Modelling Spatial Behaviour in Music Festivals Using Mobile Generated Data and Machine Learning

Luis Francisco Mejia Garcia[*1], Guy Lansley[†2] and Ben Calnan[‡3]

[1]Department of Civil, Environmental & Geomatic Engineering, University College London
[2]Department of Geography, University College London
[3]Movement Strategies

January 09, 2017

Summary

This study explores the utility of location data collected from a mobile phone app as a means of modelling spatial behaviour for consumer analysis, focusing on data from a music festival. Our aim was to harvest geo-temporal variables from the app data to model when individuals visit catering services across the site. Using Random Forest and Artificial Neural Networks machine learning algorithms, we presented an efficient means of simulating the popularity of bar areas within the festival site across time. The research demonstrates that with an appropriate methodology, mobile app data can provide useful insight for service provision planning.

KEYWORDS: crowd dynamics, spatio-temporal, mobile data, machine learning, feature engineering

1. Introduction

On-site catering facilities contribute a substantial proportion of the revenue made by large festivals and events. However, whilst good insight is required for effective service provision planning, data on consumers within festivals is very limited as their locations are not permanent. Although it is possible to achieve data linked to ticket sales, little is known about what services and facilities attendees will use within the site and at what times.

Some festivals now utilise mobile apps which provide attendees with real-time event information whilst they also routinely collect sophisticated spatio-temporal data from their handheld devices. New techniques could harvest patterns from phone apps by linking their spatial data to places and subsequently predicting how geo-temporal variables associate with visits to facilities across the site.

This study presents an exploratory analysis of newly available spatio-temporal data sourced from a mobile phone app made available to attendees at a popular music festival in the USA. The research demonstrates how machine learning techniques on geo-mobile data can model the popularity of catering facilities across time within a large event space.

2. Background

There is a long tradition of harvesting data on consumers for better location planning and stock distribution. Most work concentrates on linking consumers to their residential locations through either models or consumer databases built by more established retailers (Birkin *et al.,* 2004; Kitchen, 2014).

[*] luis.garcia.15@ucl.ac.uk
[†] g.lansley@ucl.ac.uk
[‡] bcalnan@movementstrategies.com

More recently, there has been a growing emphasis on understanding the spatial behaviour of consumers within a retail environment in order to make more intricate planning decisions (Marjanen, 1995).

To model crowd dynamics some practitioners have considered new forms of data which arise from handheld devices. Previous research has found mobile phones to be a useful indicator of spatio-temporal population movements, whereby locations of devices are triangulated from the nearest cell phone towers (Gonzalez *et al.*, 2008). It is also possible to harness more precise location information from devices' GPS capabilities when data are collected via third party apps. Social media apps, for example, are an excellent source of geo-social data which can be used to infer activities across space and time (Lansley and Longley, 2016). Given the large volumes of individuals concentrated within a relatively small area, new technologies which can be used to predict crowd dynamics could be particularly valuable for festival planners.

3. Data

The data was collected from a large festival which took place in California over a weekend in August 2015. Ticket scan data confirmed that there were 29,000 attendees on Saturday, and 26,000 on Sunday. The festival app was made available to all attendees and it regularly recorded the location of users within the site during both days.

In total, locations for 9,300 devices were recorded on Saturday and 7,600 on Sunday. Assuming each individual only installed the app on one device, the data represents about 32% of festival attendees. The app used the mobile phones' positioning systems to record users' locations every 5 seconds and produced a vast database of over a 100 million time-stamped records. Although we are considerate of the impacts of variable signal reception, battery life issues and organisational factors, the data was assumed to be a representative indicator of crowd dynamics during the festival. Figure 1 presents four heat maps of the footprints recorded at different times across Saturday.

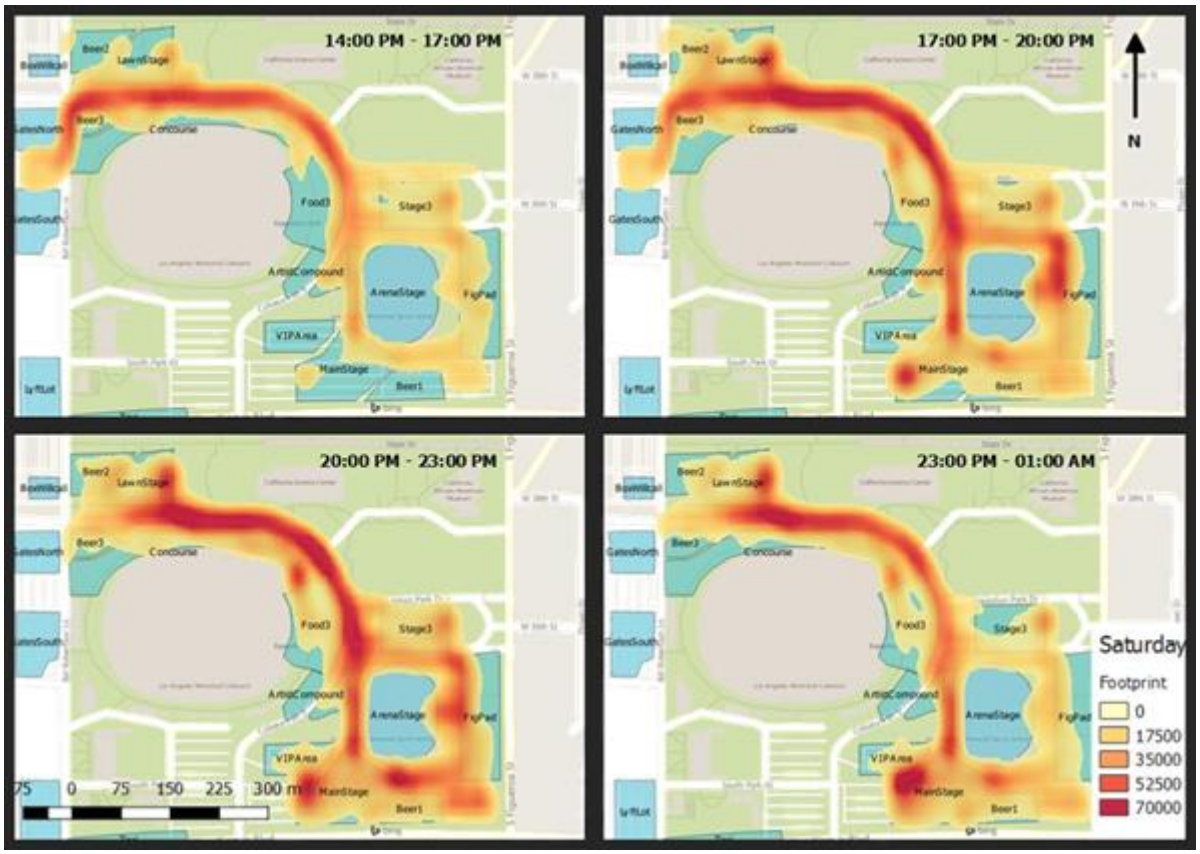

Figure 1 Heat map of the footprints in the festival area on Saturday

4. Methodology

We utilised a pipeline methodology which pooled a number of techniques to transform the data into useful variables and then detect patterns in order to predict the occupancy of bar areas across time (Figure 2).

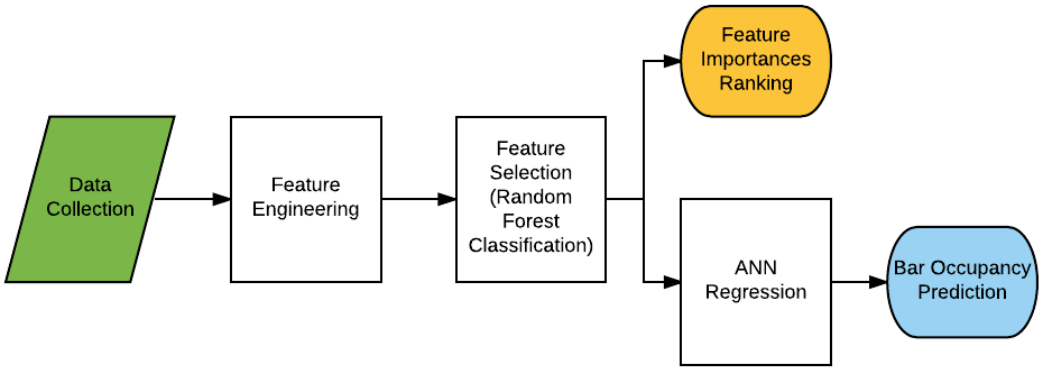

Figure 2 The key methodological steps

The first objective was to contextualise the time-stamped location data so it was possible to provide inferences from the data. Place information on the festival site was spatially joined to the data, this included the location of bars and performance areas. This was possible due to the high precision and accuracy of coordinates produced from mobile devices. The schedules of artists performing at each stage were also linked. Devices which were located within any of the bar areas were flagged in order to create a bar occupancy variable.

The app data could then be combined to form a number of features which can be used in machine learning algorithms (Zhu *et al.*, 2013). This feature engineering process, along with the subsequent feature selection, improves the efficiency of machine learning algorithms and makes their interpretation much easier (Hasan *et al.*, 2016). In this case, 6 numerical variables (features) were calculated and were represented in checkpoints, every 30 minutes, along the 2 days of the festival (Table 1). Each feature should have some association with bar occupancy habits of each user.

Table 1 The generated features

Feature	Description
Time spent in the festival	The duration of time at the festival until the checkpoint
Next artist seen	Numerical representation of the popularity[§] of the artist that performed on the next stage visited by the user.
Last artist seen	Numerical representation of the popularity[§] of the artist that performed on the previous stage visited by the user
Last time in the bar	The duration of time between the most recent visit to a bar area and the checkpoint
Distance to the nearest bar	The Euclidean distance from the last place where the user was located to the closest bar area
Gender	This variable is obtained through the social media account linkage provided through the mobile application.

[§] Created from metrics produced by Next Big Sound: https://www.nextbigsound.com/ (2016)

Machine learning algorithms were subsequently trained using these features to identify the most useful features for estimating visits to the bar areas across the festival. The feature selection stage ranked the influence of the created features through a classification model with a random forest algorithm. Random forest is a suitable machine learning approach for feature selection, especially in datasets with high dimensionality with multiple input variables, as it can identify relevant features with the presence of noise (Breiman, 2001). The classification process revealed the required information to rank the features by their importance using the Gini index for node purity. This process was executed to include all of the checkpoints across both days.

Finally, a regression model created with an Artificial Neural Network (ANN) was implemented to estimate the number of people located in any of the 4 bar areas at given checkpoints. ANNs are computational simulations of the human brain, using multiple neurons forming synapse connections. The weight of the connections between neurons depends on the contribution they have to the final output (Verlinden *et al*., 2008). They are useful for predictions based on large numbers of variables and high-level non-linear relationships. The model did not include the gender feature as this was deemed to be insignificant by the feature selection process. To test the utility of the data and the methods from this research, the algorithm was trained with 85% of the data with the intention of testing the model on the remaining 15%. The testing data represented all checkpoints from 16:00 to 19:00 on the Sunday.

5. Results

Figure 3 displays the relative influence of the 6 features across Sunday. The two most influential features are the time spent at the festival and the time since the last visit to the bar. The distance to the nearest bar feature was also influential, particularly coming up to the closing hour of the festival. The popularity of the next artist spectated was also useful, implying that there is a slight surge in sales from attendees prior to watching some of the more popular acts. There are also slightly more subtle peaks in visits following the performances of large acts.

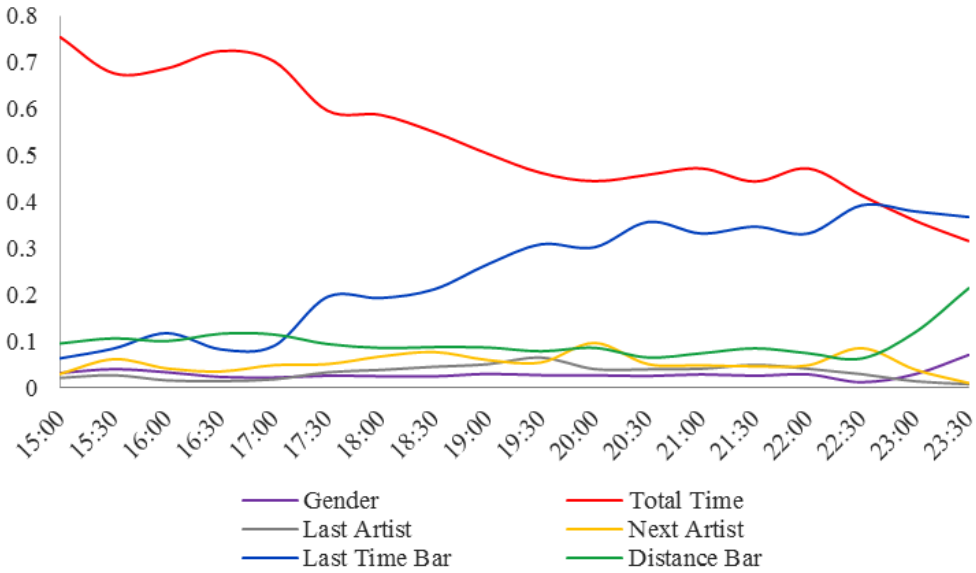

Figure 3 Feature Importances on Sunday

The predictive model created from the artificial neural networks was used to simulate bar occupancy between 14:00 and 19:00 on Sunday. The model was found to be within 75% of the actual results recorded from this time period despite the fact the model was not calculated with any data recorded during this time-frame on a Sunday. Figure 4 displays the predicted occupancy rate across all four bar

areas within the site. The model identifies the distinctive dip in visits at 17:00.

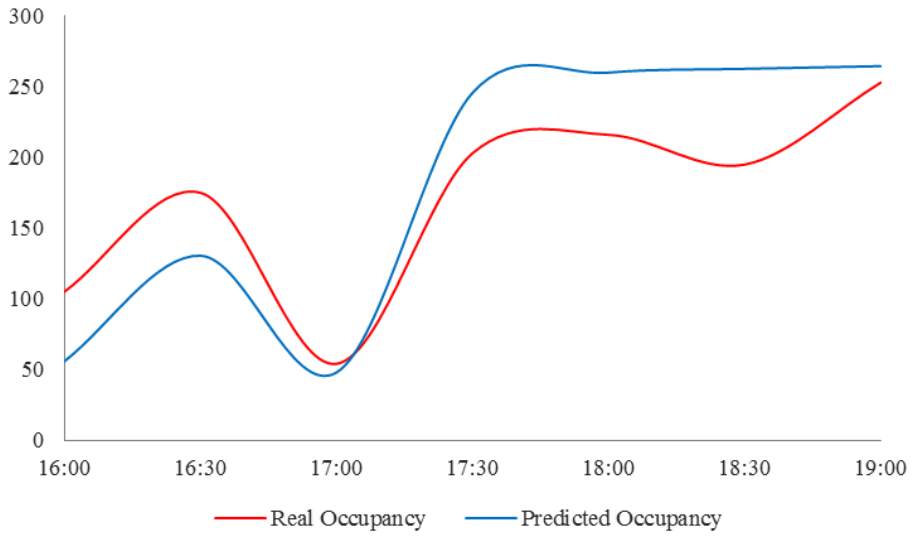

Figure 4 Bar occupancy prediction by ANN on Sunday afternoon

The model indicates that there are both spatial and temporal elements which may influence an individuals decision to visit the bar areas. Using this information planners could change how they allocate their services across space and time. For instance, it was possible to quantify the influence of distance to the frequency of visits to bar areas. An average decrease of 30% of the distance between the users and the bar areas across the whole sample could increase the occupancy by 7%.

6. Conclusions

This research has demonstrated the utility of spatio-temporal data created from handheld devices as a means of estimating crowd dynamics. By applying appropriate geographic heuristics and utilising machine learning applications, useful information on broader trends and their possible implications for the retail environment can be determined from large samples of mobile data.

Feature engineering is an appropriate tool for deducing geo-temporal variables useful for modelling simple geospatial behavioural patterns. In this analysis, bar occupancy was a proxy for consumer behaviour as it was not possible to acquire catering sales data. The subsequent models were able to successfully predict temporal trends in occupancy across the festival site despite being based on a small number of features and within a very short time-frame. Whilst this exploratory project converted the wealth of data into relatively rudimentary variables, the successful results demonstrate that there is a real opportunity for companies to implement similar processes in other scenarios.

7. Acknowledgements

This work is funded by the UK ESRC Consumer Data Research Centre (CDRC) grant reference ES/L011840/1. We would like to thank the third-party data provider for allowing us to undertake research on their data.

8. Biography

Luis Francisco Mejia Garcia is a Masters student in Spatio-temporal Analytics and Big Data Mining based at University College London (UCL). He has 5 years of experience in software development in

leading technology companies such as IBM, AT&T and TCS. Luis has a bachelor's degree in Mechatronics Engineering from the Instituto Tecnologico de Estudios Superiores de Monterrey in Mexico.

Guy Lansley is a Research Associate at the Consumer Data Research Centre and the Department of Geography, UCL. His previous research at UCL has included exploring the temporal geo-demographics derived from social media data, and identifying socio-spatial patterns in car model ownership in conjunction with the Department for Transport. His current work entails exploring population data derived from large consumer datasets.

Ben Calnan is a managing consultant at Movement Strategies, a leading consultancy in people movement and crowd dynamics. He has interests in web mapping, geo-temporal demographics and infographics. Ben has over 11 years of experience working with GIS products and holds a Masters Degree in Geographic Information Science from UCL.

References

Birkin, M., Clarke, G., Clarke, M., Culf, R. (2004). Using spatial models to solve difficult retail location problems. In: Stillwell J, Clarke G (eds) *Applied GIS and spatial analysis*. Wiley, Chichester. 35–54.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1); 5-32.

Hasan, M., Nasser, M., Ahmad, S. and Molla, K. (2016). Feature Selection for Intrusion Detection *Using Random Forest. Journal of Information Security*, 07(03); 129-140.

Gonzalez, M.C., Hidalgo, C.A. and Barabasi, A.L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196); 779-782.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, London

Lansley, G. and Longley, P. (2016) The geography of Twitter topics in London. *Computers, Environment and Urban Systems*. 58; 85–96

Marjanen, H. (1995). Longitudinal study on consumer spatial shopping behaviour with special reference to out-of-town shopping. *Journal of Retailing and Consumer Services*, 2(3); 163-174

Verlinden, B., Duflou, J., Collin, P. and Cattrysse, D. (2008). Cost estimation for sheet metal parts using multiple regression and artificial neural networks: A case study. *International Journal of Production Economics*, 111(2); 484-492.

Zhu, Y., Zhong, E., Lu, Z. and Yang, Q. (2013). Feature engineering for semantic place prediction. *Pervasive and Mobile Computing*, 9(6), pp.772-783.