# Using Wi-Fi probe requests from mobile phones to quantify the impact of pedestrian flows on retail turnover

## Abstract

This paper discusses the opportunities afforded by novel population sensing technologies in the field of 'smart' urban management. In particular, it focuses on the application of these new sources of data in retail analysis.

Our goal is to integrate data derived through novel pedestrian counting and point-of-sale systems to build a statistical model that captures the relationship between retail turnover and footfall in the UK. The point-of-sales data are provided by two UK-based food & beverage retailers. To accurately measure the pedestrian activity around retail units, we make use of the data generated by the 'SmartStreetSensor' project: a deployment of a large network of sensors installed across 105 towns and cities in the UK that collect Wi-Fi probe requests generated by mobile devices. We propose and implement novel methods for processing these raw signals into accurate estimates of pedestrian activity without compromising participants' privacy.

The resulting data is then integrated into seasonal ARIMA and dynamic regression models that can be used to predict future sales. Our results indicate that the dynamic regression model that accounts for fluctuations in footfall data outperforms seasonal ARIMA model that uses only past values and behaviours of transaction data to predict future sales. Thus, we conclude that footfall does have a strong impact on retail sales and therefore integrating footfall measures into sales forecasting can significantly improve the forecasting results. We also examine differences between the two retailers and observe a stronger correlation at the Fast Food Retailer locations compared to the correlation at Family Restaurant locations.

**Keywords: human activity patterns, Wi-Fi probe requests, retail location analysis, regression model**

DRAFT

## 1. Introduction

The emergence of people sensing technologies has led to a diverse range of new data sources that are greatly extending the ability to capture and analyse how people move through, and interact with, urban environments. The information, which has traditionally been collected through manual counting and surveys, can now be obtained using novel population sensing technologies that are common in smart cities - such as mobile devices, Wi-Fi sensors or Bluetooth beacons - at a much lower cost and over long periods.

The focus here is to demonstrate the potential of applying movement data in retail analytics – an area of research of increasing national importance to the UK. The value of accurate footfall measures in site selection process is well known (Brown, 1993; Wood & Browne, 2007) since they can offer a basis for predicting store revenues and performance (Waddington et al., 2019). Beyond where to locate a store, developing an understanding of the activity-patterns in an area allows retailers to make informed decisions around optimal trading times (Parker et al., 2017), efficient staffing schedules (Begley et al., 2018; Chapados et al., 2014; Chuang et al., 2016) and can uncover early warning of changes that can negatively impact trading success (Wehrle, 2017). Beyond the specifics of individual retailers, such measures can provide the basis for intelligence-led planning decisions that seek to mediate the impacts of online retail on physical retail spaces and, in the UK context at least, inform the significant government incentives for traditional retailing environments to diversify into other areas (Ministry of Housing Communities & Local Government, 2019).

Despite the potential of better footfall metrics, there remains a relative lack of data-driven studies to provide robust empirical evidence about the relationship between granular footfall measures and retail turnover. Our goal, therefore, is to use the most granular data available to build a statistical model that represents the relationship between retail turnover and footfall.

This research benefits from access to store-level transactions data for 34 retail units split across a Fast Food Retailer (11 retail units) and a Family Restaurant (23 retail units). All units are located in the UK and occupy a diverse range of urban retail centres. The commercial sensitivity of the data means that the retailers supplying the data have chosen to remain anonymous. Having two food & beverage retailers in the sample allows us to make comparisons but also to draw conclusions on the impact of passing footfall on this retail type.

To accurately capture the activity patterns around retail units, we utilise the 'SmartStreeetSensor' project that deployed a network of sensors at storefronts to capture the Wi-Fi probe requests from passing mobile devices. These probe requests are then used to estimate the levels of footfall at any given time. Our priority here is to develop a scalable, nonintrusive and passive collection method without compromising participants' privacy.

We ask the following research questions:

***RQ1:*** Does integrating footfall data to sales forecasting models improve the model's performance?

***RQ2:*** Is there a significant difference between the impact that footfall has on a Family Restaurant compared to the Fast Food Retailer?

This paper remainder of the paper is structured as follows. First, we discuss technologies that are used to measure pedestrian flows and outline the main advantages of estimating the footfall from Wi-Fi probe requests. Then we discuss the opportunities afforded by 'smart' technologies for the retail sector and for furthering existing research in this area. Next, we describe the processes of setting up the footfall sensors and estimating pedestrian flows using Wi-Fi probe requests before describing seasonal ARIMA and dynamic regression models that are common forecasting techniques used in retail to estimate turnover. In the results section, we provide a visual analysis of the collected footfall data and compare the performance of the two forecasting methods. We conclude with the further discussion about potential practical applications and further research objectives in the final chapter.

## 2. Literature Review

### 2.1. People sensing technologies

Much of the urban planning literature has revolved around estimating the collective movement of people through the cities in order to estimate demands on infrastructure (Hancke et al., 2012). Traditionally, this data has been gathered by manual traffic counting, photoelectric sensors, surveys, and videotaping; however, the emergence of novel information and communication technologies has greatly extended our ability to capture the data pertaining to human activities. Nowadays innovative technologies enable dynamic and continuous data collection and applications. Akhter et al. (2019) offer a summary of the most extensively used methods in human counting, such as video and thermal cameras and passive infrared (PIR) sensors radio. A further methodology used to track human trajectories is Radio Frequency

DRAFT

Identification (RFID) technology where tags carrying a unique identifier attached to an object (e.g. shopping carts as in Hui et al. (2009); Kholod et al. (2010) or conference badge as in Cattuto et al. (2010)) transmit signals captured by a system of pre-installed readers. While those monitoring techniques provide means for reducing expensive manual surveys (Bai et al., 2017), they still suffer from an inability to accurately identify distinct individuals (PIR), require bespoke infrastructure (RFID), are prone to measurement errors in outdoor environments (thermal cameras) or violate the privacy of the pedestrians (video cameras).

The advent of data from devices and services routinely carried by individuals has created viable alternatives for collecting data on human activity patterns with greater granularity and across large areas (D'Silva et al., 2017). Mobile devices are equipped with sensors (e.g. accelerometer and compass) and capabilities (e.g. Cellular radio, Bluetooth, Wi-Fi, GPS) that can be used for distributed urban sensing. The first set of research that utilised human mobility data derived through mobile devices used the cellular data from call detail records (CDRs) (e.g. Reades et al. (2007), Becker et al. (2011)), but as this data is collected and stored primarily by a small number of big telecommunications firms, the access to this data source for research purposes is limited.

DRAFT

GPS is another popular technology used to capture data on human mobility at large scales. Two primary sources of GPS data are GPS loggers carried by volunteers and GPS-enabled mobile applications installed in smartphones (Li et al., 2018). GPS data enables research on ambient population and mobility patterns in urban environments (Deville et al., 2014; Sila-Nowicka et al., 2015). GPS tracking data is also a popular source of data in tourism research (Li et al., 2018) where it has been used to find out how tourists move around a city (Edwards & Griffin, 2013) and to predict the next destination of individual tourists (Zheng et al., 2017). However, since GPS data is collected at the device level, it requires user permission to be accessed (Soundararaj et al., 2019a), radically reducing the sample size. Furthermore, GPS does not perform satisfactorily in indoor areas (Heidari & Pahlavan, 2008).

In the past decade, Wi-Fi has emerged as one of the most used technologies in providing high-speed internet access to mobile devices such as smartphones, tablets and laptops in public and private spaces (Torrens, 2008). This has resulted in multiple Wi-Fi networks being available at almost every location in dense urban environments. Traversing through this overlapping mesh of Wi-Fi networks, modern mobile devices with Wi-Fi network interfaces regularly broadcast a special type of signal known as a 'probe request' in order to discover the Wi-Fi networks available to them. This helps these devices to connect and switch between the Wi-Fi networks seamlessly. Probe requests are captured by Wi-Fi networks regardless of whether the device connects to a specific network (Johnson et al., 2019) making it a non-intrusive and passive data collection method, thus improving the participation rate. In the early studies, Wi-Fi signals were mainly used to study mobility at hyperlocal scales such as university campuses (Henderson et al., 2004; Sevtsuk et al., 2008), at event venues (Bonne et al., 2013) and in public transportation terminals (Shlayan et al., 2016), but as argued by Kontokosta & Johnson (2017), with enough infrastructure to collect the Wi-Fi probe requests, we can even aim to generate a real-time census of the city. Data derived through Wi-Fi-networks have also been used in predictive analytics to estimate user destinations based on the locations they have visited in the past (Danalet et al., 2014).

A media access control address (MAC address) assigned by the manufacturer to the mobile devices, when hashed, can act as a unique identifier without compromising participants' privacy. This has enabled a set of research looking at individual travel patterns (Rekimoto et al., 2007; Sapiezynski et al., 2015) and links between location (Phan et al., 2005). User trajectories have been used to create origin-destination matrices of customer journeys that enable a detailed analysis of passenger demand (Ji et al., 2017; Transport for London, 2017) and replace the need for manual counting (as in Ceder(1984)).

However, this data collection method is not without pitfalls. Worries have been expressed about potential misuse and threats to the device owner's privacy. In terms of regulation, legislation such as Europe's General Data Protection Regulation (GDPR) and some vendors have introduced randomisation of MAC addresses of their customers' devices (Vanhoef et al., 2016).

## 2.2. Applications of people sensing technologies in retail sector

The emergence of people sensing technologies has led to a diverse range of new data sources that provide objective measures on people's movement (Cukier & Mayer-Schönberger, 2015) for informed decision making. The competitive advantage of successful exploitation of new technologies (McAfee & Brynjolfsson, 2012) has also been recognised in retail sector.

Novel examples of the use of innovative 'smart' technologies include the use of Bluetooth beacons (Betzing, 2018) to monitor consumer in-store journeys and location-based marketing notifications that are delivered to consumers' mobile devices (Banerjee & Dholakia, 2008; Van De Sanden et al., 2019). Recent academic studies that have made use of new sources of data of people's movement include the applications of GPS traces to study consumer behaviour (Sila-Nowicka & Fotheringham, 2016) and walking patterns in retail areas (Hahm et al., 2017). In addition, several studies have looked at the importance of urban morphology and street networks on retail prosperity in urban spaces (Kang, 2016; Sevtsuk, 2014) and found that micro-location characteristics and retail composition in the area are important to explaining the retail landscape. Arunraj et al. (2016), Appelqvist et al. (2016) and Badorf & Hoberg (2020) studied the impact of weather on retail sales and found that the magnitude of the weather effect is not uniform and depend on the store location and the sales theme.

DRAFT

The benefits of accurate measures of footfall has been widely discussed in the context of retail location analysis. These data provide retailers robust evidence in site selection processes for assessing potential revenue and performance of a new venue (Brown, 1993; Waddington et al., 2019). As Wood & Browne (2007) and Berry et al. (2016) have highlighted, prior understanding of the fluctuations in footfall patterns is particularly important for smaller comparison goods retailers in urban areas, who are unlikely to have much influence on traffic volume and are therefore dependent on existing pedestrian flows.

Assessing footfall patterns around existing retail units helps retailers to make informed decisions around store operation (Fan, 2019). For example, a number of research studies have demonstrated the benefits of traffic-based scheduling to optimise staffing costs (Begley et al., 2018; Chapados et al., 2014; Chuang et al., 2016) and a recent report on high street vitality (Parker et al., 2017) emphasised the importance of matching the store trading hours with the human activity-patterns in the area as one of the key priorities in improving the store performance.

In addition, continuous and up-to-date footfall data provides robust empirical evidence for uncovering early warnings of changes that can negatively impact trading success (Wehrle, 2017). Adapting to the changes in the retailing environment has proven to be of critical importance over the last decade. The retail sector has been fundamentally changed by the growth of online retail that now takes up 21.3% (Office for National Statistics, 2018) of all the retail sales and the fall in market share has created major issues for traditional store-based retailers. Well-established retailers (e.g. Marks & Spencer) have had to downsize their store networks while others (e.g. Debenhams, Mothercare, Jamie's Italian, Patisserie Valerie) have gone into administration (Centre for Retail Research, 2019). Vacancy rates on British high streets are 10.3% (BBC, 2019) marking the highest level since January 2015 and in some regions the footfall has dropped by 17.9% (ITV, 2020) over the last decade. Further discussion of the challenges in high street retailing is out of scope in this study but has been thoroughly studied in the papers by (Grimsey et al., 2018; Parker et al., 2017; Portas, 2011).

Despite the benefits research has attributed to the use of movement data in retail analytics, there is a lack of data-driven studies that have provided robust empirical evidence about the impact of footfall on retail turnover. Previous attempts (Graham et al., 2019; Matzler et al., 2010) to analyse the relationship between pedestrian flows and retail turnover, have had to rely on manual counts, modelled data and consumer interviews, because the academic research in this field is often restricted by limited data access due to the perceived commercial value and also potentially sensitive nature of the data. Private sector companies have reported a close correlation between spend and footfall (Ipsos Retail Performance, 2018; Springboard, 2020; The Local Data Company, 2020), but they haven’t made empirical evidence public.

### 2.3. Dynamic Regression Model

DRAFT

Footfall and retail sales typically contain both daily trends and seasonal patterns, presenting challenges in developing effective regression models. Over the last few decades several approaches such as Monte Carlo method (Nelson & Schwert, 1982), K-nearest neighbour algorithms (Habtemichael & Cetin, 2016) and artificial neural networks have been studied to address these components (Ramos et al., 2015). More recently, machine learning-based techniques such as tree-based methods, Support Vector Regression (Smolak et al., 2020) and Random Forests and deep-learning based algorithms such as Recurrent Neural Network and Long Short-Term Memory have gained traction. In this study we select a dynamic regression model (Hyndman & Athanasopoulos, 2018) approach since it is most commonly used in short-term forecasting and valued for its accuracy (Smolak et al., 2020). Dynamic regression applies an Autoregressive Integrated Moving Average (ARIMA) process (Box & Jenkins, 1970) to model both trend and seasonal patterns and then fits a linear regression model to calculate the dependency between variables. We compare the predictive power of dynamic regression model against a seasonal ARIMA (SARIMA) model where the variable of interest is forecasted using only its past values. Similar comparisons between SARIMA and dynamic regression model performance have been conducted in previous studies by Arunraj et al. (2016) and Elamin & Fukushige (2018), however, to best of our knowledge, this is the first attempt to improve retail sales forecasting performance by integrating footfall counts to the forecasting model.

## 2. Data & Methodology

### 3.1. Data

This research employs two datasets: Wi-Fi probe requests and retailer transactions. The former is used to generate estimates of pedestrian activity at selected locations and the latter is aggregated to calculate total sales volumes at the same locations during the corresponding times. This section describes these data sources in detail along with the methods and techniques used to clean, process and link them before using them to conduct a comparative analysis.

DRAFT

### 3.1.1. Footfall Data

The ‘Smart Street Sensor’ project is a collaboration between the Economic and Social Research Council (ESRC) Consumer Data Research Centre (CDRC) and The Local Data Company and aims to produce a national level dataset of footfall in the United Kingdom’s retail areas with unprecedented spatial and temporal granularity. The project deploys a network of sensors installed in the front of retail stores across the UK.

All mobile devices with Wi-Fi capability regularly broadcast special signals called probe requests directed towards all Access Points in the vicinity in order to keep a list of available access points. Using a Wi-Fi transponder, the sensors collect all these probe requests and transfer them to a centralised location. Before being sent to the server these raw probe requests are aggregated by their MAC addresses for every 5 minutes and the MAC address itself is obfuscated using a cryptographic hashing algorithm. The final aggregated information sent for each unique MAC address at 5-minute intervals are listed in Table 1.

Since 2015, the project had a footprint of approximately 1000 locations across 105 towns and cities across the UK. In addition to these sensors, the project also collected manual counts of pedestrians at each location for 15-minute interval when these sensors were installed. This 15-minute manual counting was collected to allow for validation as well as calibration (detailed in Section 3.1.1.2).

| Field | Description |
|---|---|
| Packets | Total number of packets collected interval. |
| VendorPart | The first part of the MAC address showing the manufacturer of the hardware. |
| MacAddress | The second part of the MAC address that is transformed in to a cryptographic hash. |
| Signal | The minimum signal strength reported among the packets for the unique MAC address. |
| PacketType | Code corresponding to the type of the packet captured. Since only management type packets are collected, this is always ’1’ |

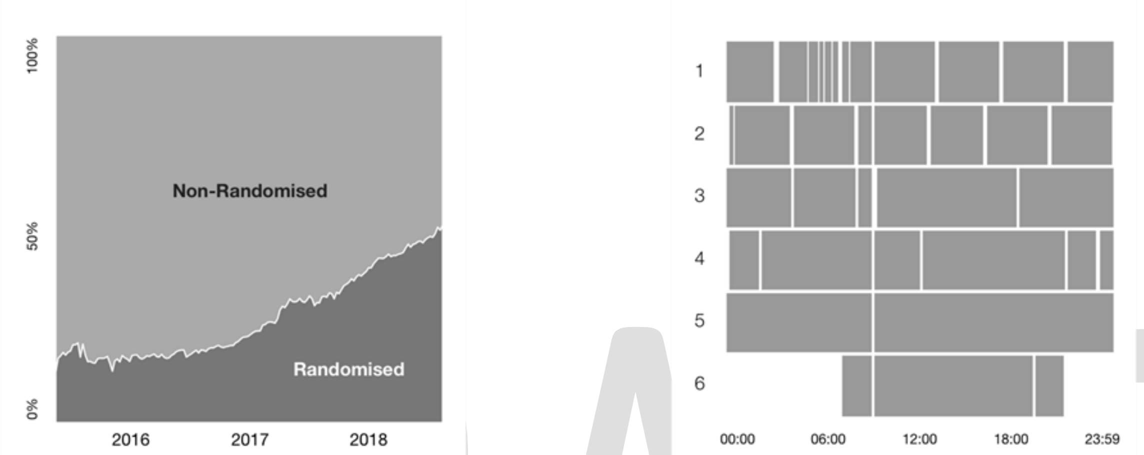
**Table 1:** The aggregated information sent by Smart Street Sensor on probe requests with Unique MAC addresses every 5 minutes.

#### *3.1.1.1. Cleaning the Data*

The first source of uncertainty arises from devices that stay around the sensors for extended periods, thus generating multiple probe requests over multiple intervals. Though this can be solved by aggregating them based on unique MAC addresses, it is exacerbated by MAC address randomisation, which started with the introduction of iOS8 in 2015 but has been increasing steadily and reached a critical point when iOS 10 implemented a more aggressive randomisation technique. The impacts of this are shown in Figure 1a.

In addition, the dataset suffers from missing data both sporadically over short time periods (=less than 30 minutes) as well as for longer durations. An example of this is shown in Figure 1b where the six sensors across a single street - Tottenham Court Road (London) - show missing data across a day. The list of causes includes connectivity failures, the reboot cycle of the sensors and accidental unplugging by store staff. Gaps of longer duration (=longer than 30 minutes) tend to result from hardware failures and also the opening times of retail establishments that cut their power when closed. Errors can also arise from the uncertain field of measurement – that is the size of the area that each sensor can detect probe requests within - which makes it challenging to convert the sensor-based counts to a particular corridor of pedestrian footfall.

It is also noted that the dataset suffers from systemic biases due to varying mobile phone ownership across locations and across time. For example, the ownership of mobile devices with Wi-Fi capability has increased steadily over the past decade leading to a steady inflation of the number of probe requests collected.



(a) The proportion of randomised vs non randomised from 2016-18 showing increase in randomised MAC addresses in Smart Street Sensor Data.

(b)The coverage of data over a day from 6 sample sensors at Tottenham Court Road, London showing short-term and long-term missing data.

**Figure 1:** Uncertainties in the data

For non-randomising devices it is straightforward to account for devices that dwell for long periods within reach of the sensor as we can simply remove all the packets that have MAC addresses repeating in any rolling-window of 30 minutes. This is possible since we have ensured that the uniqueness of the hashes are preserved within a one-week period using a weekly rotation of random salt value. But the above methods don't work with devices that randomise MAC addresses and causes massive over-counting. Since 2015 there have been multiple attempts at bypassing the randomisation to derive unique device identification using. Most of these utilise techniques such as manufacturer profiling (Martin et al., 2016), scrambler attack (Bloessl et al., 2015), timing attacks (Matte et al., 2016) or using information elements (Vanhoef et al., 2016). Though effective, these techniques often require intrusive collection of data, thus risking the privacy of users being surveyed and are therefore discounted here. An alternative method for solving this problem using the sequence numbers has been explored by Soundararaj et al. (2019a) but was found too computationally expensive for the volume of data used here. Instead, a simpler approach was implemented that utilised the ratio of the number of probe requests generated by the devices that don't randomise their MAC address against those that do to calculate a "compression" factor for each five-minute interval at every location and use it to adjust the randomised probe requests. Assuming that, on average, both randomising and non-randomising devices emit similar number of probe requests in a given time interval at a certain location, we can estimate the number of randomising devices (Nr) for a given interval from the number of non-randomising devices (Nnr) and the number of probes requests generated by both randomised (Pr) and non-randomised (Pnr) devices as explained in equation 1,

DRAFT

$$N_r = \frac{N_{nr}}{P_{nr}} \times P_r \quad (1)$$

The result of such a simple cleaning method proves effective especially against the changes in the software of the mobile devices over the long term. The results for this adjustment for a particular sensor in Cardiff is shown in Figure 2.

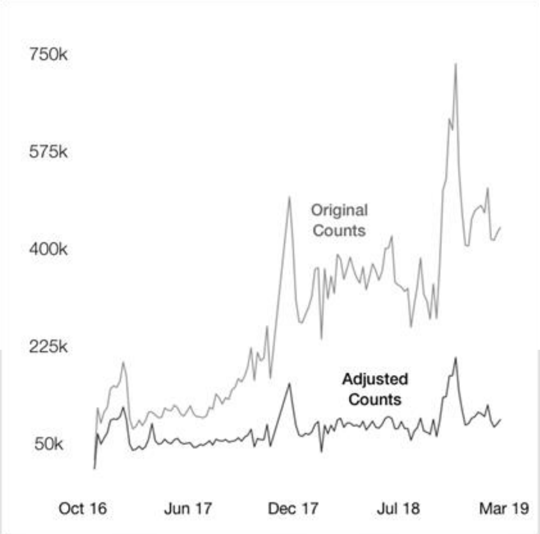


**Figure 2:** Results comparing the weekly counts of the number of devices at a chosen location before and after adjusting the number of devices with randomised MAC addresses

We can observe that the adjusted device number estimates still preserve the seasonal variations while avoiding the huge increase of probe requests caused by the changes in method of randomisation around 2017.

The missing data are filled by imputing the values from the historic data at the locations using imputeTS (Moritz & Bartz-Beielstein, 2017). The gaps shorter than 15 minutes are imputed using a straightforward spline-based method from the data preceding and following them. The longer gaps are filled in using seasonally decomposed missing value imputation while treating the data as time series data with seasonal variations at appropriate scales. For example, the hourly gaps are filled by assuming that the time series varies seasonally every 24 hours and the daily gaps are filled by assuming that the time series varies with seasonality of every 7 days.

Finally, these estimates of the number of mobile devices at each location are converted into pedestrian footfall estimates by using the “adjustment factor” - a simple ratio derived for each location by comparing the manual counts conducted at each location to the counts reported by the sensor at the corresponding times. Calibrating with ground truth was necessary since the proportions of mobile device ownership amongst the passing population was an external uncertainty to our study and could arise from a variety of spatio-temporal and demographic factors. This calibration can be carried out periodically to improve the quality of the estimation. In addition to this the overall, long-term inflation of number of devices due to mobile ownership has been adjusted assuming an underlying 0.2% weekly increase caused by the increase in smartphone penetration across UK population (Deloitte, 2018) resulting in a more continuous and reasonable estimate of number of devices present at these locations.

DRAFT

The overall data processing pipeline is shown in Figure 3 which starts with the central data repository which contains the raw data from the sensors to 5-minute aggregated footfall estimates. The pipeline was built to suit the scale, size and complexity of this particular dataset using standard Unix tools and parallelised whenever possible (Soundararaj et al., 2019b).

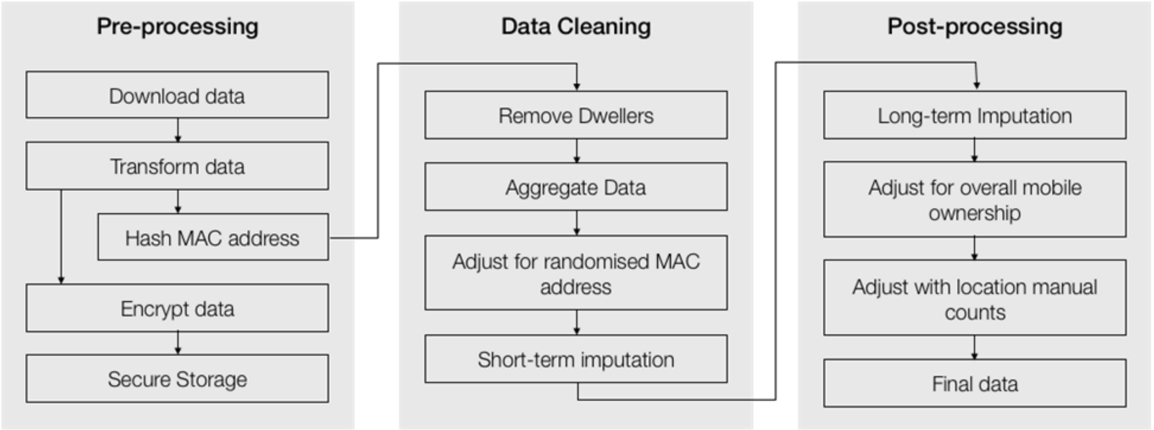


**Figure 3:** The complete data processing pipeline that takes the raw probe requests from Smart Street Sensor. The pre-processing part of the pipeline mainly concerns with producing a safe version of the raw data by removing the personally identifiable information present in them. The data cleaning involves all the methods discussed above to produce an estimate of the number of devices present around a given sensor. The post-processing is concerned with converting the device numbers into other estimates pertaining to their use which is pedestrian footfall in this case.

### 3.1.2. Transactions Data

This research benefits from also having access to store level transactions data. The data is provided under the CDRC Data Sharing Agreement and is hosted in a secure environment (Consumer Data Research Centre, 2020). Data is provided under conditions of nondisclosure and anonymity and therefore retailers are referred to using the broad categorisation of their retail type. It pertains to 11 Fast Food Retailer stores, and 23 Family Restaurant retail units. The transactions data covers the year 2017 (01 January 2017 – 31 January 2017) and is aggregated to daily transaction volumes representing the total number of transactions made at the retail unit in a day. Each retail unit included in the transactions dataset is equipped with a footfall sensor allowing an integrated analysis of footfall flows and retail turnover. Spatial distribution of the sample is shown in Figure 4.

DRAFT

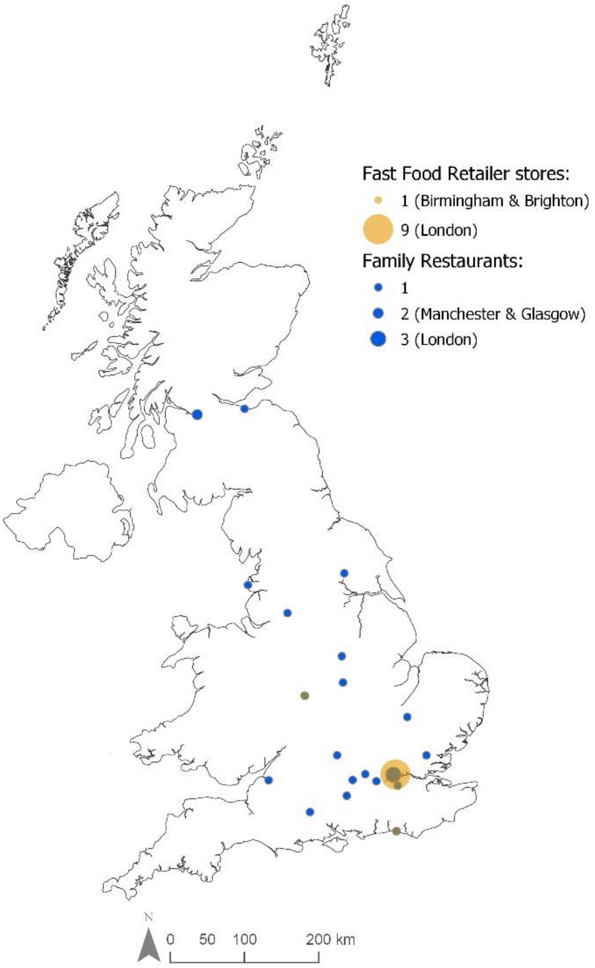


**Figure 4:** Spatial distribution of the available data. Out of the 34 locations in the sample, 12 (9 Fast Food Retailer, 3 Family Restaurant) are in London and 2 in Brighton (1 Fast Food Retailer, 1 Family Restaurant) as well as in Birmingham (1 Fast Food Retailer, 1 Family Restaurant). Remaining 18 locations are distributed across the country.

### 3.1.3. Linking transactions data to footfall data

Although, transactions data is available for the whole year 2017, the availability of the footfall data varies across the locations as the footfall sensors were installed gradually throughout the year 2017 and the data is prone to missing values (discussed in Section 3.1.1.1.).

Therefore, we extract the longest consecutive period without missing values in year 2017 for each sensor and link the aggregated daily total footfall counts to the transactions data using date and sensor number as common denominators. The temporal availability of the sample data is visualised in the Appendix A.

## 3.2. Methodology

To understand if footfall has an impact on retail turnover we compare the performance of two time-series modelling approaches - univariate (= only sales) seasonal Autoregressive Integrated Moving Average model (SARIMA) and seasonal Autoregressive Integrated Moving Average with Explanatory Variable (SARIMAX) (Hyndman & Athanasopoulos, 2018), also referred to as dynamic regression (Nagy & Simon, 2018; Pankratz, 1991). The former is an autoregressive model, meaning the variable of interest is modelled using linear combination

DRAFT

of past values of the variable. The latter adds an external variable (= footfall) into the model using linear regression and then models the data using SARIMA model.

A SARIMA model is notated as follows:

$$\text{ARIMA}\quad \underbrace{(p,d,q)}_{\substack{\uparrow\\ \left(\begin{array}{l}\text{Non-seasonal part}\\ \text{of the model}\end{array}\right)}}\ \underbrace{(P,D,Q)_m}_{\substack{\uparrow\\ \left(\begin{array}{l}\text{Seasonal part}\\ \text{of the model}\end{array}\right)}}$$

where:

p = non-seasonal autoregressive (AR) order
d = non-seasonal differencing
q = non-seasonal moving averages (MA) order

P = seasonal AR order
D = seasonal differencing
Q = seasonal MA order
m = number of periods per season

The first step in the SARIMA modelling is identifying the parameters for the above described components. For the purpose of cross-validation, the time series datasets are classified into training data and testing data. The testing data includes 14 last observations (=2 week) of each time-series. The ratio of testing data relative to the training data depends on the availability of the data at a specific retail unit (s. Figure in Appendix A) and varies between 4.7% and 15.4% of the total length of available data.

The parameters for the models are defined using only training data. We define the parameters based on autocorrelation (ACF) and partial autocorrelation functions (PACF) plots (Figure 5). The seasonal part is defined as follows: because the ACF plot (Figure 5) has significant spikes at lag 7 and at further lags which are multiples of 7, we set the seasonal period *m* to 7. Since the seasonal pattern is stable over time we set D = 1. For the non-seasonal part, we set the d to 1, which indicates the order of differencing. Differencing is required when the time series explicit a trend and is therefore not stationary. Rest of the parameters are determined through trial and error and examining the significant lags at the ACF and PACF plots. We find that the set of parameters that yield in average the best forecasting results across all 34 retail locations are SARIMA(1,1,0)(0,1,1)$_{[7].}$ To confirm that parameters are suitable, we conduct plot Ljung-Box test for each time series to check that the residuals have no remaining autocorrelations. We acknowledge that the selected set of parameters might not be the most optimal for each time series studied in this research, but our aim is to compare the performance of two models under the same conditions and we are not looking to maximise the forecasting performance. Furthermore, although Fast Food Retailer and Family Restaurant have variation in their sales patterns (e.g. Family Restaurant is busier over the weekend, Fast Food Retailer is busier over the working week), the seasonality at lag 7 and autocorrelation in all data sets are similar.

DRAFT

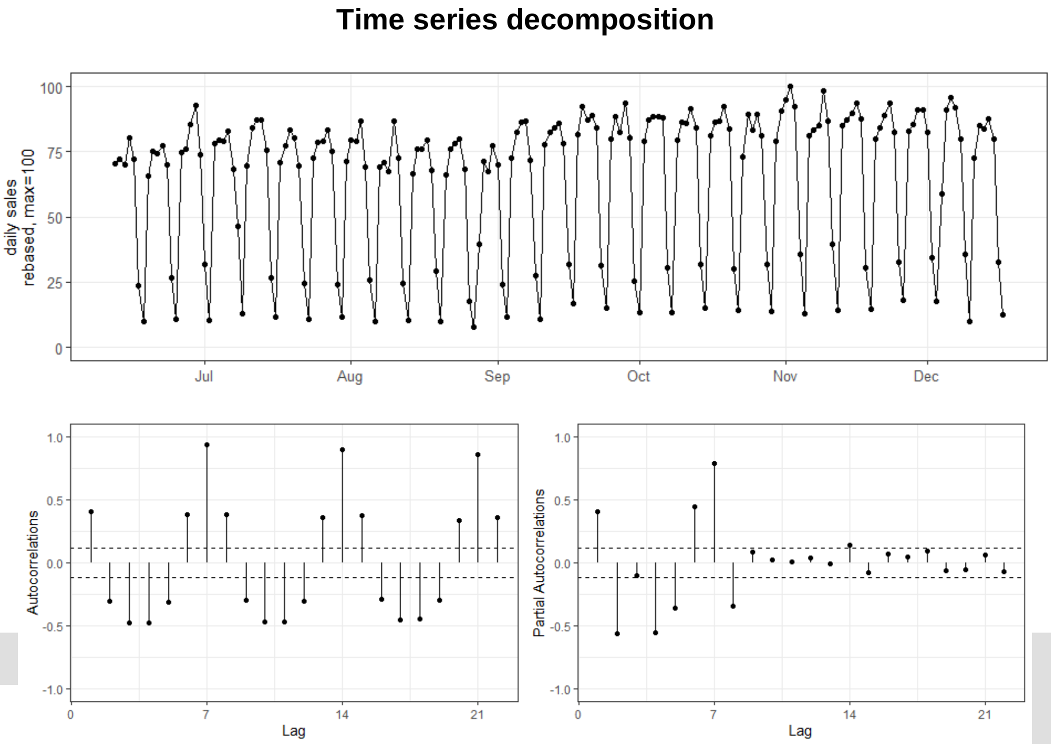


**Figure 5:** Trend and seasonality in the data –
example based Fast Food Retailer sales data from London (Strand).

The model is then used to predict 14 days' worth of observations at each retail location. We use the *forecast* function available in the *forecast* R package developed by Hyndman & Athanasopoulos (2018). To compare the performance of the models, we calculate the average difference between the forecasted values and the observed values (=testing data) expressed as mean absolute percentage error (MAPE). The model with lower MAPE values is considered to be the more accurate. Finally, the Wilcoxon signed-rank test is used to test for statistically significant difference between MAPE measures of the SARIMA and SARIMAX models.

# 4. Results

## 4.1. Exploratory analysis of footfall data

Figure 6 shows the normalised weekly footfall of 10 different locations across Cardiff for the years 2017 and 2018. The patterns in the footfall reveal events that were happening in Cardiff and the unusually high or low footfall in the corresponding weeks. The most significant event was in February 2018, when all sensors reported the lowest numbers they have ever recorded. This coincided with the cold wave in UK nicknamed 'Beast from the East' (Wikipedia, 2018), which brought adverse weather conditions all over the UK and led to a significant reduction in footfall. The other identifiable events are bank holiday weekends which result in higher than normal footfall and FIFA World Cup which took place in the summer of 2018.

DRAFT

DRAFT

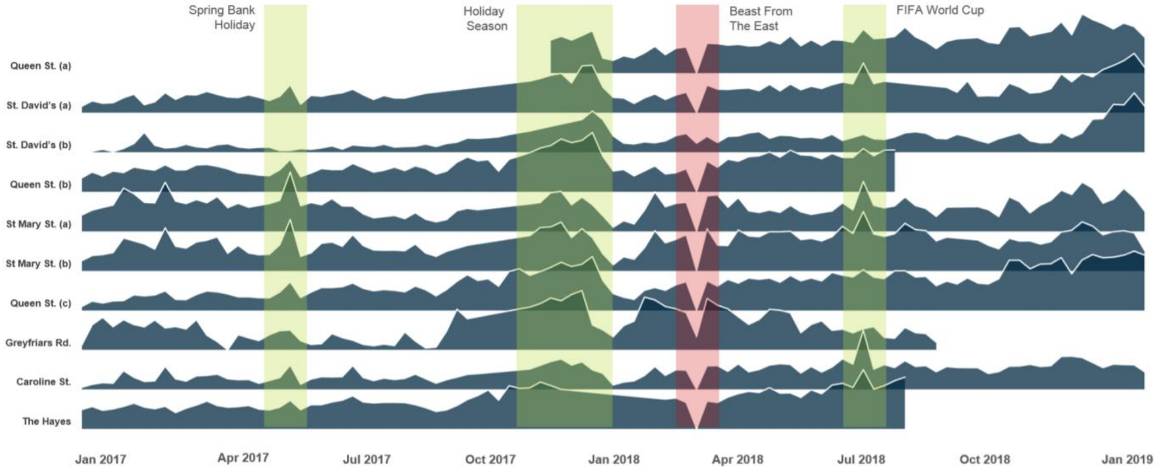


**Figure 6:** Long term footfall profiles at 10 locations in Cardiff in 2017 and 2018. Bank holiday weekends, the festive season and FIFA World Cup increased footfall, whereas the ‘Beast from the East’ cold weather in February 2018 triggered a major decrease.

Footfall patterns can also reveal the function of the place. For example, Figure 7 shows the daily footfall profile of three locations in London for two weeks in 2019. It can be observed that all three locations have completely different patterns of usage. Leicester Square is mostly a night-time destination where the footfall peaks around evening while Regent Street is a mostly office location with three distinct peaks corresponding to morning commute, evening commute and lunch. These insights can be crucial for retailers operating in these places for optimising their business operation in terms of store opening times, scheduling shifts etc.

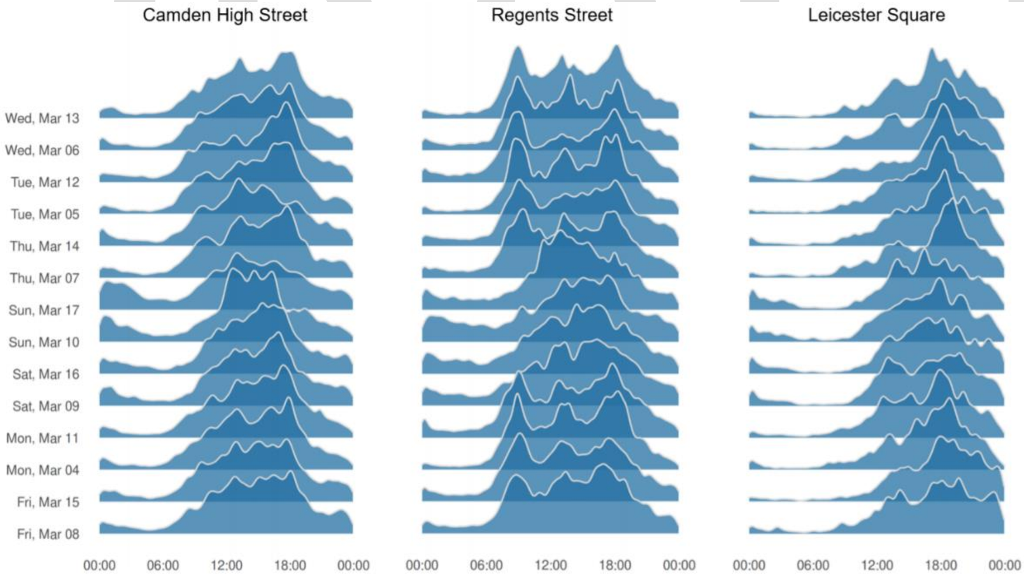


**Figure 7:** Footfall profiles at locations across London demonstrating the difference in their nature. The graph shows hourly profiles from 08th March 2019 to 13th March 2019 across 3 locations in London.

## 4.2. Exploratory analysis of the relationship between footfall and transactions

The dynamic regression model assumes linear relationship between variables. Therefore, to confirm that the relationship between sales and footfall is indeed linear we plot the variables on a scatterplot, fit a regression line between the variables and calculate the correlation. The results for some of the retail units are visualised in Figure 8. The fitted lines have a positive slope, reflecting the positive, linear relationship between sales and footfall. At Fast Food Retailer locations, Sunday and Saturday values are significantly lower than the values from Monday to Friday forming clusters seen on the scatterplots. This confirms the weekly (7 day) seasonality observed on the ACF and PACF plots in section 3.2.

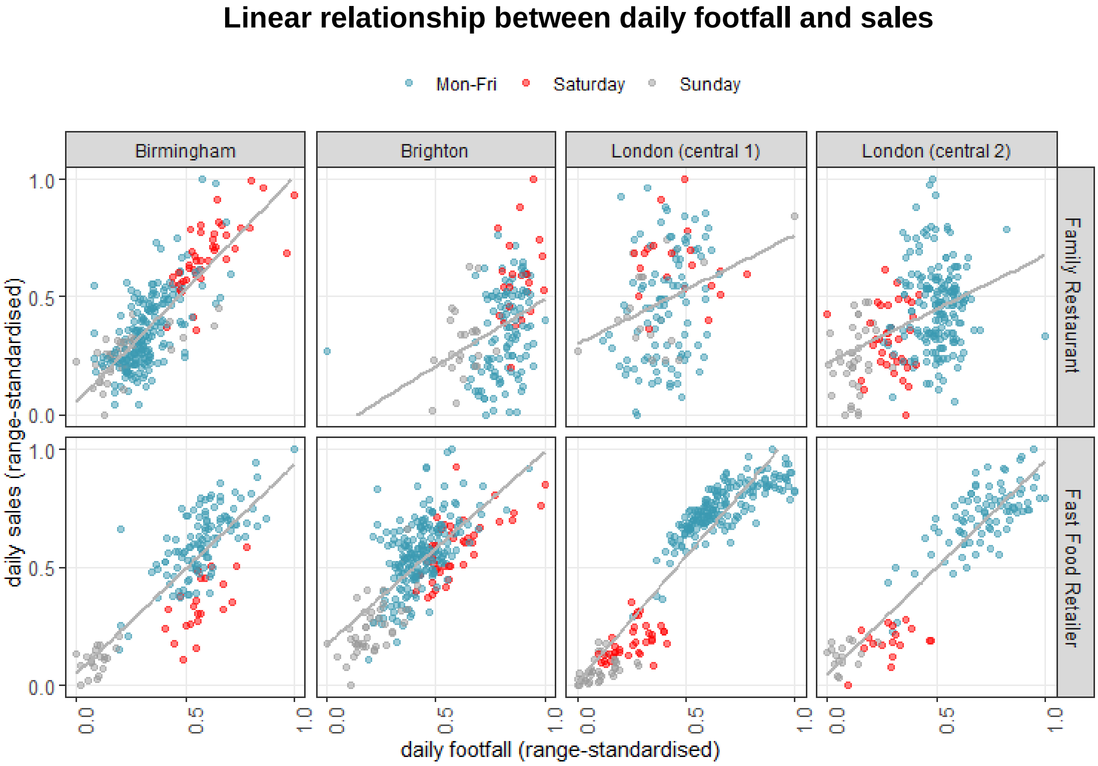


**Figure 8:** The scatterplots show linear relationship at all retail locations, which is stronger at Fast Food Retailers, especially at the retail units located in London.

In order to calculate the correlation coefficients (r), we need to remove the seasonality from the data because seasonal patterns can cause spurious regression outputs (Granger & Newbold, 1974). This is achieved through seasonal adjustment whereby we subtract an observation from the previous observation from the same season (in this case from the same weekday in the prior week). Seasonal-differencing is also applied in forecasting models in Section 4.3.

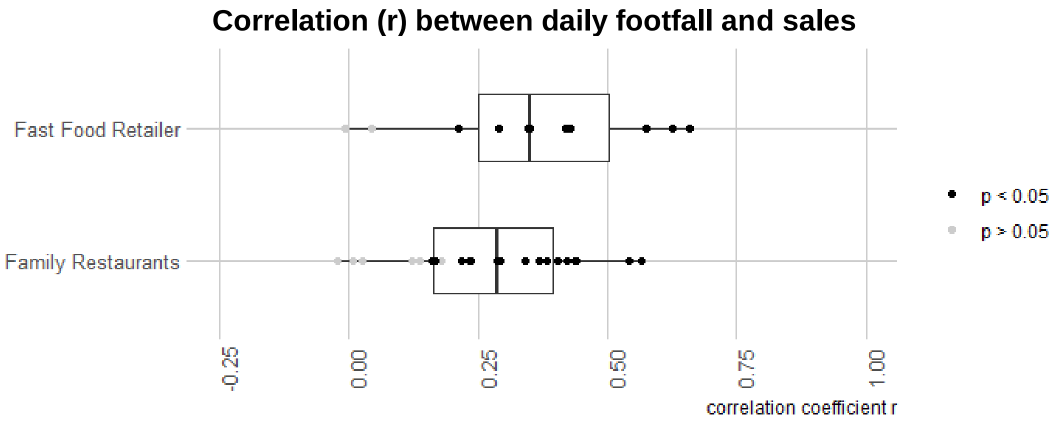
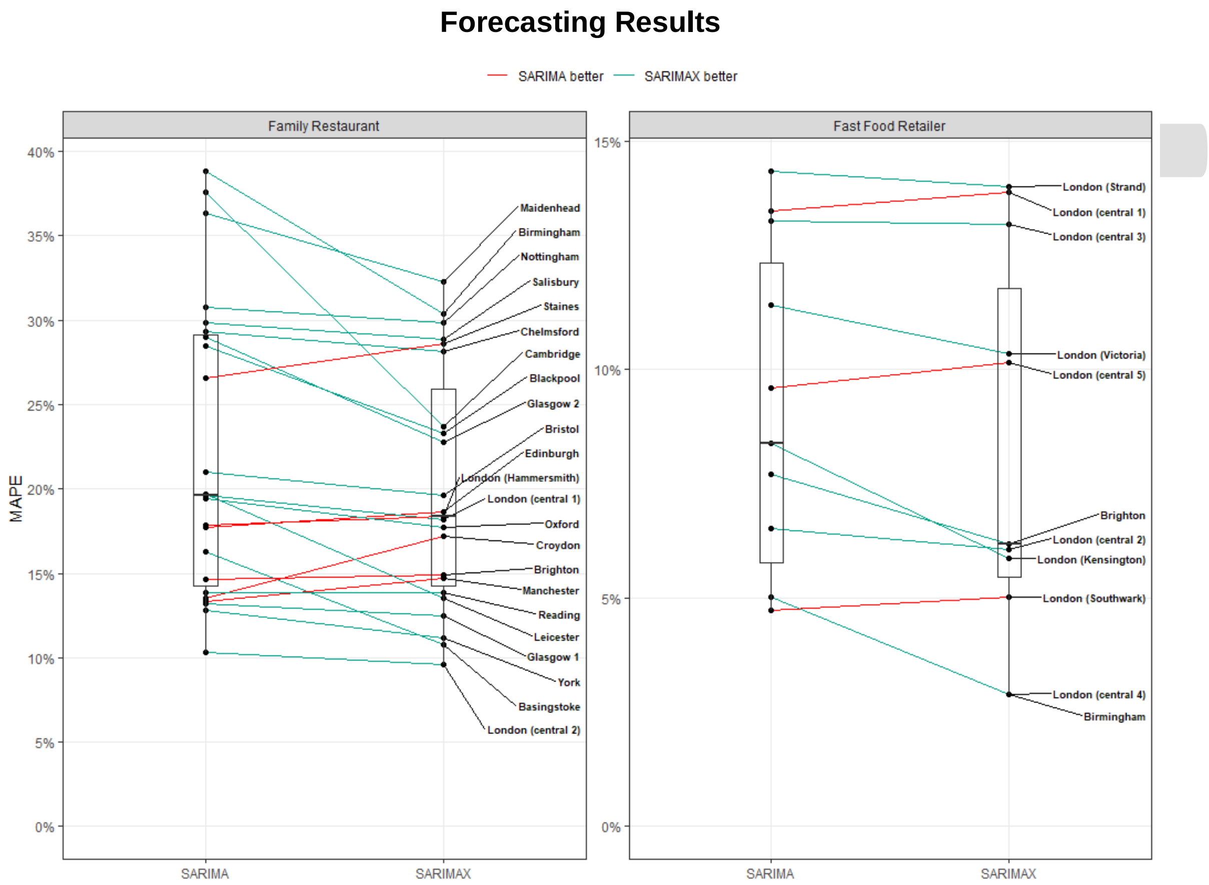


**Figure 9:** Correlation (r) between daily footfall and sales. There is a significant correlation between footfall and sales at 9 (out of 11) Fast Food Retailer locations and the median

correlation is 0.42 (excluding non-significant correlations). The correlation is significant at 17 Family Restaurants and the median correlation is 0.34.

### 4.3. Forecasting results

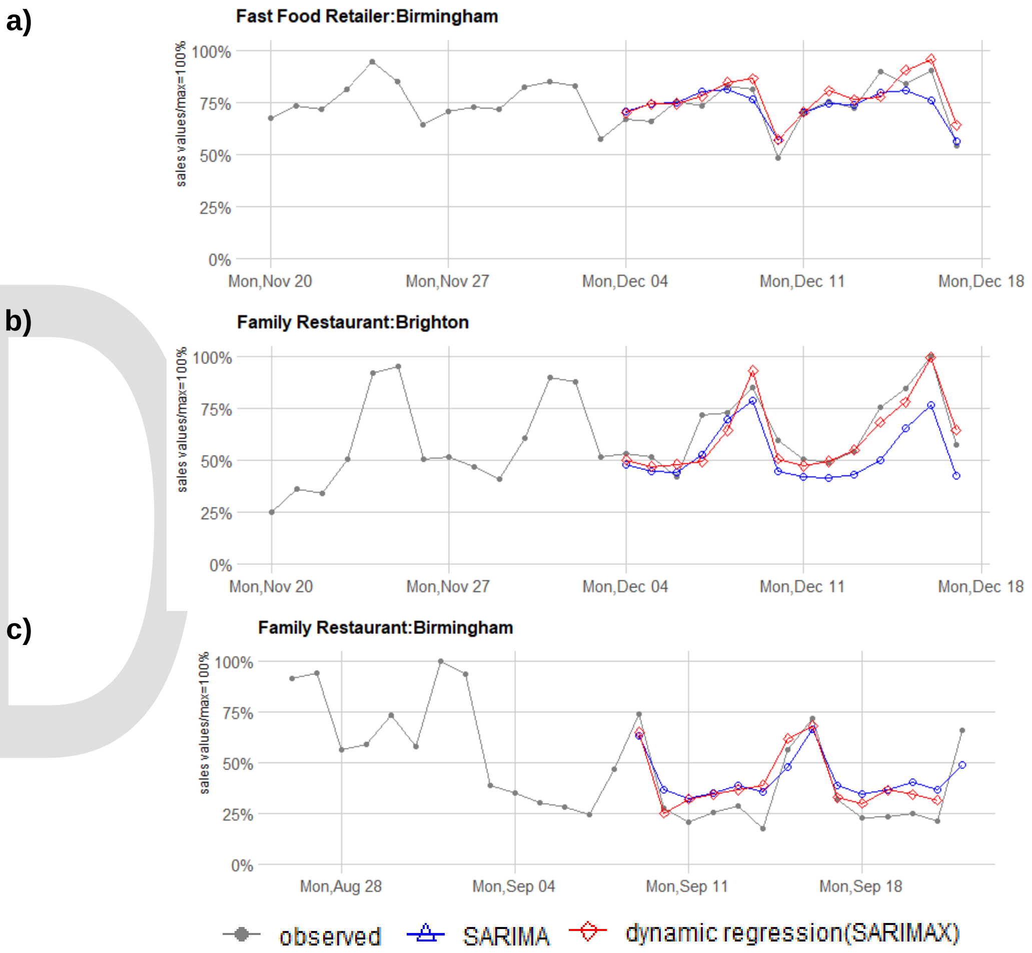
Figure 10 shows the comparison between the two forecasting models. In the case of Family Restaurant, the forecasting was improved in 17 cases out of 23 and MAPE dropped from 19.6% to 18.4%. In the case of Fast Food Retailer, adding footfall data to the model improved forecasting in 8 cases out of 11 and the MAPE dropped from 8.4% to 6.2%. Therefore, we conclude that footfall has an impact on the transactions and integrating footfall counts to sales forecasting models improves the forecasting results. The Wilcoxon signed-rank test confirms that MAPE measures of dynamic regression model (SARIMAX) are significantly lower than the MAPE values of univariate SARIMA model that uses only past sales data to predict future values.



| | Family Restaurant | | Fast Food Retailer | |
|---|---|---|---|---|
| | SARIMA | SARIMAX | SARIMA | SARIMAX |
| Median MAPE: | 19.6% | 18.4% | 8.4% | 6.2% |
| Wilcoxon signed-rank test: | p=0.004573 | | p= 0.009204 | |

**Figure 10:** Lower MAPE values indicate a lower error percentage and therefore a better model. p-value by Wilcoxon signed-rank test is in both cases <0.05 which confirms that MAPE measures by the dynamic regression model were significantly lower than by SARIMA model.

Lewis (1982) has defined MAPE values that are less than 10% as “highly accurate forecasting” and values between 10%-20% as “good forecasting”. Therefore, Fast Food Retailer forecasting results can be considered as “highly accurate” and the 2.2% decrease in MAPE values when adding footfall data to forecasting is a significant improvement. We conclude that footfall has a strong impact on Fast Food Retailer sales performance. In average, the performance of the forecasting model at Family Restaurant locations can be considered “good forecasting”; however, there are 9 locations where the MAPE values even after adding footfall to the forecasting model stay between 20%-50% which is considered as “reasonable forecasting”. The forecasting results could potentially be improved by adding further independent variables (e.g. weather, staffing levels, etc.) to the forecasting model. There are no locations were the MAPE value exceed 50% which would be considered inaccurate forecasting.



D T

**Figure 11:** Observed and predicted values. The observed values in data are shown with a grey line and dots. The blue line and circles show the predicted values by seasonal ARIMA model and the red line and squares show the predicted values estimated by the dynamic regression model.

Figure 11 shows the comparison between observed and predicted values. Example a) shows that dynamic regression predictions are more reactive to sudden changes than results of the SARIMA. Besides forecasting future sales, the difference between predicted and observed values could be used to evaluate the retail location performance in the given time period. A significantly lower observed value compared to predicted value as seen in example c) could be seen as an indicator of poor sales performance.

# 5. Conclusions

We contributed to research on the big data analytics in 'smart cities' by developing a novel technology for estimating levels of pedestrian flows based on Wi-Fi probe requests using a network of sensors installed at the storefronts. We concluded that this method provides a good balance between precision and cost, is scalable and does not compromise the privacy of those involved.

Our main objective in this study was to apply this pedestrian counting data to study the impact that passing footfall has on the retail sales. Our results indicate that footfall has a positive impact on retail turnover in most locations and integrating footfall measures into sales forecasting can significantly improve the forecasting results. However, there are spatial variations (e.g. Family Restaurant units in London are less impacted by the passing footfall than the retail units outside of London) as well as variations between retail types (compared to Family Restaurant, Fast Food Retailer's trade is more impacted by the passing footfall). The prediction models could be used to evaluate potential turnover of at prospective locations where footfall data is available by training the model based on transactions data from a similar location (= similar footfall pattern, similar retailing environment).

DRAFT

Based on our findings, we assume that other micro-site characteristics such as the socio-demographic profile, in some cases are as important as footfall. In the future research, we aim to extend the analysis to further retail types as the data becomes available, add more independent variables to the dynamic regression model and study the forecasting results in spatial context in order to understand other factors besides footfall which might impact the turnover. The most important variables to further include in the research are information about the retail composition, socio-economic variables about the residential population (e.g. Output Area Classification) as well as about the working population (e.g. Workplace Zone Classification). This data is easily accessible but would require a different modelling approach than SARIMA models used in this research. Further variables that would provide interesting insights are the break-down of the sales data by the product type and indication about take away purchases. This information was not provided by the retailers.

We conclude from our results that footfall has a strong impact on retail turnover as suggested in earlier studies (Berry et al., 2016; Graham et al., 2019; Wood & Browne, 2007); however, we expand the previous research by stating that footfall in not equally important for all retail types and at all locations and socio-economic variables should be accounted for as well.

# 6. Declarations of interest

None

# 7. References

Akhter, F., Khadivizand, S., Siddiquei, H. R., Alahi, M. E. E., & Mukhopadhyay, S. (2019). IoT Enabled Intelligent Sensor Node for Smart City: Pedestrian Counting and Ambient Monitoring. *Sensors*, *19*(15), 3374. https://doi.org/10.3390/s19153374

Appelqvist, P., Babongo, F., Chavez-Demoulin, V., Hameri, A. P., & Niemi, T. (2016). Weather and supply chain performance in sport goods distribution. *International Journal of Retail and Distribution Management*. https://doi.org/10.1108/IJRDM-08-2015-0113

Arunraj, N. S., Ahrens, D., & Fernandes, M. (2016). Application of SARIMAX Model to

Forecast Daily Sales in Food Retail Industry. *International Journal of Operations Research and Information Systems*, *7*(2), 1–21. https://doi.org/10.4018/ijoris.2016040101

Badorf, F., & Hoberg, K. (2020). The impact of daily weather on retail sales: An empirical study in brick-and-mortar stores. *Journal of Retailing and Consumer Services*, *52*. https://doi.org/10.1016/j.jretconser.2019.101921

Bai, L., Ireson, N., Mazumdar, S., & Ciravegna, F. (2017). Lessons learned using Wi-Fi and bluetooth as means to monitor public service usage. *UbiComp/ISWC 2017 - Adjunct*, 432–440. https://doi.org/10.1145/3123024.3124417

Banerjee, S. (Sy), & Dholakia, R. R. (2008). Mobile Advertising: Does Location Based Advertising Work? *International Journal of Mobile Marketing*, *March*, 1–23.

BBC. (2019). *High Streets hit as shop vacancy rate worst since 2015*. https://www.bbc.co.uk/news/business-49311298

Becker, R. A., Cáceres, R., Hanson, K., Loh, J. M., Urbanek, S., Varshavsky, A., & Volinsky, C. (2011). A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*. https://doi.org/10.1109/MPRV.2011.44

Begley, S., Fox, R., Lunawat, G., & MacKenzie, I. (2018). *How analytics and digital will drive next-generation retail merchandising*. McKinsey & Company. https://www.mckinsey.com/industries/retail/our-insights/how-analytics-and-digital-will-drive-next-generation-retail-merchandising

Berry, T., Newing, A., Davies, D., & Branch, K. (2016). Using workplace population statistics to understand retail store performance. *International Review of Retail, Distribution and Consumer Research*, *26*(4), 375–395. https://doi.org/10.1080/09593969.2016.1170066

Betzing, J. H. (2018). Beacon-based customer tracking across the high street: Perspectives for location-based smart services in retail. *Americas Conference on Information Systems 2018: Digital Disruption, AMCIS 2018*.

Bloessl, B., Sommer, C., Dressier, F., & Eckhoff, D. (2015). The scrambler attack: A robust physical layer attack on location privacy in vehicular networks. *2015 International Conference on Computing, Networking and Communications, ICNC 2015*. https://doi.org/10.1109/ICCNC.2015.7069376

Bonne, B., Barzan, A., Quax, P., & Lamotte, W. (2013). WiFiPi: Involuntary tracking of visitors at mass events. *2013 IEEE 14th International Symposium on a World of Wireless, Mobile and Multimedia Networks, WoWMoM 2013*. https://doi.org/10.1109/WoWMoM.2013.6583443

Box, G. E. P., & Jenkins, G. M. (1970). Time series analysis: Forecasting and control. *San Francisco: Holden-Day.*

Brown, S. (1993). Retail location theory: evolution and evaluation. In *The International Review of Retail, Distribution and Consumer Research* (pp. 185–229). https://doi.org/10.1080/09593969300000014

Cattuto, C., Van Den Broeck, W., Barrat, A., Colizza, V., Pinton, J.-F., & Vespignani, A. (2010). Dynamics of Person-to-Person Interactions from Distributed RFID Sensor Networks. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0011596

Ceder, A. (1984). Bus frequency determination using passenger count data. *Transportation Research Part A: General*. https://doi.org/10.1016/0191-2607(84)90019-0

Centre for Retail Research, N. (2019). *Who’s Gone Bust in Retail?*

DRAFT

http://www.retailresearch.org/whosegonebust.php

Chapados, N., Joliveau, M., L'Ecuyer, P., & Rousseau, L. M. (2014). Retail store scheduling for profit. *European Journal of Operational Research*, *239*(3), 609–624. https://doi.org/10.1016/j.ejor.2014.05.033

Chuang, H. H. C., Oliva, R., & Perdikaki, O. (2016). Traffic-Based Labor Planning in Retail Stores. *Production and Operations Management*, *25*(1), 96–113. https://doi.org/10.1111/poms.12403

Consumer Data Research Centre. (2020). *CDRC data*. https://www.cdrc.ac.uk/data-services/cdrcdata/

Cukier, K., & Mayer-Schönberger, V. (2015). The Rise of Big Data: How It's Changing the Way We Think about the World. In *The Best Writing on Mathematics 2014*. https://doi.org/10.1515/9781400865307-003

D'Silva, K., Noulas, A., Musolesi, M., Mascolo, C., & Sklar, M. (2017). If I build it, will they come? Predicting new venue visitation patterns through mobility data. *SIGSPATIAL'17, November 7-10 2017, Los Angeles Area, CA, USA*. https://doi.org/10.1145/3139958.3140035

Danalet, A., Farooq, B., & Bierlaire, M. (2014). A Bayesian approach to detect pedestrian destination-sequences from WiFi signatures. *Transportation Research Part C: Emerging Technologies*. https://doi.org/10.1016/j.trc.2014.03.015

Deloitte. (2018). *Mobile Consumer Survey 2018*. https://www2.deloitte.com/uk/en/pages/technology-media-and-telecommunications/articles/mobile-consumer-survey.html

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D., & Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1408439111

Edwards, D., & Griffin, T. (2013). Understanding tourists' spatial behaviour: GPS tracking as an aid to sustainable destination management. *Journal of Sustainable Tourism*. https://doi.org/10.1080/09669582.2013.776063

Elamin, N., & Fukushige, M. (2018). Modeling and forecasting hourly electricity demand by SARIMAX with interactions. *Energy*, *165*, 257–268. https://doi.org/10.1016/j.energy.2018.09.157

Fan, D. (2019). *Retail Revenue Prediction Models with Spatial Data Science*. https://carto.com/blog/retail-revenue-prediction-data-science/

Graham, C., Khan, K., & Ilyas, M. (2019). Estimating the value of passing trade from pedestrian density. *Journal of Retailing and Consumer Services*, *46*, 103–111. https://doi.org/10.1016/j.jretconser.2017.10.005

Granger, C. W. J., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*. https://doi.org/10.1016/0304-4076(74)90034-7

Grimsey, B., Hopkinson, M., Hood, N., Pascoe, E., Shellard, C., Sadek, J., Cassidy, K., Dehullu, V., & Baker, M. (2018). *The Grimsey Review 2*. http://www.vanishinghighstreet.com/

Habtemichael, F. G., & Cetin, M. (2016). Short-term traffic flow rate forecasting based on identifying similar traffic patterns. *Transportation Research Part C: Emerging Technologies*. https://doi.org/10.1016/j.trc.2015.08.017

DRAFT

Hahm, Y., Yoon, H., Jung, D., & Kwon, H. (2017). Do built environments affect pedestrians' choices of walking routes in retail districts? A study with GPS experiments in Hongdae retail district in Seoul, South Korea. *Habitat International*, *70*, 50–60. https://doi.org/10.1016/j.habitatint.2017.10.002

Hancke, G., Silva, B., & Hancke, Jr., G. (2012). The Role of Advanced Sensing in Smart Cities. *Sensors*, *13*(1), 393–425. https://doi.org/10.3390/s130100393

Heidari, M., & Pahlavan, K. (2008). Performance evaluation of wifi RFID localization technologies. In *RFID Technology and Applications*. https://doi.org/10.1017/CBO9780511541155.007

Henderson, T., Kotz, D., & Abyzov, I. (2004). The Changing Usage of a Mature Campus-wide Wireless Network. *MobiCom'04, Sept. 26-Oct. 1, 2004, Philadelphia, Pennsylvania, USA.*

Hui, S. K., Fader, P. S., & Bradlow, E. T. (2009). Path Data in Marketing: An Integrative Framework and Prospectus for Model Building. *Marketing Science*, *28*(2), 320–335. https://doi.org/10.1287/mksc.1080.0400

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.). https://otexts.org/fpp2/

Ipsos Retail Performance. (2018). *People Counters: People Counting Systems &amp; Technology*. Ipsos Retail Performance. https://www.ipsos-retailperformance.com/en/services/people-counters/

ITV. (2020). *ITV investigation finds high street footfall is on the decline in the Meridian region*. https://www.itv.com/news/meridian/2020-01-16/itv-investigation-finds-high-street-footfall-is-on-the-decline-in-the-meridian-region/

Ji, Y., Zhao, J., Zhang, Z., & Du, Y. (2017). Estimating Bus Loads and OD Flows Using Location-Stamped Farebox and Wi-Fi Signal Data. *Journal of Advanced Transportation*. https://doi.org/10.1155/2017/6374858

Johnson, N. E., Mandiola, P., Blankinship, C., Bonczak, B., & Kontokosta, C. E. (2019). Validating the Use of Wi-Fi Signals to Estimate Hyperlocal Urban Populations. *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*. https://doi.org/10.1109/BigData47090.2019.9006517

Kang, C. D. (2016). Spatial access to pedestrians and retail sales in Seoul, Korea. *Habitat International*, *57*, 110–120. https://doi.org/10.1016/j.habitatint.2016.07.006

Kholod, M., Nakahara, T., Azuma, H., & Yada, K. (2010). The influence of shopping path length on purchase behavior in grocery store. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-642-15393-8_31

Kontokosta, C. E., & Johnson, N. (2017). Urban phenology: Toward a real-time census of the city using Wi-Fi data. *Computers, Environment and Urban Systems*. https://doi.org/10.1016/j.compenvurbsys.2017.01.011

Lewis. (1982). Industrial and Business Forecasting Methods: A Practical Guide to Exponential Smoothing and Curve Fitting. *Butterworth Scientific*. https://doi.org/10.1002/for.3980010202

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management*. https://doi.org/10.1016/j.tourman.2018.03.009

Martin, J., Rye, E., & Beverly, R. (2016). Decomposition of MAC address structure for

DRAFT

DRAFT

granular device inference. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/2991079.2991098

Matte, C., Cunche, M., Rousseau, F., & Vanhoef, M. (2016). Defeating MAC address randomization through timing attacks. *WiSec 2016 - Proceedings of the 9th ACM Conference on Security and Privacy in Wireless and Mobile Networks*. https://doi.org/10.1145/2939918.2939930

Matzler, K., Mooradian, T. A., King, L., & Linder, A. (2010). Converting browser to buyers: An approach to measure and increase conversion rates in retailing. *Innovative Marketing*, *6*(1), 34–38.

McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*.

Ministry of Housing Communities & Local Government. (2019). *New Task Force to help revitalise high streets and town centres - GOV.UK*. https://www.gov.uk/government/news/new-task-force-to-help-revitalise-high-streets-and-town-centres

Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: Time series missing value imputation in R. *R Journal*, *9*(1), 207–218. https://doi.org/10.32614/rj-2017-009

Nagy, A. M., & Simon, V. (2018). Survey on traffic prediction in smart cities. *Pervasive and Mobile Computing*, *50*(October), 148–163. https://doi.org/10.1016/j.pmcj.2018.07.004

Nelson, C. R., & Schwert, G. W. (1982). Tests for Predictive Relationships Between Time Series Variables: A Monte Carlo Investigation. In *Source: Journal of the American Statistical Association* (Vol. 77, Issue 377).

Office for National Statistics. (2018). *Internet sales as a percentage of total retail sales (ratio) (%)*. https://www.ons.gov.uk/businessindustryandtrade/retailindustry/timeseries/j4mc/drsi

Pankratz, A. (1991). *Forecasting with Dynamic Regression Models*. John Wiley and Sons.

Parker, C., Ntounis, N., Millington, S., Quin, S., & Castillo-Villar, F. R. (2017). Improving the vitality and viability of the UK High Street by 2020: Identifying priorities and a framework for action. *Journal of Place Management and Development*, *10*(4), 310–348. https://doi.org/10.1108/JPMD-03-2017-0032

Phan, D., Xiao, L., Yeh, R., Hanrahan, P., & Winograd, T. (2005). Flow map layout. *Proceedings - IEEE Symposium on Information Visualization, INFO VIS*. https://doi.org/10.1109/INFVIS.2005.1532150

Portas, M. (2011). *The Portas Review: An independent review into the future of our high streets*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/6292/2081646.pdf

Ramos, P., Santos, N., & Rebelo, R. (2015). Performance of state space and ARIMA models for consumer retail sales forecasting. *Robotics and Computer Integrated Manufacturing*, *34*, 151–163. https://doi.org/10.1016/j.rcim.2014.12.015

Reades, J., Calabrese, F., Sevtsuk, A., & Ratti, C. (2007). Cellular Census: Explorations in Urban Data Collection. *IEEE Computer Society*. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4287441

Rekimoto, J., Miyaki, T., & Ishizawa, T. (2007). LifeTag: WiFi-based continuous location logging for life pattern analysis. *Lecture Notes in Computer Science (Including*

*Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-540-75160-1_3

Sapiezynski, P., Stopczynski, A., Gatej, R., & Lehmann, S. (2015). Tracking human mobility using WiFi signals. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0130824

Sevtsuk, A. (2014). Location and Agglomeration: The Distribution of Retail and Food Businesses in Dense Urban Environments. *Journal of Planning Education and Research*, *34*(4), 374–393. https://doi.org/10.1177/0739456X14550401

Sevtsuk, A., Huang, S., Calabrese, F., & Ratti, C. (2008). Mapping the MIT campus in real time using WiFi. *Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City*, *January 2008*, 326–338. https://doi.org/10.4018/978-1-60566-152-0.ch022

Shlayan, N., Kurkcu, A., & Ozbay, K. (2016). Exploring pedestrian bluetooth and WiFi detection at public transportation terminals. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*. https://doi.org/10.1109/ITSC.2016.7795559

Sila-Nowicka, K., & Fotheringham, A. S. (2016). A route map to calibrate spatial interaction models from GPS movement data. *GIScience 2016 Workshop on Analysis of Movement Data*. http://eprints.gla.ac.uk/128780/1/128780.pdf

Sila-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U., & Fotheringham, A. S. (2015). Analysis of human mobility patterns from GPS trajectories and contextual information. In *International Journal of Geographical Information Science* (Vol. 30, Issue 5). http://eprints.gla.ac.uk/128784/http://eprints.gla.ac.uk

Smolak, K., Kasieczka, B., Fialkiewicz, W., Rohm, W., Siła-Nowicka, K., & Kopańczyk, K. (2020). Applying human mobility and water consumption data for short-term water demand forecasting using classical and machine learning models. *Urban Water Journal*, *17*(1), 32–42. https://doi.org/10.1080/1573062X.2020.1734947

Soundararaj, B., Cheshire, J., & Longley, P. (2019a). Estimating real-time high-street footfall from Wi-Fi probe requests. *International Journal of Geographical Information Science*. https://doi.org/10.1080/13658816.2019.1587616

Soundararaj, B., Cheshire, J., & Longley, P. A. (2019b). Medium data toolkit - A case study on Smart Street Sensor Project. *Proceedings of the 27th Conference on GIS Research UK (GISRUK)*, *June*.

Springboard. (2020). *Insights | Footfall analytics and visitor demographics*. https://www.spring-board.info/insights

The Local Data Company. (2020). *Footfall Counting*. https://www.localdatacompany.com/products/footfall

Torrens, P. M. (2008). Wi-Fi geographies. *Annals of the Association of American Geographers*. https://doi.org/10.1080/00045600701734133

Van De Sanden, S., Willems, K., & Brengman, M. (2019). In-store location-based marketing with beacons: from inflated expectations to smart use in retailing. *Journal of Marketing Management*, *35*, 1514–1541. https://doi.org/10.1080/0267257X.2019.1689154

Vanhoef, M., Matte, C., Cunche, M., Cardoso, L. S., Piessens, F., Iminds-Distrinet, †, & Leuven, K. U. (2016). *Why MAC Address Randomization is not Enough: An Analysis of Wi-Fi Network Discovery Mechanisms*. https://doi.org/10.1145/2897845.2897883

Waddington, T., Clarke, G., Clarke, M. C., Hood, N., & Newing, A. (2019). Accounting for

DRAFT

Temporal Demand Variations in Retail Location Models. *Geographical Analysis*, *51*(4), 426–447. https://doi.org/10.1111/gean.12179

Wehrle, D. (2017). *Can footfall data predict future sales performance?* https://www.retail-week.com/retail-voice/can-footfall-data-predict-future-sales-performance/7027770.article?authent=1

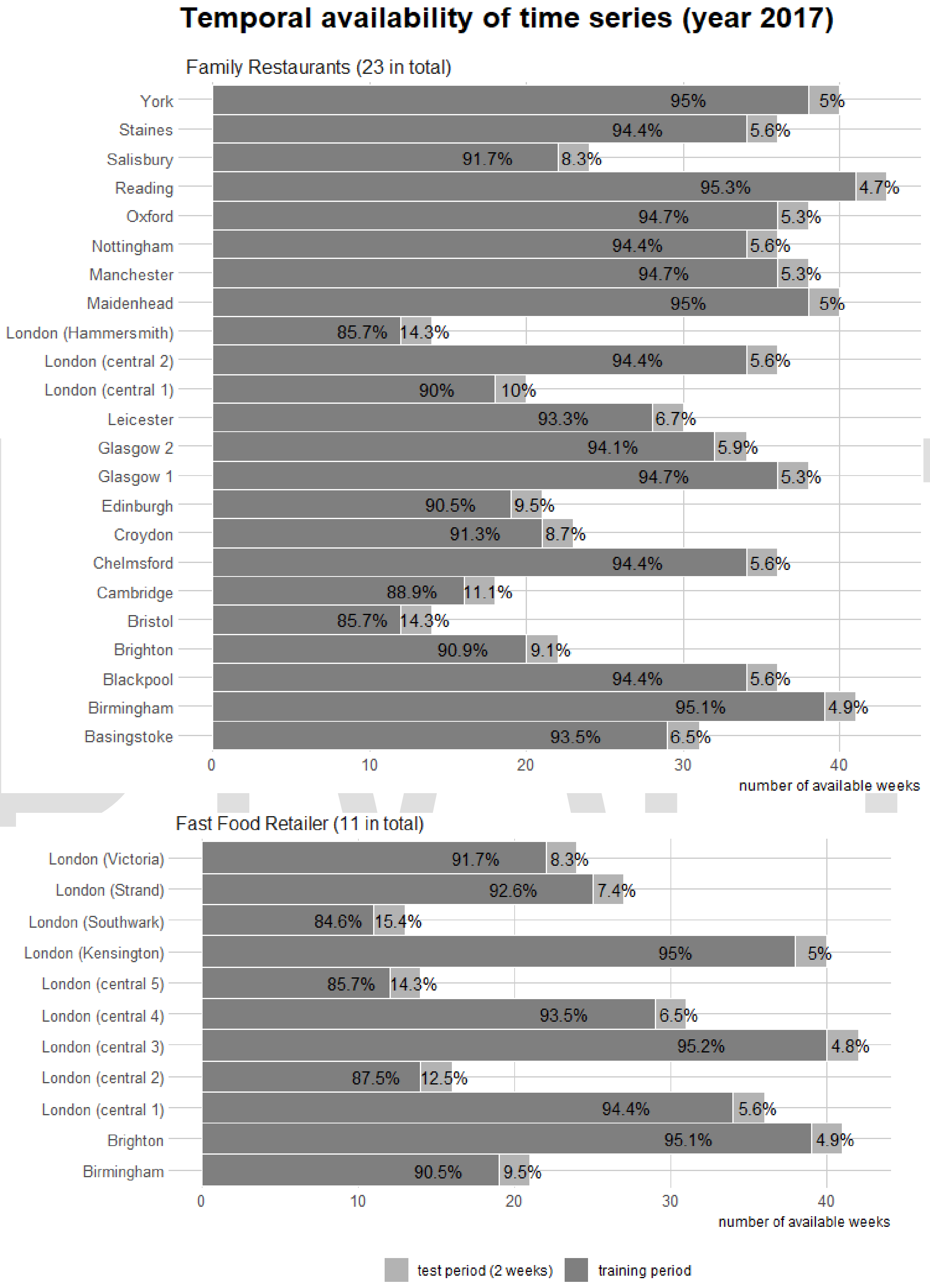
Wikipedia. (2018). *2018 Great Britain and Ireland cold wave*. https://en.wikipedia.org/wiki/2018_Great_Britain_and_Ireland_cold_wave

Wood, S., & Browne, S. (2007). Convenience store location planning and forecasting — a practical research agenda. *International Journal of Retail & Distribution Management*. https://doi.org/10.1108/09590550710736184

Zheng, W., Huang, X., & Li, Y. (2017). Understanding the tourist mobility using GPS: Where is the next place? *Tourism Management*. https://doi.org/10.1016/j.tourman.2016.08.009

DRAFT

# Appendix A: Temporal availability of the sample data


Temporal availability of time series (year 2017)
Family Restaurants (23 in total)
York 95% 5%
Staines 94.4% 5.6%
Salisbury 91.7% 8.3%
Reading 95.3% 4.7%
Oxford 94.7% 5.3%
Nottingham 94.4% 5.6%
Manchester 94.7% 5.3%
Maidenhead 95% 5%
London (Hammersmith) 85.7% 14.3%
London (central 2) 94.4% 5.6%
London (central 1) 90% 10%
Leicester 93.3% 6.7%
Glasgow 2 94.1% 5.9%
Glasgow 1 94.7% 5.3%
Edinburgh 90.5% 9.5%
Croydon 91.3% 8.7%
Chelmsford 94.4% 5.6%
Cambridge 88.9% 11.1%
Bristol 85.7% 14.3%
Brighton 90.9% 9.1%
Blackpool 94.4% 5.6%
Birmingham 95.1% 4.9%
Basingstoke 93.5% 6.5%
number of available weeks
Fast Food Retailer (11 in total)
London (Victoria) 91.7% 8.3%
London (Strand) 92.6% 7.4%
London (Southwark) 84.6% 15.4%
London (Kensington) 95% 5%
London (central 5) 85.7% 14.3%
London (central 4) 93.5% 6.5%
London (central 3) 95.2% 4.8%
London (central 2) 87.5% 12.5%
London (central 1) 94.4% 5.6%
Brighton 95.1% 4.9%
Birmingham 90.5% 9.5%
number of available weeks
test period (2 weeks)
training period