


Towards a Comprehensive Temporal Classification of Footfall Patterns in the Cities of Great Britain


Karlo Lugomer¹

Department of Geography, University College London
Pearson Building, Gower Street, WC1E 6BT, London, United Kingdom
karlo.lugomer.14@ucl.ac.uk

 <https://orcid.org/0000-0002-0820-3772>

Paul Longley²

Department of Geography, University College London
Pearson Building, Gower Street, WC1E 6BT, London, United Kingdom
p.longley@ucl.ac.uk

 <https://orcid.org/0000-0002-4727-6384>

Abstract

The temporal fluctuations of footfall in the urban areas have long been a neglected research problem, and this mainly has to do with the past technological limitations and inability to consistently collect large volumes of data at fine intra-day temporal resolutions. This paper makes use of the extensive set of footfall measurements acquired by the Wi-Fi sensors installed in the retail units across the British town centres, shopping centres and retail parks. We present the methodology for classifying the diurnal temporal signatures of human activity at the urban microsite locations and identify characteristic profiles which make them distinctive regarding when people visit them. We conclude that there exist significant differences regarding the time when different locations are the busiest during the day, and this undoubtedly has a substantial impact on how retailers should plan where and how their businesses operate.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases temporal classification, temporal profiles, time series cluster analysis, Wi-Fi sensors, retail analytics

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.43

Category Short Paper

Acknowledgements This research was sponsored by the UK Economic and Social Research Council (ESRC) Consumer Data Research Centre (ES/L011840/1) and a Ph.D. studentship co-funded by the ESRC and the Local Data Company.

1 Introduction

Spatial classifications have been a subject of a wide range of research papers in geography and GIScience. The popularity of clustering can be justified by the vast amount of readily available spatial data and need for interesting characteristics and patterns extraction [4]. Such classifications aim to describe the extent to which place A is similar to place B and to

¹ Award ES/L011840/1

² Award ES/L011840/1



© Karlo Lugomer and Paul Longley;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 43; pp. 43:1–43:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

use the derived clustering solution to make predictions about the characteristics of locations where data are incomplete and thus inform the industrial or public planning policymakers.

While the geographical classifications have been extensively covered in the past literature, little has been done to characterise bigger samples of places based on the recorded activity patterns on the finer temporal scales. In the past, this could have been done only by manual surveying, which is a costly and laborious process and does not enable the continuous data acquisition. These shortcomings have been addressed after the rapid development and wide-scale adoption of smartphones and Wi-Fi, GPS and Bluetooth technologies, which together made possible the collection of high volumes of data at small time periods, while, regarding spatial resolution, coming even to the granularity of an individual.

Knowing about where people go at which times in the weekly, daily or (sub-)hourly time frames has great practical importance for many fields. A good example of a sector where this is particularly relevant is retailing. Knowing what time of the day a specific retail unit can expect to see the highest number of potential customers passing by is vital to understanding whether that particular location is suitable for a specific category of retail business. For example, pubs and bar operators will be more interested in the places where footfall is significant in the evenings. This is contrary to the coffee shop operators, which will seek to exploit the large flow of morning commuters and midday lunch and coffee consumers.

This paper aims to use the footfall measurements collected by the Wi-Fi sensors to characterise urban microsite locations based on the features of the recorded temporal signatures of footfall. In other words, we are interested in finding out whether urban locations tend to differ in terms of diurnal temporal distribution of footfall and if so, how common each profile is. This classification presents the first step in acquiring a broader understanding of how urban places function and why people tend to find themselves at particular places at particular times of the day or days of the week.

2 National footfall data set

The data for this project were acquired through the network of Wi-Fi sensors installed by the Local Data Company (LDC) in the different UK cities from July 2015 until August 2017. They were placed in the three different categories of retail centres: shopping centres, out-of-town retail parks and, most commonly, urban town centres, i.e. high streets.

The initial set of retail centres for sensor installations was chosen based on the research sample design tailored to incorporate different cities of Great Britain, capturing centres of different sizes and diverse set of geodemographic characteristics of their catchment areas. The criteria for the sample locations outside London were dominant Output Area Classification (OAC) Supergroup, which is based on the cluster analysis of the 2011 Census variables [3]; town centre size expressed by the number of businesses and the town centre type, i.e. position of the centre in the national hierarchy. The primary criterion for the locations in Greater London was, on the other hand, the population size of retail centres' respective catchment areas. The Wi-Fi sensors were placed inside the retail units as close to the storefront window as possible.

2.1 Data acquisition

The Wi-Fi sensors work by receiving the probe requests sent out by the smartphones that are scanning for the available Wi-Fi networks. When a pedestrian carrying a smartphone with Wi-Fi and background scanning turned on passes by the Wi-Fi sensor, the sensor records the data contained in that probe request. The data includes the time stamp, the device

signal strength and the MAC address, which is hashed at the sensor level to preserve the privacy of the device owners. The idea is to derive the accurate measurements of the number of passers-by, monitor their fluctuations over time and use them to characterise locations based on their temporal distribution.

2.2 Data pre-processing

The approach described in the previous subsection comes with limitations, as derived footfall is prone to measurement errors due to factors which cause overcounting or undercounting [7].

Overcounting is caused by the fact that Wi-Fi sensors typically capture probe requests from devices which dwell locally (for example, workers in the retail unit and surrounding offices, devices other than smartphones such as printers, etc.). Undercounting stems from the fact that some passers-by do not have Wi-Fi probing capabilities enabled on their smartphones or they are simply missed due to the presence of some physical obstructions or signal interferences. The overcounting factors can be eliminated automatically by filtering methods and undercounting factors can to a certain extent be accounted for by the calibration in which passers-by are counted manually on site. After that, the ground truth is compared to the filtered sensor measurements, an adjustment factor is calculated by dividing those two figures and then used to adjust the measures. A more detailed treatment of those factors and ways to eliminate them is given in [7] and [9].

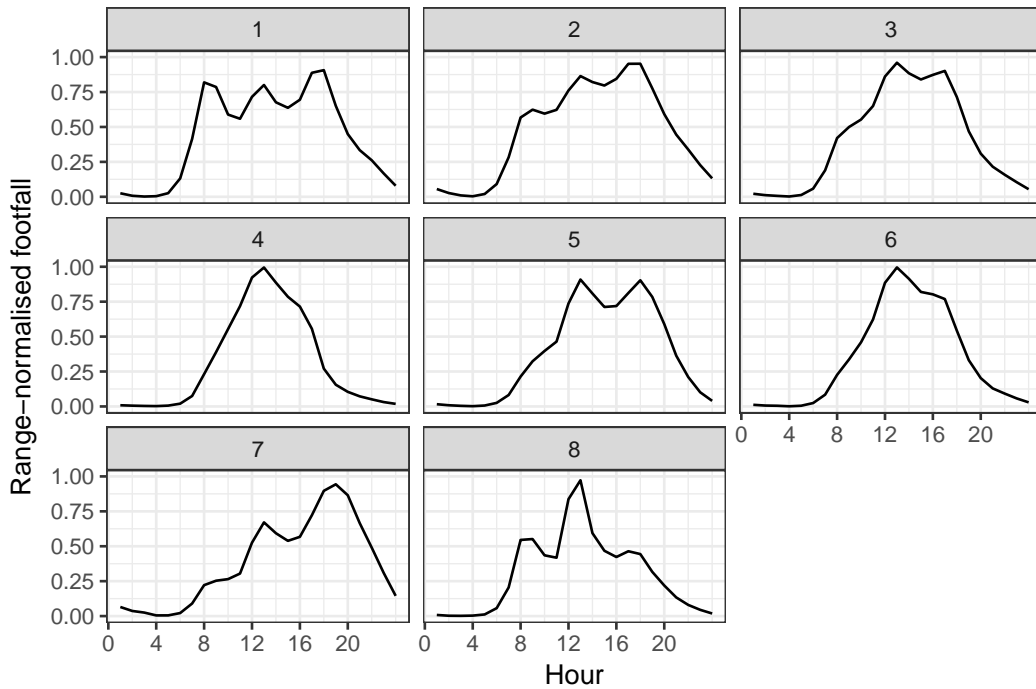
After identifying the devices of interest which serve as the proxy for people, the data were cleaned from outliers, as in this case we are interested in detecting the general functional characteristics of the location, rather than unusual events. The missing data were inputted by linear interpolation or inferred by taking the historical data for the corresponding hours and days of the week in cases where gaps of missing data were too wide for reliable interpolation. One representative weekly profile was then generated for every location by taking the median of every hour separately. The result comprised averaged time series each comprising 168 hours of the week for each of the 605 selected locations.

3 Clustering methodology

Since temporal profiles of different days of the week differ, it is not sensible to create a temporal classification for a "typical, average day" for each location. When the variation of footfall across time is visually inspected at the chosen location, Mondays through Thursdays generally display mutually similar profile shapes, whereas Fridays begin to differ if that location has pronounced nightlife activity. Same is true for Saturdays; however, due to the absence of the majority of workers, the daytime activity profile is usually different. In the first instance, the classification was therefore conducted for the footfall between Mondays and Thursdays for each location. The previously cleaned data were range-normalised.

The next step was to choose from the myriad of distance measures and clustering algorithms suitable for the time series clustering [2][5][6][8].

According to [1], the distance measures are commonly classified as (dis)similarities in either time, shape or change. The similarity in time can be regarded as a special case of similarity in shape, so the two go under the collective term shape-based methods [8][1]. In our case, we are interested in detecting the clusters of similar shapes of footfall profiles, however, at the same time, knowing at what time peaks or troughs occurred is also relevant. That said, we examined the shape-based methods more closely and Dynamic Time Warping (DTW) and Euclidean distances (ED) were found to be the most useful for our particular problem. A further justification for this is found in the recent detailed comparison of the



■ **Figure 1** Temporal profiles of microsite locations (*data source*: Local Data Company (2015–2017)).

different distance measures [2], in which it was concluded that despite some plausible progress made in the time series classification domain, DTW remains hard to beat and it is at the same time computationally less intensive than some of the newly proposed methods such as the Collection of Transformation Ensembles (COTE). In addition, it was found that, on reasonably large data sets comprising thousands (and in some cases only hundreds) of series, the difference between the classification error rate of the DTW and the ED diminishes [11]. In our case, the cleaned data set comprises 605 locations, which means that while warping may be advantageous, the ED could still suffice. Both ED and DTW with a relatively small width of the warping window equal to one hour were tested and coupled with several different partitional and hierarchical methods (k-means, PAM and Ward’s method). The ED fed into Ward’s algorithm provided the best trade-off between the mathematical validity, as measured by clustering validity indices [10] and interpretability.

4 Results and discussion

The optimal clustering solution was found to comprise eight distinct temporal profiles, as shown in the Figure 1.

The number of cases across clusters is unevenly distributed (Table 1), however, since we aimed to detect the interesting functional differences between places, trying to balance the number of cases would produce clusters in which such interesting properties would have been inherently lost.

According to the Table 1, the most common temporal profile in the retail centres of Great Britain (27.93% of the sampled microsite locations) is a two-peaked profile with a maximum around midday and late afternoon - appropriately labelled as *Consistent afternoons*. Unlike with similar profiles, such as One-directional commute, the drop of footfall during the early

■ **Table 1** The breakdown of cluster cases.

Cluster	Proposed name	Cases	Percentage (%)
1	Commute and lunch	84	13.88
2	Gradual rise	80	13.22
3	Consistent afternoons	169	27.93
4	Midday top	119	19.67
5	One-directional commute	29	4.79
6	Lunch time with minor afternoon commuter inflow	90	14.88
7	Quiet mornings, busy evenings	19	3.14
8	Busy lunchtimes with both commuting peaks	15	2.48
	Total	605	100.00

afternoon, i.e. between 2 pm and 5 pm is almost insignificant, which means that such locations benefit from consistently high footfall throughout most of the day. The second most common temporal profile (*Midday top*, comprising 19.67% locations) is a simple one-peaked profile with maximum activity recorded around midday. Such locations likely attract lunch goers. Next cluster is *Lunch time with minor afternoon commuter inflow*, comprising 14.88% of the locations. It is a one-peaked profile with a minor secondary peak in the late afternoon, which is not strictly speaking a peak, but rather a part of the profile where a drop of footfall slows down due to the impact of late afternoon commuters. However, in these locations, commuters are not as numerous as is the case in some other locations, so secondary peaks are not formed.

Similarly numerous, clusters 1 (*Commute and lunch*) and 2 (*Gradual rise*) account for 13.88% and 13.22% of the locations, respectively. Both are three-peaked profiles and are characterised by busier customer traffic during all three characteristic periods during the day - morning rush hour, lunchtime and afternoon rush hour. The difference is that Gradual rise cluster expects more customers towards the end of the day and intra-day differences of footfall volume are not as pronounced. Commute and lunch, on the other hand, has more pronounced peaks and intermediate drop and corresponding locations may expect the similar volume of passing footfall during all three periods, with a peak in the late afternoon recording slightly higher footfall than other two peaks.

The profiles captured by the remaining three minor clusters are not as commonly encountered across the British retail space, however, since they are functionally specific, it is worth further investigating their temporal distribution of footfall.

As was already mentioned, *One-directional commute* cluster is characterised by the two-peaked profiles of microsite locations (4.79%) with a more significant drop in customer traffic after the lunchtime, as compared to the similarly shaped Consistent afternoons cluster. Interestingly, these locations do not record any peak during the morning rush hour but do record one during the afternoon rush hour. Next, *Quiet mornings, busy evenings* cluster (3.14%) is to a certain extent similar to the Gradual rise locations, but morning footfall is much smaller, and differences between the peaks are much more pronounced. Moreover, the maximum footfall is, on average, reached between 7 pm and 8 pm, which seemingly makes these locations more attractive for the dinner and pub goers. And finally, occurring at only 15 of the sampled locations (2.48%), *Busy lunchtimes with both commuting peaks* is characterised by its distinctive dominant lunchtime peak and two smaller peaks during the rush hours.

5 Conclusion and future work

The initial aim of this paper was to test whether different microsite locations in urban areas display different diurnal footfall patterns and if that was the case, to further inspect if the readings from the Wi-Fi sensors could serve to derive the temporal classification of footfall patterns. This cluster analysis proved that there exist significant differences in footfall patterns among urban microsite locations. We identified eight clusters of distinct functional characteristics and described each of them.

As part of the future work, we aim to combine the identified profiles with the ancillary data on local vacancy rates, retail occupancy structure, i.e. local compositions of store types, in addition to the relative distributions of footfall that were presented here. The geodemographic characteristics of the retail centre catchment areas or the underlying Workplace Zones will also be considered as the relevant factors worth further investigation. The ultimate goal is to identify and explain the functional characteristics of the national set of retail centres based on both structural and dynamical properties of space.

References

- 1 Anthony Bagnall, Eamonn Keogh, Stefano Lonardi, Gareth Janacek, et al. A bit level representation for time series data mining with shape based similarity. *Data Mining and Knowledge Discovery*, 13(1):11–40, 2006.
- 2 Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31(3):606–660, 2017.
- 3 Christopher G Gale, A Singleton, Andrew G Bates, and Paul A Longley. Creating the 2011 area classification for output areas (2011 oac). *Journal of Spatial Information Science*, 12:1–27, 2016.
- 4 Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of intelligent information systems*, 17(2-3):107–145, 2001.
- 5 Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.
- 6 T Warren Liao. Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874, 2005.
- 7 Karlo Lugomer, Balamurugan Soundararaj, Roberto Murcio, James Cheshire, and Paul Longley. Understanding sources of measurement error in the wi-fi sensor data in the smart city. In *Proceedings of the 25th GIS Research UK (GISRUK) Conference, Manchester, UK, April 18-21, 2017*, 2017.
- 8 Pablo Montero, José A Vilar, et al. Tsclust: An r package for time series clustering. *Journal of Statistical Software*, 62(1):1–43, 2014.
- 9 Roberto Murcio, Bala Soundararaj, and Karlo Lugomer. Movements in cities: Footfall and its spatio-temporal distribution. In Paul Longley, James Cheshire, and Alex Singleton, editors, *Consumer Data Research*, chapter 6, pages 85–96. UCL Press, London, 2018.
- 10 Mark A Newell, Dianne Cook, Heike Hofmann, and Jean-Luc Jannink. An algorithm for deciding the number of clusters and validation using simulated data with application to exploring crop population structure. *The Annals of Applied Statistics*, pages 1898–1916, 2013.
- 11 Jin Shieh and Eamonn Keogh. i sax: indexing and mining terabyte sized time series. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631. ACM, 2008.