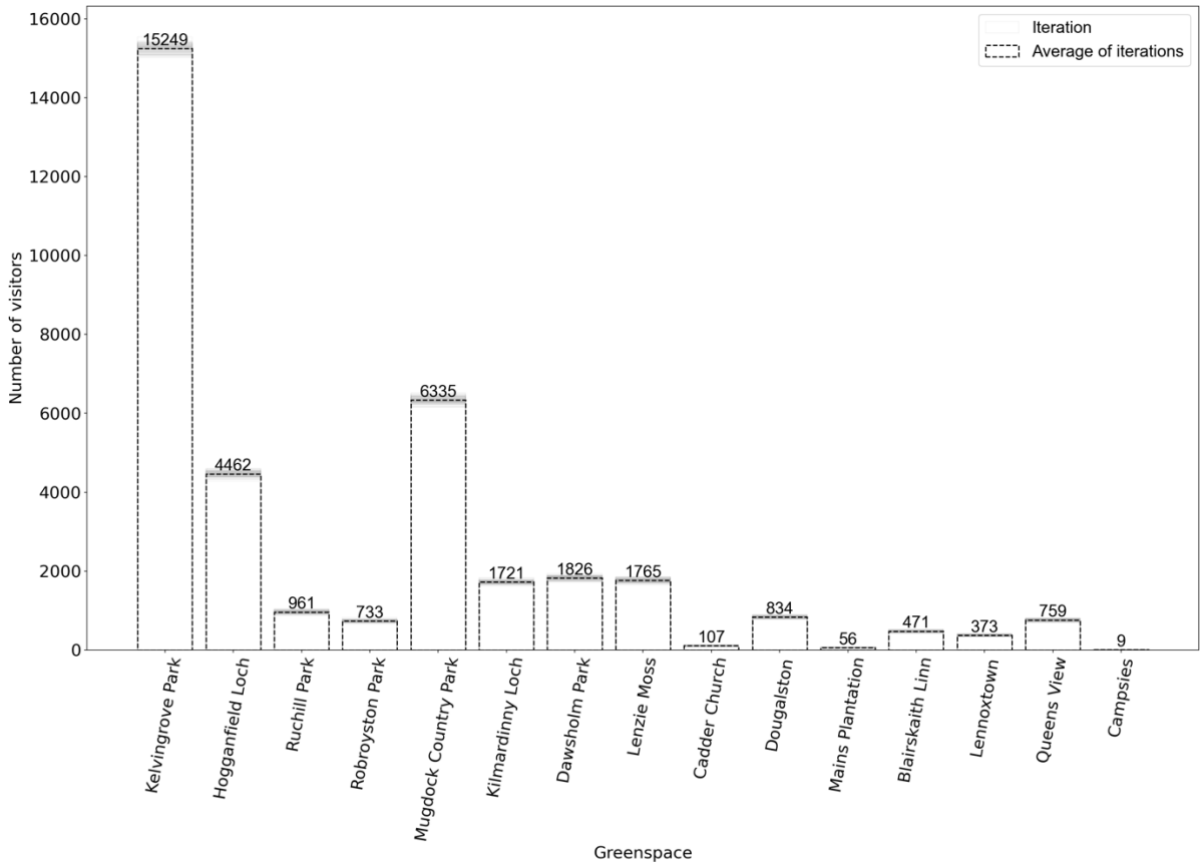

Figure 4 Predicted visits on a typical Saturday in the ABM. Most agents (15249) would visit Kelvingrove Park.

Using Kelvingrove Park as an example, the number of predicted visits per hour on a Saturday are shown in figure 5. Mid-morning (10:00-11:00) and mid-afternoon (15:00-17:00) have the highest predicted number of visitors and night-time hours (23:00-06:00) have the lowest predicted number of visitors.

The Poisson and multinomial predictors were integrated into the ABM and used to seed human agents at greenspace entrances (Figure 6).

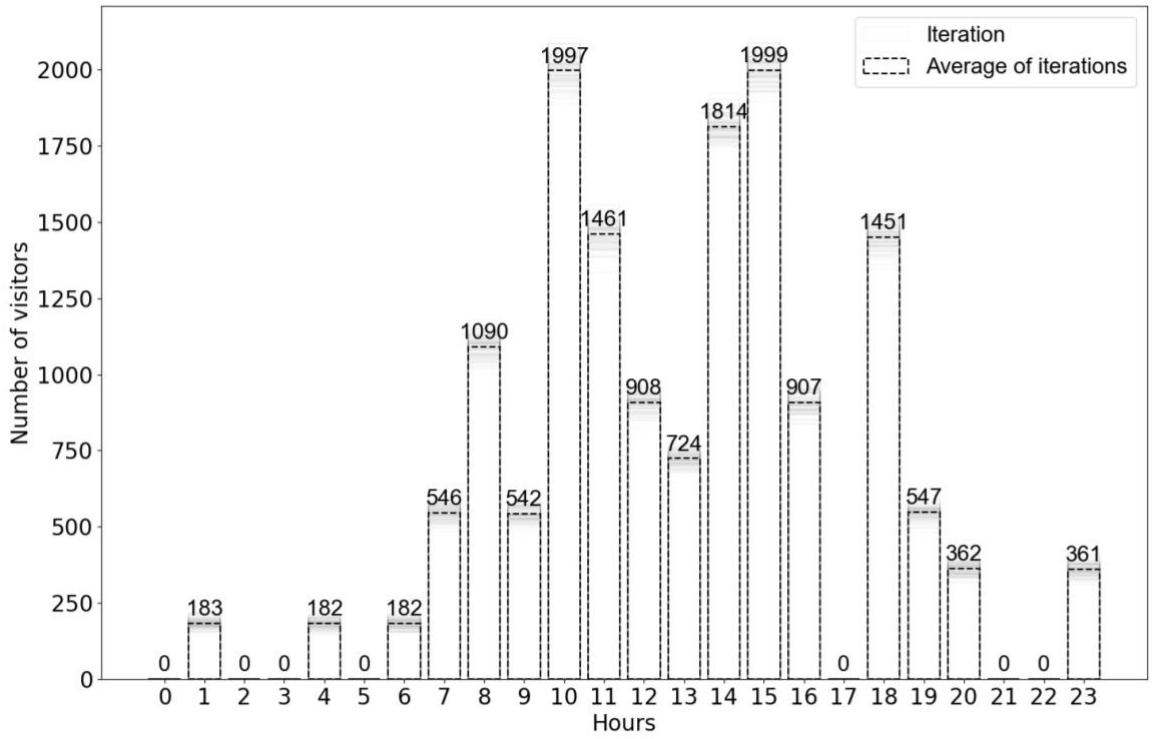

Figure 5 Predicted visits per hour in Kelvingrove Park on a Saturday. Most agents would visit during daylight hours (07:00-17:00).

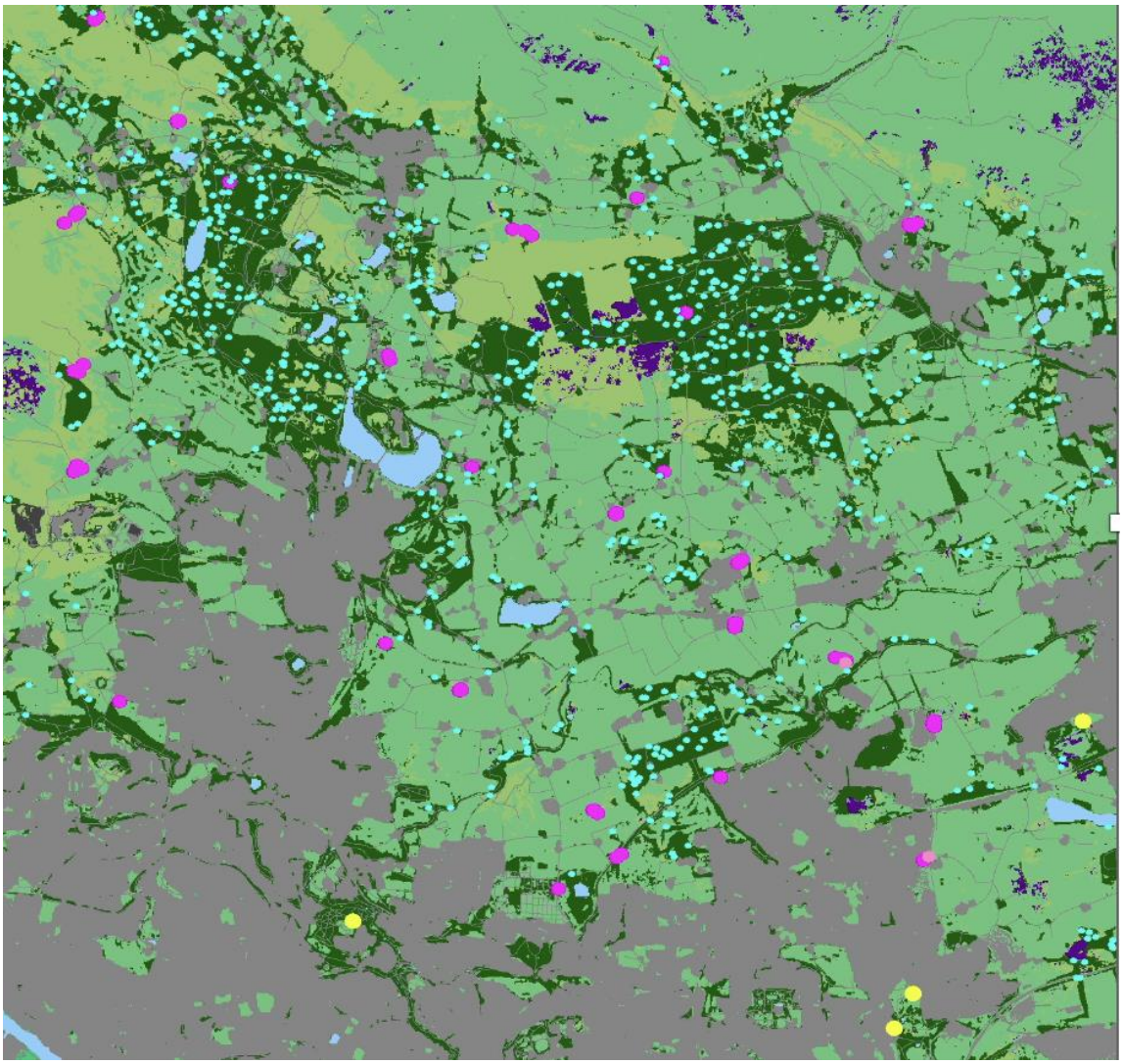

Figure 6 Humans (yellow) seeded at greenspace entrances in the ABM.

4. Conclusions

Human behaviour needs to be characterised for various research applications. Data can contain spatial and socio-demographic misrepresentations, making it important to use multiple datasets when characterising human behaviour. Here, we integrated several datasets and used them to build Poisson and Multinomial predictors to estimate visits to greenspaces. Daily predicted visits follow expected patterns, with higher visitor numbers at the weekend (when most people do not work). Predicted visits per greenspace also follow expected patterns, with more visits to popular tourist locations (Kelvingrove Park, Mugdock Country Park) and reflects the underlying data used (higher numbers of Google Reviews, Facebook check-ins, Flickr photos). Hourly predicted visits also follow expected patterns, with more visits during the day, compared to the night (when most people are asleep or avoiding greenspaces for safety). The approach used is a simple way to characterise human behaviour using data with varying levels of veracity and unknown uncertainties. Our implementation can be improved by incorporating more datasets and calculating weights for datasets used to predict hourly visits.

5. Acknowledgements

This research, conducted as part of the Maximising ecosystem services in urban environments (MEaSURE) project (https://nercmeasureproject.co.uk/), was funded by the Natural Environment Research Council (NE/W003120/1). Thanks to Dr Jessica Hall for drafting the survey questions.

Providers of data used in this paper are as follows: Flickr (https://www.flickr.com/services/api/) , Meta Platforms, Google LLC, BestTime (https://besttime.app/) , TripAdvisor LLC. This research benefitted from the following Python libraries: SciPy, NumPy, MatPlotLib, Pandas.

References

Gao, S., Janowicz, K., Montello, D.R., Hu, Y., Yang, J.-A., McKenzie, G., Ju, Y., Gong, L., Adams, B., Yan, B., 2017. A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science* 31, 1245–1271.

Huck, J., Whyatt, D., Coulton, P., 2015. Visualizing patterns in spatially ambiguous point data. *Journal of Spatial Information Science* 2015, 47–66.

Huck, J.J., Whyatt, J.D., Davies, G., Dixon, J., Sturgeon, B., Hocking, B., Tredoux, C., Jarman, N., Bryan, D., 2023. Fuzzy Bayesian inference for mapping vague and place-based regions: a case study of sectarian territory. *International Journal of Geographical Information Science* 37, 1765–1786.

Janowicz, K., Harmelen, F. van, Hendler, J.A., Hitzler, P., 2015. Why the Data Train Needs Semantic Rails. AI Magazine 36, 5–14.

Jepsen, J., Topping, C., 2004. Modelling roe deer (Capreolus capreolus) in a gradient of forest fragmentation: Behavioural plasticity and choice of cover. *Canadian Journal of Zoology* 82, 1528–1541.

Li, L., Goodchild, M.F., Xu, B., 2013. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. Cartography and Geographic Information Science 40, 61–77.

Li, S., Gilbert, L., Harrison, P.A., Rounsevell, M.D.A., 2016. Modelling the seasonality of Lyme disease risk and the potential impacts of a warming climate within the heterogeneous landscapes of Scotland. *Journal of The Royal Society Interface* 13, 20160140.

Sinclair, M., Maadi, S., Zhao, Q., Hong, J., Ghermandi, A., Bailey, N., 2023. Assessing the socio-demographic representativeness of mobile phone application data. *Applied Geography* 158, 102997.

Biographies

Kirsty Watkinson is a Research Associate for the MEaSURE project at the University of Manchester. Kirsty's role involves the development of an agent-based model and investigation of ecosystem service-urban relationships. Kirsty is also interested in volunteered geographic information and the implications of data scarcity on geographical analysis.

Jonathan Huck is a Senior Lecturer in Geographical Information Science in the Department of Geography at the University of Manchester. He is interested in the application of maps, GIS and emergent technologies to geographical problems, particularly in the areas of health, conflict and the environment.