Off the Beaten Path: An analysis into the visiting habits of Londoners and tourists

Flavio Ribeiro de A. F. Mello

1. DOMAIN INTRODUCTION

London is a global city both in economical and touristic terms. It can be said that both local and tourist populations compete for limited space, specially in the city centre which hosts a number of world famous tourist attractions and a large portion of the city's jobs, offices and services. From a brick and mortar business standpoint, one may gain competitive advantage by better understanding the surrounding region's footfall profile in terms of mobility trends of tourists and residents or, more profoundly, what type of business attracts each population group. During the last decade, smartphones with GPS have progressed from luxury to ubiquitous. Social media posts are automatically geotagged and timestamped and can be used to provide insight into demographic spatial movement.

With this motivation, this study sets out to answer the following questions:

- Is it possible to detect spatial differences in central London visiting habits of tourists and residents based on social media geotagged and timestamped posts?
- Do these differences vary over time?
- Is it possible to establish a relationship between said differences and business categories present in each region?

For tackling these questions two main datasets will be used: i) London Tweet residents vs tourists dataset [7], as provided by City University Teaching Staff, containing data from tweets inside Greater London from 05/11/2012 to 24/10/2013. Although the tweet dataset represents data collected over the course of a year, the data is already aggregated into hours of the week, thus it is not possible to detect trends that emerge over timespans longer than a week. ii) Business location and metadata, obtained through Yelp's Fusion API[5]. There will be need for some manipulation in both datasets. Combining residents and tourists tweet datasets and have the relevant locations selected. While the business location will need to be consolidated through Yelp's Fusion API[5], linked to the region divisions provided by the tweet dataset, and have geographically weighted values extracted from the resulting links.

1.1 Terminology

Throughout the study, the following terminology applies:

Resident: A visitor which is identified as a London resident by having tweets inside London in **at least** 30 different days during the observed period.

Tourist: A visitor which is identified as not being a London resident by having tweets inside London in **less** than 30 different days during the observed period.

Location: Each of the 606 Voronoi polygons across greater London the dataset uses to bin tweet events.

2. ANALYTICAL TASKS AND IMPLEMENTATION APPROACH

After the preprocessing steps mentioned above, the study proceeds to investigate the data with 3 progressive tasks:

2.1. Initial Investigation

Description: Display choropleth maps of central London with locations colored by tourist versus resident visitor percentage using a diverging color scheme to evidentiate the degree each location skews towards either population.. Display maps for different time aggregations to investigate whether there is any kind of relationship between time and visitor distribution trends. As shown in lecture 5, the mixture of time and space variation can be complex to observe in a single image, thus the study relies on the representation of time-varying spatial situations to indicate presence of time influence in the location's population distribution.

Computational Component: Calculate for each location the percentage of tourist visitors relative to the total visitors registered over all the time spans mentioned above. This effectively normalizes the values with respect to total visitors per location, a step needed to account for varying magnitudes of total visitors per location, as seen in lecture 8.

Visual Component: In the map representing the complete time span, search for patterns shown in lecture 3. Hot/cold spots, clusters and linear trends. In the sequential maps representing time slices, search for changes between slices: locations that change hue, representing a change in the majority of visitors; and locations that change intensity, becoming either more or less dominant for either population. At this stage, it is also interesting to verify whether the results fit known/expected trends (e.g.: Do tourist hotspots such as Big Ben have a tourist majority? Do known business centres like City of London have a resident majority?)

2.2. Clustering

Description: Clustering can be an efficient method in gaining insight into spatio-temporal data, as shown in lecture 5 and its practical exercise. In order to surface underlying temporal patterns have each location represent a week-long time series and cluster them by similarity of temporal visitor distribution with the intent to find locations with similar visitor distribution trends.

Computational Component: Calculate tourist to resident ratio for each location/hour pair in the complete dataset. Apply k-means clustering to the resulting data to find locations(rows) with similar hour by hour (columns) weekly distribution trend.

Visual Component: Run clustering algorithm with different number of clusters(n_clus) and inspect the results by displaying all locations'

time series along with each cluster centroid, color-coded by cluster. Choose number of clusters which results in groups with noticeably different trends, without having redundant, or too similar, clusters. For the chosen value of n_clus, revisit the time-series with centroid plot and interpret the results. Display classified locations in a choropleth map using a qualitative color scheme to indicate cluster id. Interpret the spatial distribution of the clusters in the context of the temporal differences discovered in the time-series cluster plots.

2.3. Feature Selection & Classifier Tuning

Description: Attempt to evidentiate relationship between location's weekly visitor trends (i.e. classified cluster) and their business profiles by modeling a classifier which predicts a given location's class based on business data. Select candidate features to be used in said classifier with the aim to reach a feature set which correctly matches the studied phenomena, while being the simplest model possible. Resulting feature set is then used to fit a classifier, and its performance evaluated visually. Depending on the results, repeat feature selection process and compare.

Computational Component: For each location, calculate occurrence distribution of categories and price levels. Calculate correlation between groups of selected candidate features. Classifier training, prediction and scoring.

Visual Component: For each candidate feature, display unidimensional scatterplots of locations per feature value, color-coding points by location's cluster. Look for features that provide clear separation between clusters. Assess selected features through the correlation table, removing redundant features. Run model with selected features and evaluate classification results in terms of prediction confidence and correct classification for each class(cluster), using a visualization similar to Ren et al[1]. Revisit model accordingly.

3. ANALYTICAL STEPS

3.1. Initial Investigation

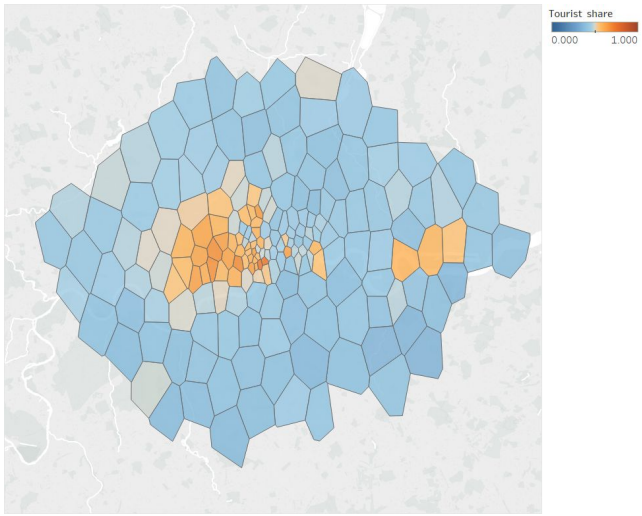

Figure 1: Total tourist visitor activity share aggregated throughout the whole year. Orange means tourist majority, blue mean resident majority

Initial investigation in the dataset shows clear spatial separation of tourist and resident regions. In general (figure 1), the tourist majority areas are located in the western portion of zone 1[3].

Further comparing the tourist to resident ratio during the week versus the weekend (figure 2) shows increased tourist to resident ratio towards east central London in the weekends. This indicates that there may be increased touristic or decreased resident activity during the weekend in these locations.

Figure 2: Tourist activity share Mon-Fri (top) vs weekend (bottom)

Over the course of a day (figure 3), there seems to be no significant changes in tourist to resident activity, with the sole exception of Blackwall.

In general, the results are consistent with known tourist hotspots in London's west end, and indicate that there is some variation to the tourist to resident activity over the week, thus legitimating further study in the time aspect of touristic activity.

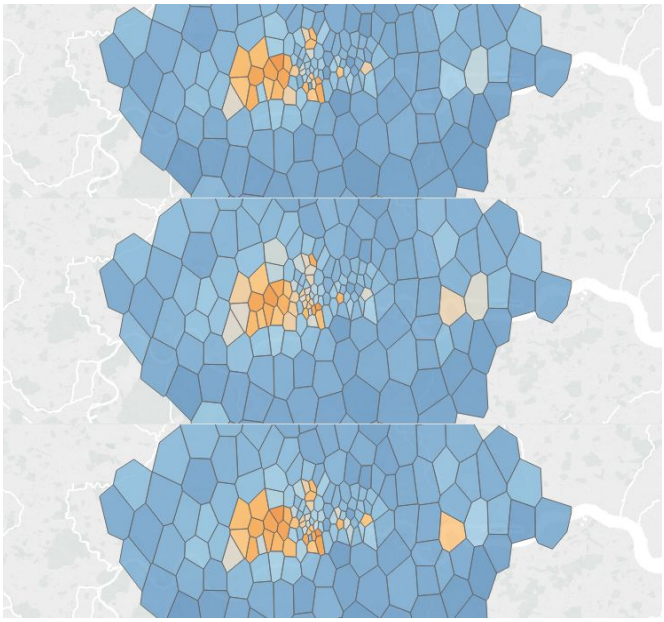

Figure 3: Tourist activity share. Morning (top), Afternoon (mid), Night (bottom)

3.2. Clustering

After running the clustering process multiple times, the optimal number of clusters was found to be 5. From the clustering results (figure 4) following interpretations were extracted:

Clus_0: Mostly resident, with little change in activity ratio during the day or weekend relative to working days.
Clus_1: Equal share of tourist to residents, with heavy increase of tourist activity during the weekends.
Clus_2: Mostly residential with more noticeable increase in tourist activity during afternoons and the weekend.
Clus_3: Predominantly resident, with noticeable increase in tourist share in the afternoon and during the weekends. Some locations may flip to tourist majority during the weekends.
Clus_4: Predominant tourist share during the whole week.
Apart from the cluster interpretations, there are two general trends observed in all clusters: increased tourist activity share in the afternoons and weekends; and increased resident share in late night/early morning activity

As seen in Figure 5, there are some clear geographical trends in the cluster allocation. The eastern section of zone 1 is dominated by Londoner activity, as are zones 2 and 3. The western portion of zone 1 hosts predominantly tourist locations or locations that become touristic during the weekend. Central zone 1 is mostly visited by Londoners during the week, with equal share during the weekend.

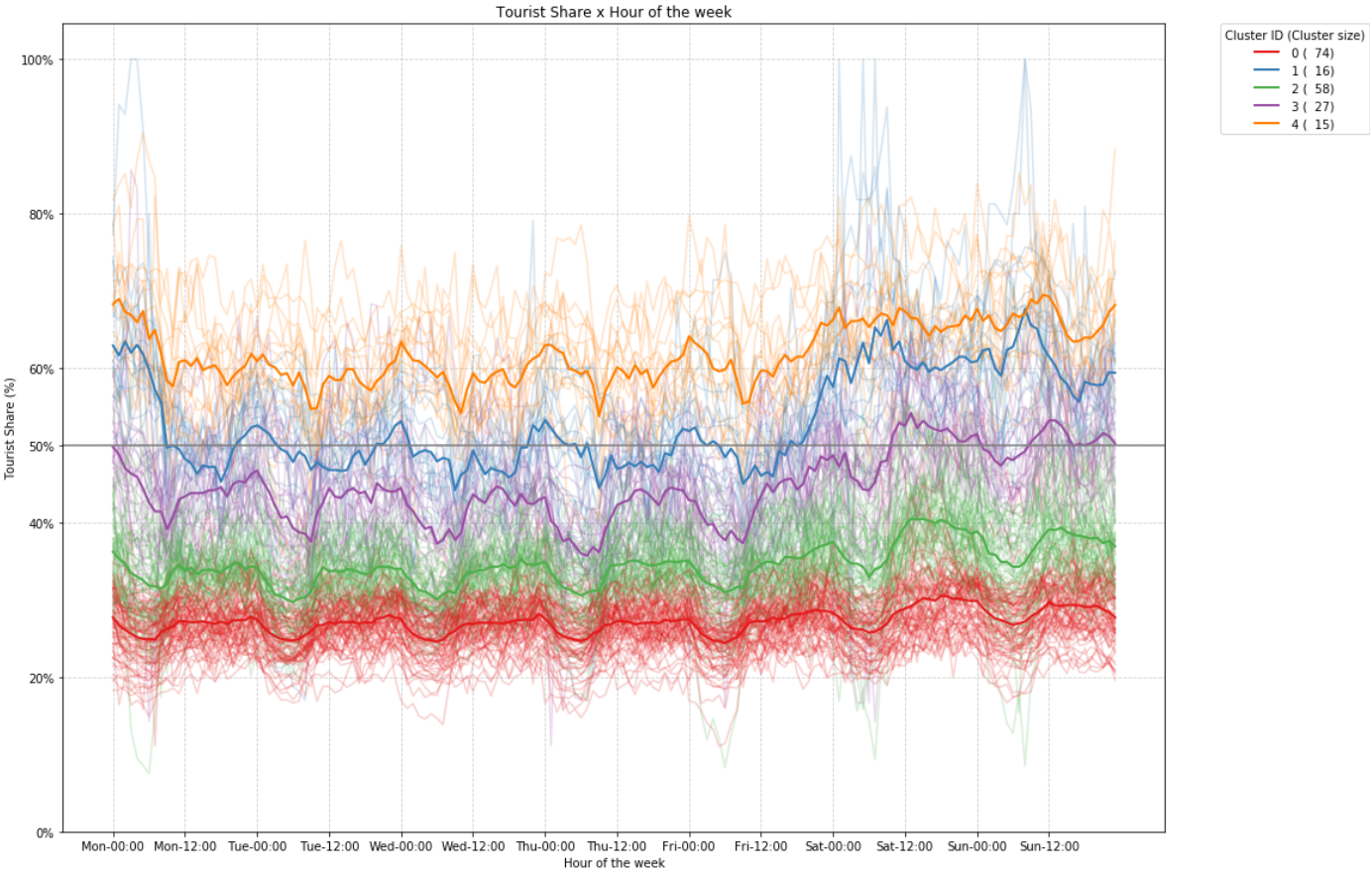

Figure 4: Tourist activity share time-series clustering results. Each line represent a single location, thicker lines represent cluster centroids

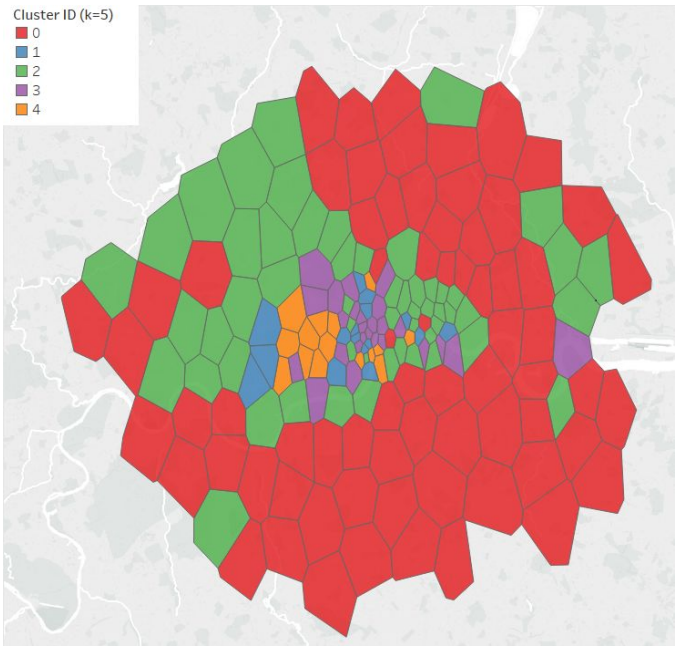

Figure 5: London locations colored by temporal trend clusters

3.3. Feature Selection & Classifier Tuning

The feature selection process started with electing features based on possible hypothetical explanations. Table 1 shows the initially selected features along with their devised explanations. All features are calculated as their representative share in all businesses found in said location.

Feature	Hypothesis
Hotels & Travel	Temporary accommodation is a necessity for tourists. Regions with more hotels could be related to increased touristic activity.
Local Services	Services like engraving and notaries cater mostly to local population.
Restaurants	Tourists usually have to resort to restaurants for every meal while on vacation. Residents may use cooking facilities in their own home.
Arts & Entertainment	London is internationally known for its museums which consistently rank in the city's top attractions in travel guides.
Food	Grocery shopping is mostly done by locals
Home Services	Tourists are usually hosted in a temporary accomodation, home-related services are not expected to be of their interest.
Shopping	Shopping is considered a touristic activity by several tourism guides. Locals have more opportunity to resort to online shopping.
Expensive (Price Level £££ or ££££)	Touristic places are commonly more expensive than their local counterparts.

Table 1: Candidate features and hypotheses

As an initial investigation into the features' suitability, the pairwise correlation was calculated for all selected features (table 2). To keep the model parsimonious, both *Shopping* and *Food* features were dropped for having somewhat stronger (>0.2) correlation to two other features.

	Arts	Food	Home Services	Hotels & Travel	Local Services	Restaurants	Shopping	Expensive
Arts	1.00							
Food	-0.12	1.00						
Home Services	0.04	0.05	1.00					
Hotels & Travel	-0.09	-0.24	-0.10	1.00				
Local Services	-0.10	-0.22	-0.05	-0.07	1.00			
Restaurants	-0.02	0.12	-0.03	-0.09	-0.28	1.00		
Shopping	0.19	0.11	0.14	-0.32	-0.06	-0.20	1.00	
Expensive	-0.11	0.11	0.08	-0.09	-0.01	0.37	0.08	1.00

Table 2: Candidate features correlation matrix

To further assess the features' suitability for classifying locations, unidimensional scatterplots were produced indicating locations' distribution relative to each feature. Visually inspecting the resulting plots indicates that, individually, the selected features are not able to consistently separate the locations into cohesive groups.

In order to test the feature's suitability for, as a group, classify locations according to their temporal trend, the investigation then proceeded to fit and predict locations' clusters based on their business related data. Due to the small sample used in the prediction process (under 200 locations) the validation methodology used was Leave One Out - each location is predicted once, with all the rest of the locations used to fit the model. This ensures that for each prediction the training set contains observations of all possible location classes.

The locations were classified using a random forest and the prediction results displayed visually in barcharts according to the prediction probability, and whether the prediction is correct or not, in a structure similar to Ren et al[1]. The hyperparameters selected for tuning were number of trees, minimum samples per leaf, and

Figure 6: Cluster distribution per feature. Note the absence of clear separation between clusters for every feature.

maximum features sampled for each tree grown. They were selected for being representative in Random Forests' observed performance[4].

Unfortunately, despite running the model with different combinations of features and hyperparameter values, the model continued to perform poorly. Figure 7 shows the best observed performance, achieved with the features mentioned above and the following hyperparameters: 5 maximum features, 2 minimum samples, 200 trees. Note how predictions are consistently done with low confidence (probability < 60%) and often mistaken. Performance was the worst when attempting to identify locations belonging to clusters 1 and 4, which are the ones with most relevant tourist activity. Taken together, these observations indicate that, in the current setting, local business metadata is not a good predictor of tourist to resident visiting habits.

4. FINDINGS

Overall, this study was able to surface spatio-temporal patterns in tourist to resident activity in London. Significant touristic activity is mostly confined to TFL's zone 1[3], more specifically the region west of Charing Cross. Overall there is significant increase in tourist activity share during the weekends. Over the course of a day, relative tourist activity seems to be higher in the afternoon, while early mornings and late nights have more relative resident activity. The study was not able to correlate the spatio-temporal patterns in terms of business category and price level.

Revisiting the original questions:

Is it possible to detect spatial differences in central London visiting habits of tourists and residents based on social media geotagged and timestamped posts?

Yes, it was possible to detect significant differences between tourists and residents in location visiting habits in London based on social media data. zones 2 and 3 are dominated by Londoner activity, while central London is divided east/west into predominantly Londoner and Tourist activity, respectively.

Do these differences vary over time?

Yes, there were observed two main time related trends over the course of the week. The first is a small, daily trend, which shows increased relative tourist activity in the afternoon; and a second, much more significant one, indicates increased tourist activity share during the weekends. This effect is particularly strong in a few select regions which observe the tourist share increase by 10% of the total visitors. Zones 2 and 3 remain hosting mostly Londoner activity, while zone 1's central locations experience significant tourist share during the weekends. The only location outside zone 1 to experience pronounced increase in tourist activity is Blackwall, this could be related to the The O2 Arena which often hosts major international artists.

Is it possible to establish a relationship between said differences and business categories present in each region?

No. No strong relationship was found between categories and temporal visitor patterns.

5. CRITICAL REFLECTION

5.1 Direct Implications

This study surfaced relevant trends in tourist activity. Even if it was unable to correlate such trends in terms of business category and metadata, from a business standpoint it is still valuable to understand which locations are majorly touristic all the time, which locations see a large influx of tourist share in the weekends, and which locations are dominated by Londoners most of the time. The "occasionally touristic" locations present a particular challenge in that the business may need to cater to different populations with different needs.

One possible cause for the apparent lack of correlation between tourist activity and business type could be the granularity of the categories used in the study — the topmost, thus more generic, categories in Yelp's category list[6]. Another possible cause could be because of London's zoning policies, with many neighbourhoods mixing residential and multiple types of commercial activities. Perhaps in cities with different zoning policies — more strictly defining allowed commercial activities per region — there would be more separation in terms of business categories found in each region.

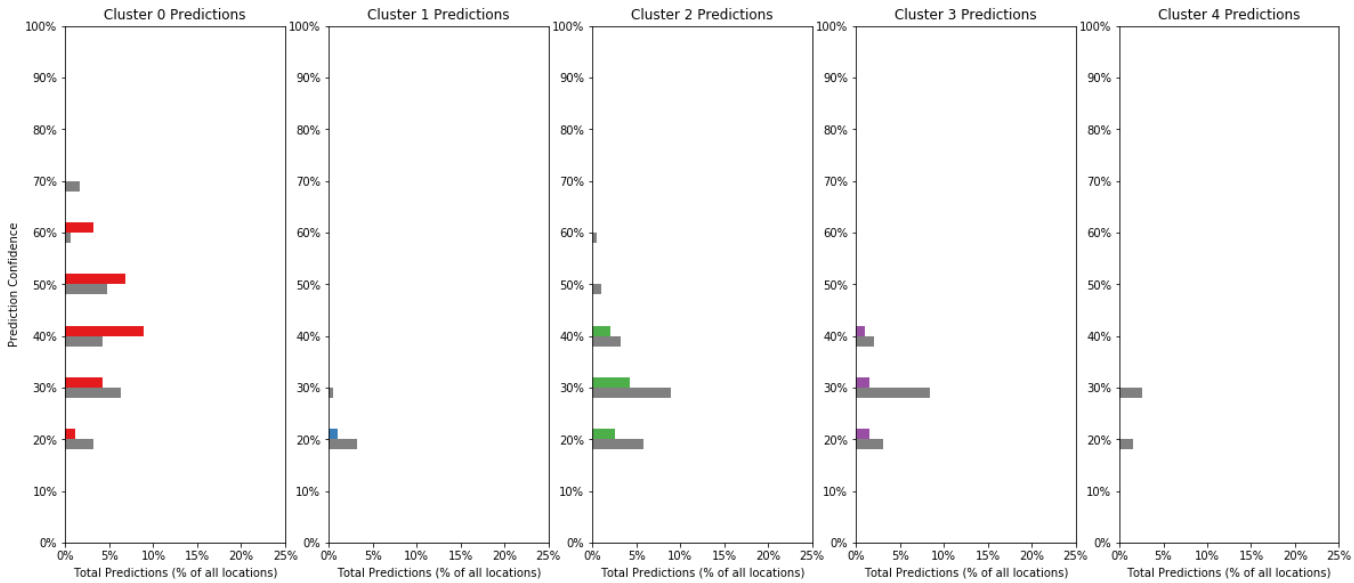

Figure 7: Prediction confidence distribution per predicted cluster. Colored bars indicate correct predictions, gray bars indicate incorrect predictions.

Both hypotheses remain open to further studies to test their validity.

Finally, a particular challenge with the classification problem is the small sample size, considering the studied area consists of only 190 locations, having some clusters with less than 20 observed locations, it becomes difficult to properly train a classifier unless there is a blatant separation in terms of the features selected. Predictions for clusters 1 and 4 were particularly poor, coincidently those were the clusters with fewer locations. One way to deal with this issue would be to perform location binning using smaller areas, however this has a limit as having too small of a granule may lead to loss of cohesiveness in terms of neighbourhoods.

At this stage, it is important to note some biases present in the data used for the study. The first is regarding the population studied: being a commercial technological platform, Twitter users do not represent the entirety of the population[2], thus the trends observed in this study represent trends of twitter active population. If there is need to account for the complete population, the sampling process should be demographically adjusted to account for this bias. The second, similar, bias is that the sampled population may not accurately represent the totality of twitter population. The dataset contained no indication of the sampling methodology used. Another bias present in the data is regarding the Yelp dataset used: the same way Twitter users are not a demographically accurate subset of the entire population, the businesses present in Yelp may not be perfectly representative of the totality of businesses found in London. It is perfectly possible for particular types of businesses to be absent from Yelp's database.

Finally, it is important to emphasize that the study set out to find correlation, and not full explanation, between visitor trends and businesses metadata. That is because relying only in the data used, it is impossible to discern the direction of the causality relationship. Are businesses located in particular regions due to increased touristic activity there, or are tourists in fact attracted to particular locations because of their offerings of businesses availability?

5.2 Visual Analytics Components

The visual analytics approaches used in this study were crucial for reaching the findings mentioned above. The constant dialogue between computational findings and visualization, specially in the form of choropleth maps were important to assess whether the results match known patterns, surfacing possible missteps in the process, which were promptly corrected. The initial investigation provided important guidance in terms of which kind of temporal study should be made — it was decided to focus on the weekly trend rather than daily trends after observing more expressive variation between working day versus weekend rather than morning, afternoon and night. Finally, at the model building stage, there was a constant dialogue between selecting features and tuning hyperparameters; and visualizing the ensuing prediction confidence results. It was this cyclical process that directed the tuning process and enabled to ultimately evaluate the model's shortcomings.

5.3 Domain Extrapolation

Given the generic nature of the data used — time and geotagged tweets, business location and metadata — there is a multitude of possible applications of this approach, both in similar and alternative domains. One could compare how cities with different economic activities (e.g.: mostly services vs shared service/tourism vs mostly tourism) observe such trends, or compare how different city topologies (e.g.: London vs. a city less reliant on mass public transit) affect these trends. Another interesting study would be to attempt to model the visiting trends in terms of tourist attractions gathered from

several different sources such as TripAdvisor, Lonely Planet and Michelin Travel Guide and evaluate how they compare in suggesting locations versus the actual observed trends.

Reframing to a city planning standpoint, using the same methodology it is possible to assess how specific populations interact with the space and use the resulting data to guide public policy, for example analysing visiting trends of mobility impaired people and use the resulting data to guide public investment/policy towards meeting accessibility standards.

More abstractly, this methodology could be replicated with any type of spatial time series to identify spatio-temporal trends between different populations — classified using available metadata and aggregated from geotagged, timestamped events — and correlate said trends with spatial multivariate data in an attempt to explain said phenomena.

REFERENCES

[1] D. Ren, S. Amershi, B. Lee, J. Suh and J. D. Williams, "Squares: Supporting Interactive Performance Analysis for Multiclass Classifiers," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 61-70, Jan. 2017. doi: 10.1109/TVCG.2016.2598828

[2] Mellon, J., & Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*. https://doi.org/10.1177/2053168017720008

[3] O'Brien, O., 2018. Mapped-based data visualisations, often of London and/or transport/transit.: oobrien/vis. URL https://github.com/oobrien/vis/blob/master/tube/data/zones1to6.json (accessed 12.22.18).

[4] Probst, Philipp & Boulesteix, Anne-Laure & Wright, Marvin. (2018). Hyperparameters and Tuning Strategies for Random Forest.

[5] Documentation - Yelp Fusion API [WWW Document], n.d. URL https://www.yelp.com/developers/documentation/v3 (accessed 12.22.18).

[6] All Category List - Yelp Fusion API [WWW Document], n.d. URL https://www.yelp.com/developers/documentation/v3/all_category_list (accessed 12.22.18).

[7] London Tweet residents vs tourists dataset, provided by City, University of London teaching staff URL http://staff.city.ac.uk/~sbbb717/moodle/2014-2015/inm433/courseworkdata/London_tweet_TS_week_hours_residents_vs_locals.zip