A clustering study to verify four distinct monthly footfall signatures: a classification for UK retail centres

Technical Report 1 (Version 5)*

Christine L Mumford[†1], Catherine R Parker[‡2], Nikolaos Ntounis[§2] and Ed Dargan[¶2]

[1]School of Computer Science & Informatics, Cardiff University
[2]Institute of Place Management, Manchester Metropolitan University

December 11, 2017

1 Note on Version 5

Adds 30 retail centres from London and Greater London, making a total of 144 centres.

2 Note on Version 4

Replaces term "Convenience" with "Multifunctional"; includes two more towns from 112 in Version 3 of this report, to 114 in this version. All the experiments have been repeated and corrections/improvements have been made to the text.

Note on Version 3

A full year of data for 2016 has now become available and all the experiments documented in earlier versions of this report have been repeated to include the extra data. Some improvements have been made to the analyses where appropriate, and the daily and hourly signature work has been properly integrated into the report.

[*] Updated versions of this report will be issued as appropriate. For example, when more data becomes available.

[†] MumfordCL@cardiff.ac.uk (Contact Author)
[‡] C.Parker@mmu.ac.uk
[§] N.Ntounis@mmu.ac.uk
[¶] Edmund.Dargan@stu.mmu.ac.uk

Note on Version 2

In Version 2 we have added our work on daily and hourly signatures for retail centres.

Overview

This report describes the application of *K-Means clustering*, *Principal Components Analysis*, and various statistical techniques to retail footfall data and verifies the existence of four distinct monthly footfall signature types, exactly as first proposed in the High Street UK2020 project (see later). Additionally, we have examined footfall profiles for days of the week, and discovered two distinct patterns for these, which have helped develop some further useful classifications. The footfall data is supplied by Springboard LTD$^{\text{TM}}$and consists of hourly records broadcast from several hundred counters located in traditional retail centres throughout the UK. Before using any of the data, we check its completeness for every counter by identifying any missing hourly data. The most recent computed completeness figure at 97 % proves the counters to be very reliable. The the main retail centre types identified from the monthly profiles are: *comparison*, *holiday*, *speciality* and *multifunctional* (previously known as convenience/community). Comparison shopping centres tend to be located in the larger town and city centres and their monthly signatures can be identified by a footfall peak in December, coinciding with the Christmas preparation period. Holiday towns are busier in the summer months and footfall drops right down in the winter, whilst multifunctional centres tend to have more of a flat profile throughout all the months of the year. Finally, speciality centres seem to be somewhat of a "hybrid" type between comparison and holiday, insofar as they have peaks in the summer and in December, although these peaks are not as pronounced as they are in pure comparison and holiday centres. Of all the signature classes, the multifunctional class appeared to be the least consistent, and contained a great variety of large, medium and small towns, in terms of their footfall volume.

Beginning with a vast amount of raw hourly data gathered from each counter (8760 readings per year), various annual, monthly, daily and hourly totals and averages are computed for each of 144 retail centres, and then further processed to produce footfall profiles for "standard years", "standard weeks" and "standard days" for each retail centre. We began our work on the monthly profiles from "standard year" patterns produced for each retail centre. Our methodology of choice is the K-Means clustering techniques, and using this technique reproduces the four signature types first proposed in the High Street UK2020 study. In addition the K-Means technique is able to classify each of the 144 retail centres as one of these four signature types. During the analysis and classification, we use *Silhouette Coefficients* to help assess the distinctness and quality of the clusters. When applying K-Means to any data, it is up to the user to choose how many clusters he/she would like, by setting the value of K to some integer value greater or equal to two. We discovered that setting $K = 2$ produces signatures for comparison and holiday, $K = 3$ identifies comparison, holiday and speciality, and $K = 4$ reveals all four of the expected signatures types. Furthermore when we assess the "strength" of the clustering (in other words, their degree of separation or distinctness), the clustering appears to be strongest when $K = 2$ indicating that comparison and holiday types produce the most distinctive profiles. These scores then drop a little between $K = 2$ and $K = 3$ and then little further for $K = 4$. Setting $K > 4$ picks up patterns very similar to the four main signatures already identified. For this reason we curtail our analysis at $K = 4$.

Observing the "standard year" profiles for individual retail centres, it is clear that some centres match one of the four standard templates very closely indeed, whilst for others the resemblance to one of the four classes can be rather more difficult to spot visually. As well as sharing characteristics with more than one type, some centres demonstrate patterns unique to themselves. To visualize aspects of the great variability between centres, we computed distance values for every retail centre from each of the four signature templates, to give measures of how closely each centre resembles the template signatures. The resulting graphical plots indeed show clearly that a simple "all or nothing" classification does not tell the whole story. One particularly interesting observation is that the signature profiles for the holiday towns are well separated from all the other retail centre profiles, and thus emphasizes the very distinct nature of footfall patterns in holiday towns. The other clusters (comparison, speciality and multifunctional) all show some degree of overlap with each other. However, by applying Principal Components Analysis (PCA) to the monthly profiles for our retail centres, we are able to separate all the clusters very well, and produce a two dimensional plot with barely any overlaps at all, clearly demonstrating that the clusters for comparison, speciality, multifunctional as well as holiday represent a viable classification for the retail centres. PCA is a completely independent process that does not rely on any information from the K-Means classification to produce its findings.

In addition to the "cleaning up" of the classification clusters, PCA supports our observations that December, July and August are key months for distinguishing between the retail centre signature types, with a December peak associated with Christmas shopping in comparison centres, and a July and August peak associated with the height of the tourist season in the UK. March was also identified as an important month, although its usefulness in distinguishing between signature types is less intuitive and warrants further investigation. An additional analysis was carried out, to investigate whether there is any relationship between total footfall with signature type. From this work we are able to deduce that comparison retail centres tend to be busier than other types of centre, and this ties in well with our observation that comparison sites seem to consist mostly of large city and town centres.

Following our analysis of monthly footfall data, we carried out some further work looking for signature patterns within the days of the week. From this we identified two distinct profiles: Type 1 with a fairly level footfall pattern for Monday through to Saturday, and a big drop in footfall on Sunday, and Type 2 with a clear peak on Saturday and a smaller drop for Sunday. On closer examination, it became clear that the Type 2 daily profile was typical of comparison and holiday towns and also the larger towns identified as multifunctional by the monthly signatures. Finally, we have briefly examined footfall patterns for the 24 hours of the day, which demonstrated that the centres are busiest in the middle of the day.

The work outlined in this report demonstrates that distinct monthly and daily (and to a lesser extent hourly) signatures exist for UK retail centres. However, this represents only the first stage of this research. The crucial question to answer is whether knowing the classification for individual retail centres can actually help stakeholders improve their offer to customers and make their centres more successful. This will form an important component of future reports in this series.

3 Introduction

The growth of internet shopping is having a profound effect upon traditional retail centres, like the High Street [16]. Nevertheless, the recent Digital High Street Report [15] demonstrates that the internet revolution can be a constructive, rather than destructive, force of change. Furthermore, the impact of internet shopping is not felt equally across all centres [14], and data suggests that large metropolitan, as well as small speciality centres, are faring better than small and medium sized centres that lack a speciality offer. Smaller centres are finding it difficult to adapt to changes in consumer behaviour. Recent exploratory research from Manchester Metropolitan University as part of the ESRC-funded High Street UK2020 project [10], has used Springboard$^{\text{TM}}$footfall data to typologise centre and town types, based upon their activity profiles. They have found initial evidence of specific footfall "signatures" representing comparison shopping centres, holiday towns, speciality centres, and multifunctional centres [9] (which they previously called convenience/community centres). Comparison shopping centres are typified by a peak in footfall in the month of December, presumably coinciding with pre-Christmas spending. On the other hand, multifunctional centres tend to have a much flatter profile all the year round, whilst holiday and speciality towns attract more of their visitors in the warmer weather of the summer months, because they have some special attraction, such as historical architecture, or they are located near the sea, or in the midst of National Parks, or other areas of natural beauty. Holiday and speciality towns can be distinguished from each other by observing a higher summer peak for holiday towns and secondary peak in December for speciality towns. Of particular interest, and one of the key motivations for this present research, is the preliminary evidence in [9] suggesting that centres with footfall patterns adhering most closely to one of the four typical activity profiles, tend to perform better than those without a clear profile. In other words, towns that have a definite "offer" for their catchment appear to attract more customers. Retailers that are located in places that attract more footfall will tend to perform better: "the strong correlation between spend and footfall across the UK indicates that footfall is a robust barometer of performance [7].

The key contributions of the present report are as follows:

- Verification of the four footfall signatures indicated in the UK2020 study.
- Identification of the key months for distinguishing between the four different footfall profiles.
- Identification of subgroups within the four main grouping: 1) larger towns typified by a strong weekend peak in the daily footfall pattern and 2) smaller towns that

are very quiet at weekends.

We use the K-Means clustering technique to classify the activity profiles of the retail centres, and validate the associated signature types. Following this, we apply Principal Components Analysis (PCA) to demonstrate that K-Means is able to produce clusters that are clearly separate from each other. Additionally, PCA is able to identify a few months that are key in distinguishing between the four signature footfall patterns. The platform used is an iMac Intel i7 quad core 3.5 GHz with 32 GB RAM.

Section 4 describes our methodology, starting with details of how hourly footfall data from all the retail centres is processed to obtain monthly totals, and moving on to the clustering and statistical analysis techniques used. Next comes Section 5 where we present the results of our K-Means clustering experiments on our retail centre data and analyse the quality of clusters obtained. An examination of how the signature type is related to total annual footfall is also included to assess whether certain types of centre (such as comparison centres) are busier than others. Finally in this Section, PCA is applied to help verify the distinctness of the footfall signature classification, and also identify key distinguishing months typifying the different signature types. The findings in this report are finally summarized in Section 6, where we also outline the next steps planned for our research.

4 Methodology

In this report we analyse monthly, daily and hourly footfall counts on a large set of data. Our main purpose is to verify or dispute the existence of the four distinct signatures previously observed on a much smaller set of data in UK2020. It is necessary that our methodology is focussed on automating data processing tasks, so that large quantities of hourly recorded data can be combined into monthly or daily totals quickly, and multiple graphs and results from statistical analyses can be produced in a matter of seconds.

In the following subsections we first describe how we store and process the raw data (Subsection 4.1), and then we go on to explain in Subsection 4.2 how the K-means clustering algorithm works on our data. Next, we define the Silhouette Coefficient and discuss how it can be used to help assess how well our data fits into the clusters to which it is assigned (Subsection 4.3), and finally we briefly introduce Principal Components Analysis, which we will return to in Section 5.3.

4.1 Data Files

Footfall data provided by Springboard UK Limited$^{\mathrm{TM}}$consists of hourly footfall counts from some 500 counters located in about 200 retail centres around the UK. Some of these counters have been operating since the start of 2006, whilst others have been installed more recently. Most of our analysis requires at least one full year of data, so some locations cannot as yet be included.

The historical data was provided by Springboard as 11 comma separated values (csv) files consisting of records in the format seen in Table 4.1. The data was validated and stored in a single Hierarchical Data Format file (HDF5)[1] [1]. Python and Pandas have been used to prepare and process the data, and scikit-learn [12] has provided the clustering toolkit and also the Principal Components Analysis module used later.

Region	Retail Centre	Camera Location	Hourly Timestamp	Footfall Count

Table 4.1: *Format of Springboard raw files*

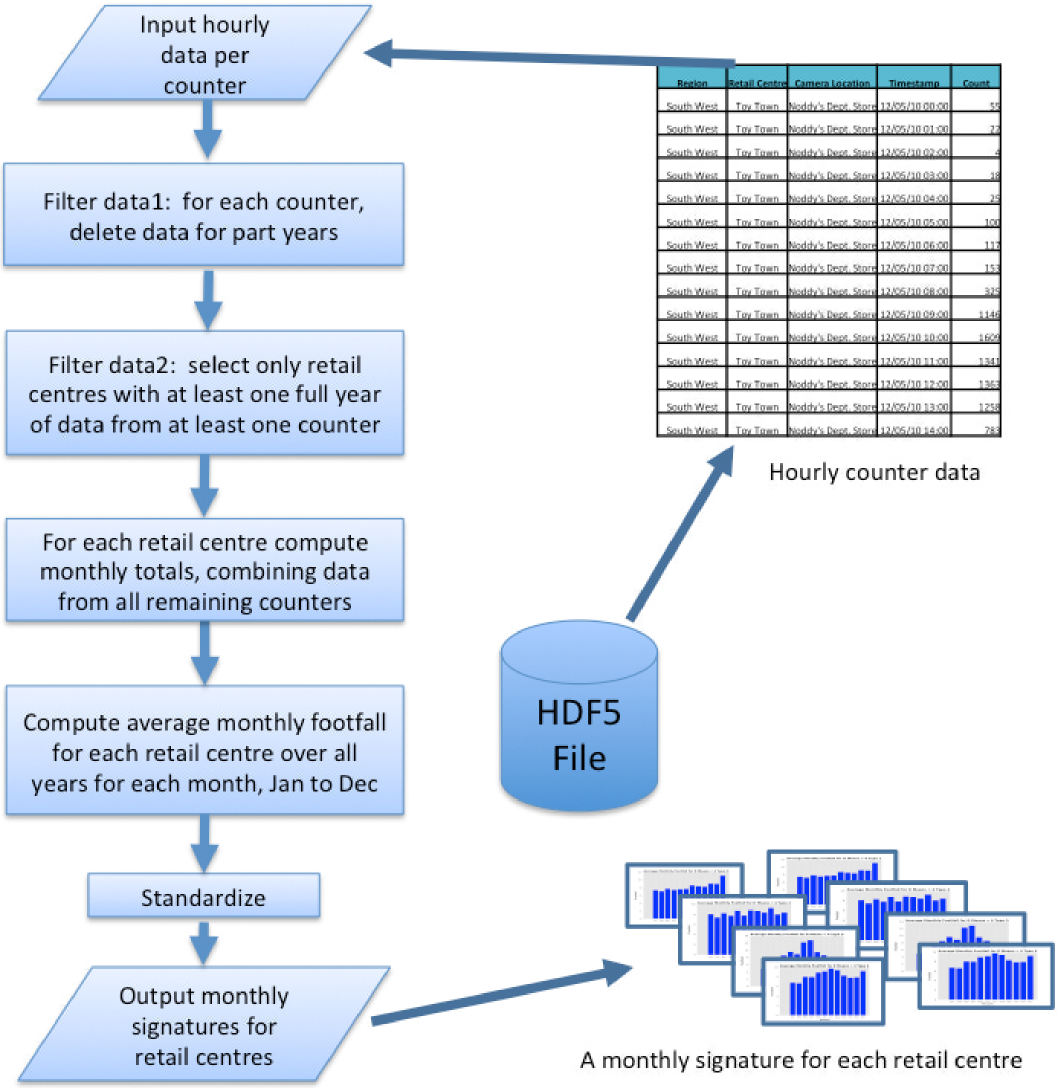
Data Preparation For this study we examine monthly, daily and hourly footfall profiles for retail centres. Before beginning the study however, validation of the data is important. We check the completeness of the data by examining the hourly counts recorded in the raw data supplied by Springboard for each of the counters. We calculate the total number of hourly records submitted since the counter was first switched on, and then divide that total by the number of elapsed hours in the same time period. From this, we compute a percentage activity for each counter. The arithmetic mean of these averages for the all the counters is 97 %, which demonstrates high reliability of the counters when taken as a whole. Nevertheless, a handful of counters have recorded rather low activity percentages and these are being investigated further. A number of factors can impede the function of the counter - including power outage or being unwittingly obscured by signs or other obstacles. Each counter is checked daily by Springboard enabling the research team to get the information necessary to decide which counters should be excluded from the data set in future.

Moving on to computing the profiles (or signatures), the first step is to find a way to combine hourly data from different counters into single monthly, daily or hourly totals for each retail centre. The second step uses these single monthly, daily or hourly totals to compute a "representative year, week or day", respectively, for each retail centre, depending on the type of analysis undertaken.

Monthly Footfall Signatures Monthly footfall signatures consist of mean footfall values for each calendar month of the year. For example, assuming there are four complete years of data for a particular retail centre, the January footfall figure will be computed by adding together the footfall counts for all the Januaries and then dividing by four. The other eleven months will be computed similarly.

For each retail centre we compute our monthly footfall totals from the original hourly data held in the HDF5 file. Using all the counters spread around a retail centre is likely to provide a more balanced picture than would be obtained by simply choosing one counter, and this approach should also prove less susceptible to issues with individual counters or temporary local road or pavement closures etc. However, we are mindful that new counters are installed part way through particular years is various locations, and we avoid the distortion that these new additions would make to monthly counts

[1] "HDF5 is a unique technology suite that makes possible the management of extremely large and complex data collections."


Input hourly data per counter
Filter data1: for each counter, delete data for part years
Filter data2: select only retail centres with at least one full year of data from at least one counter
For each retail centre compute monthly totals, combining data from all remaining counters
Compute average monthly footfall for each retail centre over all years for each month, Jan to Dec
Standardize
Output monthly signatures for retail centres
HDF5 File
Hourly counter data
A monthly signature for each retail centre

Region	Retail Centre	Camera Location	Timestamp	Count
South West	Toy Town	Noddy's Dept. Store	12/05/10 00:00	55
South West	Toy Town	Noddy's Dept. Store	12/05/10 01:00	22
South West	Toy Town	Noddy's Dept. Store	12/05/10 02:00	4
South West	Toy Town	Noddy's Dept. Store	12/05/10 03:00	18
South West	Toy Town	Noddy's Dept. Store	12/05/10 04:00	25
South West	Toy Town	Noddy's Dept. Store	12/05/10 05:00	100
South West	Toy Town	Noddy's Dept. Store	12/05/10 06:00	117
South West	Toy Town	Noddy's Dept. Store	12/05/10 07:00	153
South West	Toy Town	Noddy's Dept. Store	12/05/10 08:00	325
South West	Toy Town	Noddy's Dept. Store	12/05/10 09:00	1146
South West	Toy Town	Noddy's Dept. Store	12/05/10 10:00	1609
South West	Toy Town	Noddy's Dept. Store	12/05/10 11:00	1341
South West	Toy Town	Noddy's Dept. Store	12/05/10 12:00	1363
South West	Toy Town	Noddy's Dept. Store	12/05/10 13:00	1258
South West	Toy Town	Noddy's Dept. Store	12/05/10 14:00	783

Figure 4.1: *Flowchart to show the data preparation required for our clustering experiments*

by including only full years of data from each counter. In this way we also remove the influence of counters that are “switched off” part way through a year. Thus for every retail centre, the first step is compute an average footfall count for each hour of each day, giving us 365 × 24 hours for each year that counters have been operational in a retail centre. Within an individual retail centre, these hourly counts represent an arithmetic mean taken for each hour separately, by summing the count from each included counter active in that retail centre, and then dividing that sum by the number of active counters. The hourly averages are then added up to give totals for each month, and monthly totals for a “representative year” are next computed, as explained above. Finally, we take the representative year for each retail centre and “standardize” it, by transforming total annual footfall for each centre to be 100 %, and that 100 % is distributed over the months, January to December, in proportion to their contribution to the 100 % total. An outline of our data preparation process is illustrated in a flowchart in Figure 4.1.

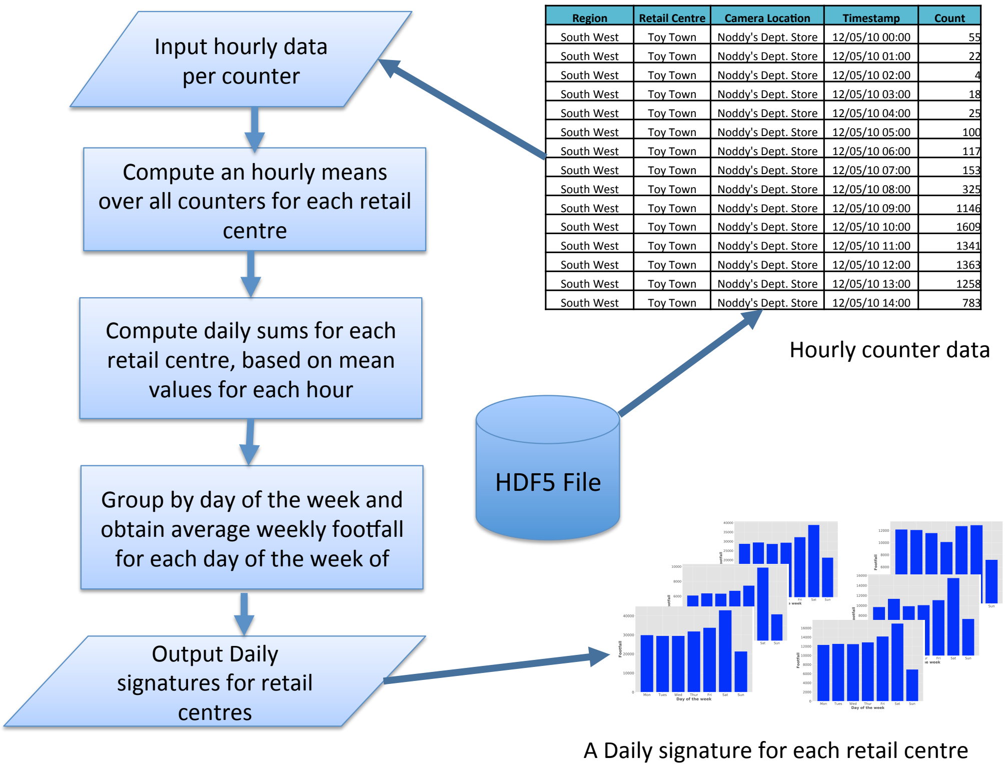
Daily Footfall Signatures For daily counts, data is processed from the same HDF5 file as is used for extracting the monthly footfall counts. However, this time it is not necessary to filter out data from counters present for only part years, because with daily footfall signatures, any distortion will be minimal. First of all hourly footfall counts are extracted for each of the retail centres in our study and averaged over all the counters in that retail centre. The hourly counts are next combined to produce daily counts, which are finally averaged for each day of the week, to obtain for each retail centre, the mean daily footfall for each day of the week covering the whole period for which a retail centre has had counters installed. Thus, our daily footfall signatures consist of averages for Monday, Tuesday,...,Sunday. The data processing to obtain the daily footfall signatures is summarized in Figure 4.2

Hourly Footfall Signatures Springboard data is processed in a similar way as described above for the daily counts. Instead of computing average daily counts for the days of the week however, each hourly signature comprises twenty four average footfall counts for each hour of the day, from 00 am, to 23 pm.

4.2 Clustering and K-Means

K-means clustering [3] is popular method for cluster analysis in data mining. K-means clustering aims to partition n data points into K clusters (where the value of K is selected in advance by the user), so that each observation belongs to exactly one cluster. The “centre of gravity” for each cluster, known as its “centroid”, serves as a representative for that cluster. Because the problem of finding the correct centroids is computationally difficult (NP-Hard), heuristic methods are used that quickly converge to local optima. Thus, generally speaking, a very slightly different solution will be obtained every time a K-means computation is carried out on the same data, due to random variation. However, these variations are very slight indeed, so can be ignored for practical purposes.

Given a set of observations $(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional real vector, K-means clustering aims to partition the n observations into K $(\leq n)$ cluster


Input hourly data per counter
Compute an hourly means over all counters for each retail centre
Compute daily sums for each retail centre, based on mean values for each hour
Group by day of the week and obtain average weekly footfall for each day of the week of
Output Daily signatures for retail centres
HDF5 File
Hourly counter data
A Daily signature for each retail centre

Region	Retail Centre	Camera Location	Timestamp	Count
South West	Toy Town	Noddy's Dept. Store	12/05/10 00:00	55
South West	Toy Town	Noddy's Dept. Store	12/05/10 01:00	22
South West	Toy Town	Noddy's Dept. Store	12/05/10 02:00	4
South West	Toy Town	Noddy's Dept. Store	12/05/10 03:00	18
South West	Toy Town	Noddy's Dept. Store	12/05/10 04:00	25
South West	Toy Town	Noddy's Dept. Store	12/05/10 05:00	100
South West	Toy Town	Noddy's Dept. Store	12/05/10 06:00	117
South West	Toy Town	Noddy's Dept. Store	12/05/10 07:00	153
South West	Toy Town	Noddy's Dept. Store	12/05/10 08:00	325
South West	Toy Town	Noddy's Dept. Store	12/05/10 09:00	1146
South West	Toy Town	Noddy's Dept. Store	12/05/10 10:00	1609
South West	Toy Town	Noddy's Dept. Store	12/05/10 11:00	1341
South West	Toy Town	Noddy's Dept. Store	12/05/10 12:00	1363
South West	Toy Town	Noddy's Dept. Store	12/05/10 13:00	1258
South West	Toy Town	Noddy's Dept. Store	12/05/10 14:00	783

Figure 4.2: *Flowchart to show the data preparation required to produce daily footfall signatures*

$C = \{C_1, C_2, \ldots, C_K\}$ so as to minimize the within-cluster sum of squares (sum of distance functions of each point in the cluster to the cluster centroid). In other words, its objective is to find for each x_i its best fitting cluster, $A(x_i)$ given by:

$$A(x_i) = arg \min_{C} \sum_{j=1}^{K} \sum_{x_p \in C_j} d(x_i, x_p) \tag{4.1}$$

Before finally settling on the choice of K-Means as the clustering algorithm to use for our study, we experimented briefly with some other approaches, principally Affinity Propagation [4] and Meanshift [2]. However, K-Means produced the most reliable results, according to the measured silhouette values (described below), and meeting deadlines for the present project precluded a thorough comparative study of clustering methods. It is worth pointing out however, that a fuller study of methods is worth considering as future work. We use Python and Scikit-learn [12] for coding our K-Means analysis.

4.3 Assessing the Quality of Clustering

A number of metrics exist to assess the quality of assignment of data to clusters, and several of these are provided in the Scikit-learn package. However, only one in the package, called the Silhouette Coefficient, is suitable when no "ground truth" labels are available. For example, ground truth labels can be used when a clustering algorithm is applied to an automated pattern recognition task, such as for hand written character identification. A subset of characters can be labelled by humans, and then a clustering algorithm can be assessed on the basis of how many hand written characters are correctly classified. In the present study however, we have no "ground truth". Indeed the very point of this clustering exercise is to find the "ground truth" and thus classify the retail centres. For this reason we shall use the Silhouette Coefficient for our study. Although, we must always bear in mind the context in which we are working, i.e. why we are applying a clustering technique to retail centre signatures in the first place. We are hoping that the classification of retail centres into distinct types will help those centres better focus their "offer" to attract more customers. If knowing what type of footfall profile a particular centre matches most closely proves to be of no help in informing how stakeholders can improve their offer and performance, then the whole exercise will have no practical value. After all, the features provided to the clustering method, which are in our case monthly signature values standardized in a way we have devised ourselves, consist of a tiny subset of subjectively selected features of retail centres, which may or may not be the most important features for our purposes. The dangers of blindly pursuing a mode of classification have been succinctly pointed out as long ago as 1912 by Charles Mercier [8]

> *"Classification is often spoken of, in books on Logic, as if there were but one ideally right mode of it, –the Natural Classification– and all other modes are wrong. This is a mistake. Classifications are made by us for our*

> *convenience; and whether a classification is right or wrong depends on whether or not it is suitable to the purpose for which it is made....... The nature of the classification that we make.......must have direct regard to the purpose for which the classification is required. In as far as it serves this purpose, the classification is a good classification, however 'artificial' it may be. In as far as it does not serve this purpose, it is a bad classification, however 'natural' it may be."*

4.3.1 Silhouette Values

Silhouette coefficients provide a technique to assess the validity and consistency of an assignment of data objects to clusters, following the application of a clustering algorithm such as K-Means. It is available in Scikit-learn, which is convenient for us. The Silhouette metric, first described by Peter J. Rousseeuw [13], provides a useful measure of how well each object lies within its assigned cluster. Silhouette values range from -1 to 1, where a value close to +1 indicates that an object is a good fit within its own cluster and a poor fit to neighbouring clusters, and a value close to 0 indicates that an object is on or very close to the decision boundary between the object's assigned cluster and a neighbouring cluster. A negative value indicates that an object has probably been assigned to the wrong cluster. If most objects have a high Silhouette value, then the clustering configuration is likely to be a good one. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters. However, as pointed out by Rousseeuw in his 1987 paper, care must be taken when interpreting Silhouette results, particularly when there is an "outlier" present, in terms of members of one of the clusters having very different properties from members of all the other clusters. In the presence of an "outlier class", Silhouette values for clustering assignments consisting of only two clusters may be high, even though most of the objects are artificially grouped into one "super cluster", with the second cluster formed by the (usually small) outlier class. We shall see that this situation is exactly what happens in our analysis of monthly footfall signatures in Section 5. Silhouette values can be calculated using any distance metric, such as the Euclidean distance or the Manhattan distance. We will be using Euclidean distances in the present study.

Assume our data have been clustered into K clusters, using K-means. For each data item, x_i, let $a(x_i)$ be the average dissimilarity of x_i with all other data items within the same cluster, A. Generally x_i is not the only member of its cluster. However, when cluster A contains only a single object, $s(x_i)$ is simply set equal to zero, as recommended in [13]:

$$a(x_i) = \frac{1}{m_A - 1} \sum_{j=1}^{m_A} d(x_i, x_j), \quad \forall x_j \in A \;\; \textit{such that } x_j \neq x_i \tag{4.2}$$

where m_A is the number of items in the same cluster as x_i, which we have called cluster A. $d(x_i, x_j)$ denotes the dissimilarity (which is in this case the Euclidean distance) between

points x_i and x_j. We can interpret $a(x_i)$ as how well x_i fits into its assigned cluster (the smaller the value, the better the assignment).
We then define the average dissimilarity of point x_i to any cluster $C \neq A$ as the average of the distance from x_i to all points in C:

$$d(x_i, C) = \frac{1}{m_C} \sum_{j=1}^{m_C} d(x_i, x_j), \quad \forall x_j \in C \tag{4.3}$$

Once a value of $d(x_i, C)$ has been computed for each cluster, $C \neq A$, we select the smallest of these values denoted by $b(x_i)$, which is the lowest average dissimilarity of x_i to any cluster, other than A. The cluster with this lowest average dissimilarity is said to be the "neighbouring cluster" of x_i because it is the next best fit cluster for point x_i.

$$b(x_i) = \min_{C \neq A} d(x_i, C) \tag{4.4}$$

We now define a silhouette:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max\{a(x_i), b(x_i)\}} \tag{4.5}$$

From Equation 4.5 we can easily see that:

$$1 \leq s(x_i) \leq 1 \tag{4.6}$$

4.3.2 Principal Components Analysis

Our monthly signature data consists of twelve variables, one for each month of the year. It would be useful if we could effectively reduce this dimensionality from twelve to something smaller, and thus identify which months are the most important for distinguishing between the different footfall profiles obtained using the K-Means clustering technique. Furthermore, if it is possible to reduce dimensionality from twelve to two, we could examine the clusters for separability on a two dimensional plot. A popular technique capable of delivering these potential benefits is Principal Components Analysis (PCA). However, until we have presented the results of our clustering experiments, and inspected their quality, the usefulness of PCA for our purposes is somewhat speculative: we need to ensure that we have clear and distinct signature profiles in the first place, before we consider attempting to apply further analysis. For this reason we delay a fuller description of the PCA methodology until Section 5.3.

5 Results

5.1 K-Means Signatures for Monthly Profiles

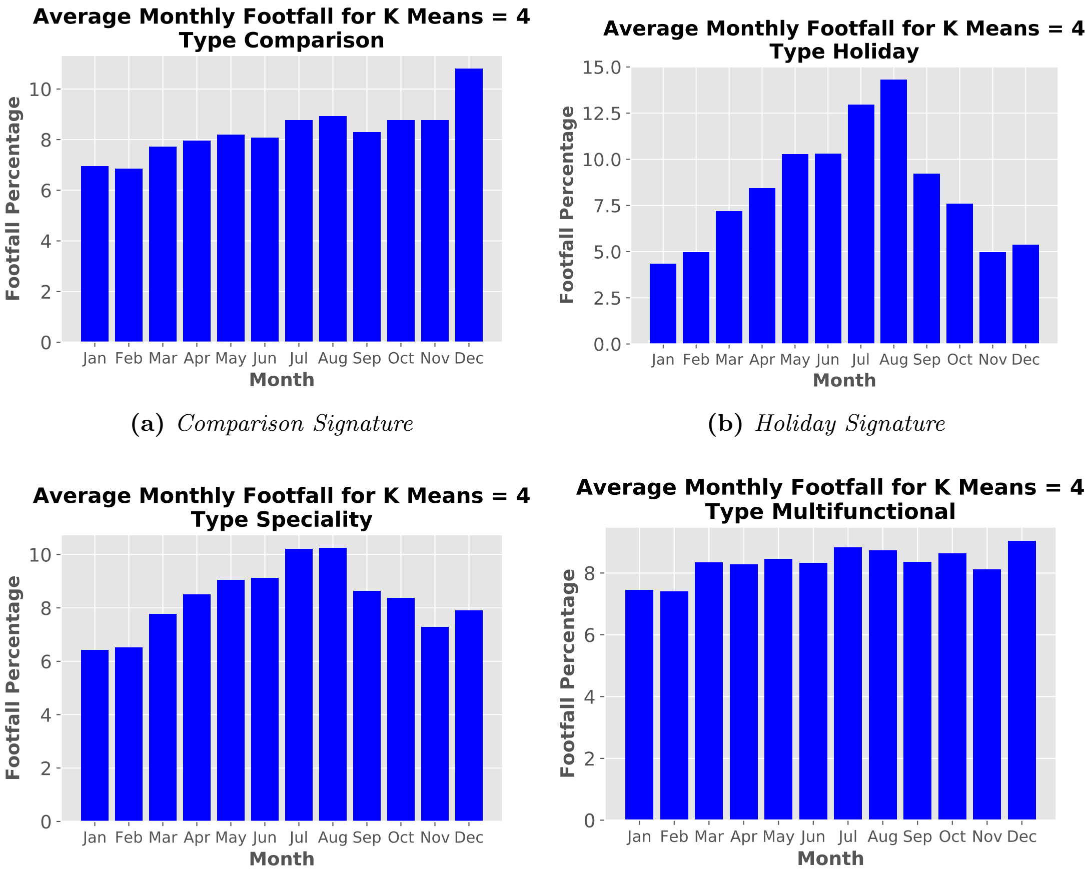
Standardized monthly footfall profiles are produced for a hundred and forty four UK retail centres, as described in Section 4.1. Each profile distributes the 100 % annual footfall over the constituent months and is stored in a .csv file. K-Means clustering is then applied separately to the two sets of a hundred and fourteen profiles – averaged (all counters) and main counters. From the HS2020 study we are expecting four distinct signatures to emerge. However, we do not make advance assumptions and experiment with K-Means clustering for values of K between 2 and 5 inclusive. We are pleased to confirm that our studies clearly confirm the existence of the four signatures proposed in HS2020, namely: comparison, holiday, speciality and multifunctional. These four signature types are illustrated in Figure 5.1.

Figure 5.2 shows Silhouette coefficients for our K-Means experiments with $K = 1 \ldots 4$. We have not displayed results for $K > 4$ because additional profiles obtained using $K = 5, 6$ etc. were similar to the signatures already obtained with $K = 4$.

Each cluster in the diagrams consists of a sorted histogram of silhouette values for the retail centres. The vertical height of each cluster on the page represents its size (i.e., the number of retail centres classified as “comparison”, or “holiday” etc.), and its width dimension shows the individual silhouette values for the retail centres that belong to that cluster. The vertical red dashed line on each diagram denotes the average silhouette value for all the retail centres (also recorded in the rectangular box at the bottom of each diagram).

It is interesting to note that the comparison and holiday signatures dominate, and appear when $K = 2$. These can be easily identified by examining the two centroids for $K = 2$. It is clear though that the cluster we have identified as “comparison” can be described as a “super cluster” (see Section 4.3.1), given that it accounts for the majority of retail centres. Under this assumption, “holiday towns” would appear to be “outliers”. When $K = 3$ the speciality signature appears, and all the signatures are present for $K = 4$. It is very noticeable that the number of retail centres in the “holiday” group remains a rather small proportion of the whole, for all values of K tried. Although confidentiality issues prevent the publication of the signature classifications for individual centres, it is noticeable on examination of these, that centres that most closely resemble the centroid for the comparison signature tend to be the larger city and town centres.

Observing the “standard year” profiles for individual retail centres, it is clear that some centres match one of the four standard templates very closely indeed, whilst others produce monthly profiles that are much more difficult to assign to one of the four classes. As well as sharing characteristics with more than one type, some centres demonstrate patterns unique to themselves. To visualize aspects of the great variability between centres, we compute euclidean distance values for every retail centre from each of the centroids from $K = 4$, to give measures of how closely each centre resembles the template signatures. The subplots in Figure 5.3 illustrate scatter diagrams of the distances from the four centroids all our hundred and fourteen retail centres. The resulting graphical



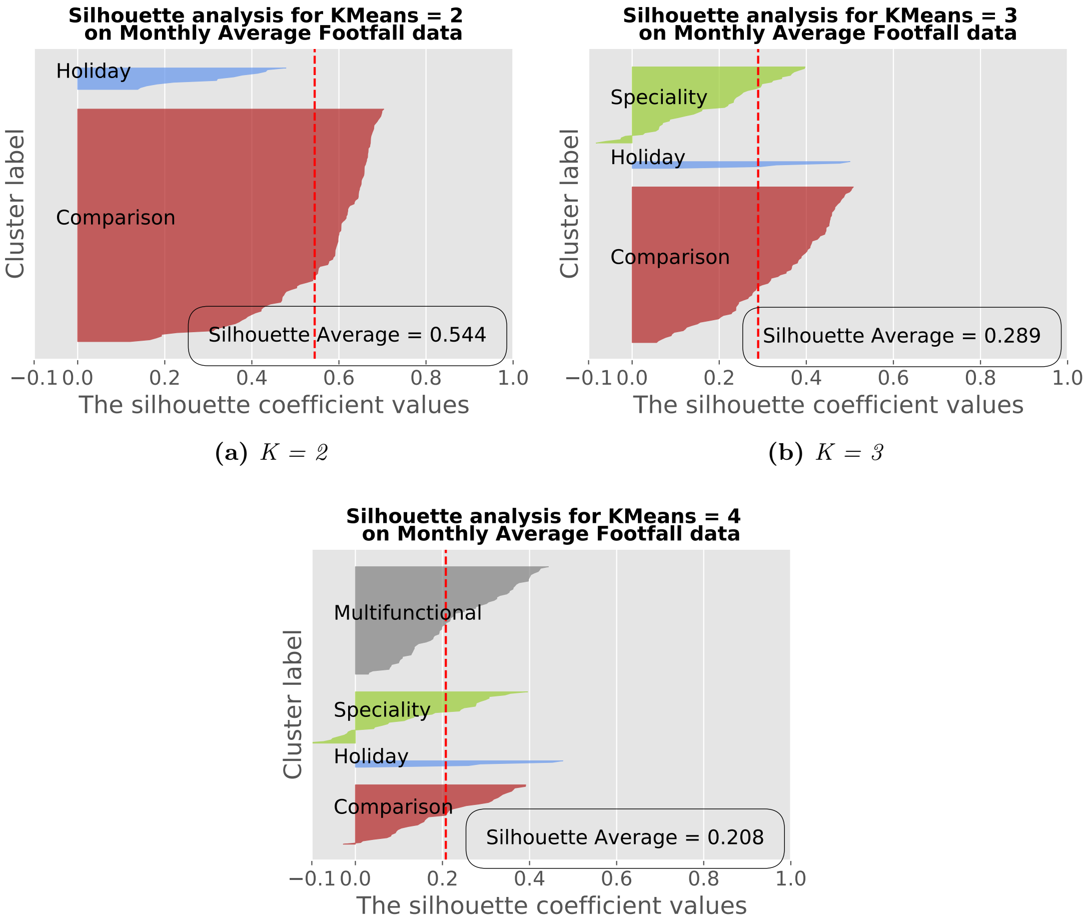
(a) *Comparison Signature*

(b) *Holiday Signature*

(c) *Speciality Signature*

(d) *Multifunctional Signature*

Figure 5.1: *The four distinct signatures that emerge from our clustering study. The histograms pictured here were obtained by running K-Means for $K = 4$ on the hundred and thirty five centres using the data for average footfall from all counters in each retail centre. The pictured signatures are the centroids.*



(a) *K = 2*

(b) *K = 3*

(c) *K = 4*

Figure 5.2: *Silhouette Values for signature classes*

plots indeed show clearly that a simple “all or nothing” classification does not tell the whole story. One particularly interesting observation is that the signature profiles for the holiday towns are well separated from all the other retail centre profiles, and thus emphasizes the very distinct nature of footfall patterns in holiday towns, supporting our findings from our K Means experiments (especially with $K = 2$). The other clusters in Figure 5.3, (comparison, speciality and multifunctional), all show some degree of overlap with each other.

5.2 Signature Versus Total Footfall

For our next task we carry out an analysis of variance (ANOVA) to test whether there is any relationship between signature type and total footfall for particular retail centres, i.e., to ascertain if some types of town are busier than others. For our total annual footfall counts, it is important that total footfall in one retail centre can be directly compared with that of another retail centre. For this purpose averaging the footfall over all counters in a retail centre is not appropriate. Instead we identify the counter in the busiest part of a retail centre and use the annual totals from that counter. For each retail centre, the annual totals from the busiest counters are then averaged to compute and average annual footfall figure for each retail centre in our study.

ANOVA results demonstrate a significant difference between mean annual footfall for retail centres depending on their signatures, as can be seen in Figure 5.4. This figure illustrates the results of a multi-comparison Tukey test, which we use following our ANOVA test to find out exactly which pairs of values show that significant differences between them. In the figure, the points show the means, and the lines indicate the 95 % confidence intervals of the expected values of the means. This analysis clearly demonstrates that comparison towns are, on average, significantly busier than holiday and multifunctional retail centres. Figure 5.5 uses box plots to illustrate the range of values for annual footfall within the individual signature classes for the raw data, and Figure 5.6 defines the graphical features for the box plots.

5.3 Principal Components Analysis to Identify Distinguishing Features in the Monthly Signatures

Now that we have completed the clustering experiments for the monthly data, and verified the four footfall signature profiles as: comparison, holiday, speciality and multifunctional, it would be useful if we could identify which months are the most important for distinguishing between the four different profiles: on visual examination of the four centroids generated by K-Means in Figure 5.1, with k = 4, a December peak is clearly a key feature of the comparison signatures, whilst July and August peaks seem to typify holiday towns, and to a lesser extent, speciality centres. Principal Component Analysis (PCA), as we discussed earlier, is a statistical procedure that we can use to provide some scientific support to the identification key months for distinguishing between the four signatures. PCA was developed by Karl Pearson [5] in 1901, but it is Harold Hotelling [6] who is responsible for giving it its name in the 1930s. Simply speaking, PCA attempts to reduce the number of variables by essentially transforming them into new variables, called the principal components. The technique works on the assumption that some of the original variables may be correlated with each other. For example, we can see that high footfall in July tends to be accompanied by high footfall in August in our retail centres. A familiar technique for reducing dimensions from two to one, is computing a line of regression. PCA extends this approach to multiple dimensions by computing a set of lines, all at right angles to each other (orthogonal), and then projecting the original variables onto these lines in the form of linear equations, for example:

$$Principal\ Component\ ij = L_1^i F_{Jan}^j + L_2^i F_{Feb}^j + \cdots + L_{12}^i F_{Dec}^j \tag{5.1}$$

where the i^{th} principal component can be computed for any particular retail centre j by evaluating the sum of the products of the weights, L^i (called Loadings in PCA), and the corresponding signature footfall value, F_{month}, for that month in retail centre j. The number of principal components is less than, or equal to, the number of original variables, and the first principal component (PC1) accounts for as much of the variability in the data as possible. The second (PC2) and subsequent principal components (PC3, PC4 etc.) then account for ever-decreasing amounts of the remaining variability. Total variability = 1 (or 100 %).

Before we begin our PCA, we transform the monthly signature for each retail centres to a percentage of its annual total, so that the monthly totals for footfall in a centre add up to 100 %. This is an important step, as we are interested in the spread of footfall throughout the months of the year, and not its total amount. Following this standardisation process, we next compute twelve principal components, to coincide with the number of variables (12 months) using the PCA library provided by Scikit-learn. Figure 5.7 indicates the cumulative percentage of variability explained by the twelve principal components. As can be observed, PC1 and PC2 explain almost 80 % of the variation in the retail centre signatures. Table 5.1 shows the loadings (or weights) for PC1 and PC2. In the Table, the loadings with the highest magnitude values are the most important. Thus, we can see that for PC1 July, August and December have the highest magnitude loadings, at 0.397, 0.531 and -0.463 respectively. The positive or negative sign shows whether loading values are directly or inversely correlated. Thus, as expected from our visual observations of the four signatures, high peaks in summer footfall in July and August for holiday towns are usually associated with low footfall in the winter months, particularly December. PC2 emphasizes March, August and December. December is clearly the peak month for comparison shopping centres. and August the peak summer month in holiday and speciality centres. The predominance of March is something of a surprise. Looking at the relatively high value for footfall in March for the multifunctional signature in Figure 5.1, we hypothesise that this could possibly be a key month for identifying multifunctional centres.

Thus the two PC equations are as follows:

$$PC1_j = (-0.291 \times F_{Jan}) + (-0.223 \times F_{Feb}) + \cdots + (-0.460 \times F_{Dec}) \tag{5.2}$$

Table 5.1: *Loadings for Principle Components Analysis*

Principle Component	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	-0.277	-0.218	-0.093	0.068	0.183	0.201	0.397	0.531	0.108	-0.099	-0.336	-0.463
2	-0.257	-0.322	-0.383	-0.255	-0.125	-0.063	0.083	0.379	0.057	0.059	0.190	0.637

$$PC2_j = (-0.260 \times F_{Jan}) + (-0.267 \times F_{Feb}) + \cdots + (+0.495 \times F_{Dec}) \qquad (5.3)$$

Finally, in Figure 5.8 we plot values of PC1 and PC2 for all of our sample one hundred and twelve retail centres, and label them with the signature classification obtained from our K-Means study. From Figure 5.8 we can see that the retail centres separate nicely into four distinct clusters, which provides independent supporting evidence for our signature classification scheme (i.e., the four distinct signatures).

5.4 Daily Signatures

Our K-$Means$ study on profiles for days of the week distinguish two main signatures, as illustrated in Figure 5.9. Clearly the features that distinguish the two profiles are, 1) the right hand figure has a more pronounced peak on a Saturday than the left hand figure, and 2) the left hand figure a much smaller level of footfall on a Sunday than the right hand figure.

To investigate the distribution of the two daily signature types amongst the monthly classification, we carry our a χ^2 contingency test. If we assume that the the daily profiles are distributed evenly amongst the four monthly classes, it is simple to compute an "expected" distribution of daily profile types within the monthly classes. These expected values are shown in brackets in Table 5.2). If we then compare the actual values for the signature types, we can soon observe that the actual pattern is rather different from the expected pattern. In fact the χ^2 contingency test tells us that the actual distribution differs significantly from the expected distribution, to be precise the χ^2 statistic is = 11.08892, and the chance of this value being as high as this by chance is 0.01, which is a probability of 1 %.

Table 5.2: χ^2 *Contingency Table for retail centres' monthly signature classification (columns) versus daily signature classification (rows). The expected frequencies are given in brackets following the actual frequencies*

	Comparison	Holiday	Speciality	Multifunctional	Total
1) Small Saturday Peak	10 (13.99)	0 (1.84)	9 (12.15)	34 (25.03)	53
2) Large Saturday Peak	28 (24.01)	5 (3.16)	24 (20.85)	34 (42.97)	91
Total	38	5	33	68	144

The most notable deviations between the actual and expected values are evident for holiday and comparison towns, where Type 2 predominates. On the other hand, multifunctional has more Type 1 towns than expected, while speciality towns show a split very much consistent with the expected values. Figure 5.10 illustrate the spread of raw data for the daily signature types, clearly showing Type 2 is busier than Type 1. Figure 5.11 shows how the daily signature types are distributed amongst the four monthly types. Note that there are no Type 1 classifications for holiday towns.

Hourly Signatures Finally, we carry out a study to identify profiles for the twenty four hours of the day, and discover just two variations, where one is rather busier in the afternoon, evening and at night than the other, as can seen in Figure 5.12.Note that we have hourly data from more retail centres than is the case for monthly data, as it is not necessary to remove partial year data when we are measuring hours of the day.

6 Conclusions and Next Steps

In this report we have demonstrated the following:

- Four clear monthly footfall signatures exist, distinguishing different types of retail centre we have named comparison, holiday, speciality and multifunctional.
- Some centres have a clearer signature than others, in terms of how closely their footfall profiles match one of the four template signatures: all centres can be classified by their closest match, but some matches are better than others.
- The majority of retail centres that have been classified as comparison types are the larger city and town centres.
- Comparison centres are the busiest - they have the highest footfall.
- Holiday towns are the most distinctive, and have footfall profiles that form clusters clearly separate from all other retail centres.
- The months of December, July, August and March are the ones that vary most between the four different monthly signature types.
- Larger towns tend to be busier at weekends.

We have demonstrated that the four distinct monthly signatures exist. However, the crucial question to answer next is whether this classification can help retail centre stakeholders enhance the experience of their customers and make the centres more successful. In other words:

- can knowledge of the type of retail centre help inform its stakeholders how to best improve the performance of the centre?

Additionally, we will be looking at trends in footfall, to see how centres change and evolve over a period of time, in terms of their footfall profile and whether changing profiles are correlated with changes in performance.

The project will also move on to investigate other features of retail centres, including their locations (for example, north versus south), catchment (size of local population), retail offer (i.e., numbers of bakers and coffee shops, chemists, clothing shops, department stores etc.). We will examine how these and other factors correlate with a centre's footfall signature and also its retail performance. Weekly, daily and hourly footfall patterns will need to be examined, especially with respect to seasonal variation. We will investigate the 25 priority factors that can be changed/influenced by High Street stakeholders as identified in [11]. Working with our partners in the seven towns will ensure that our research findings can be used to the benefit of the retail stakeholders and thus have a real impact on retail centres and communities.

Acknowledgements

This work is supported by Innovate UK, Grant Number 509847.

References

[1] The hdf group. `https://support.hdfgroup.org/HDF5/`. Accessed: 2016-11-30.

[2] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, Aug 1995.

[3] E. Forgy. Cluster analysis of multivariate data: Efficiency versus interpretability of classification. *Biometrics*, 21(3):768–769, 1965.

[4] Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[5] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572, 1901.

[6] H. Hotelling. *Analysis of a Complex of Statistical Variables Into Principal Components*. Warwick & York, 1933.

[7] Springboard LTD. The performance of retail locations in the changing retail environment. `http://www.spring-board.info/updates/article/BRC-the-retailer`, July 2013. (Accessed: 2017-01-05).

[8] Chales Mercier. *A New Logic*. Open Court Publishing Company, Chicago, 1912. `https://babel.hathitrust.org/cgi/pt?id=nyp.33433089906436;view=1up;seq=184`.

[9] Steve Millington, Nikos Ntounis, Cathy Parker, and Simon Quin. Multifunctional centres: a sustainable role for town and city centres. `http://placemanagement.org/research`, 2015. (Select: Multifunctionality: a sustainable role for centres. Accessed: 2017-01-05).

[10] Cathy Parker, Nikos Ntounis, Simon Quin, and Steve Millington. Identifying factors that influence vitality and viability. http://www.placemanagement.org/media/57742/HSUK2020-End-of-Project-Reportcompressed.pdf.

[11] Cathy Parker, Nikos Ntounis, Simon Quin, and Steve Millington. Identifying factors that influence vitality and viability.

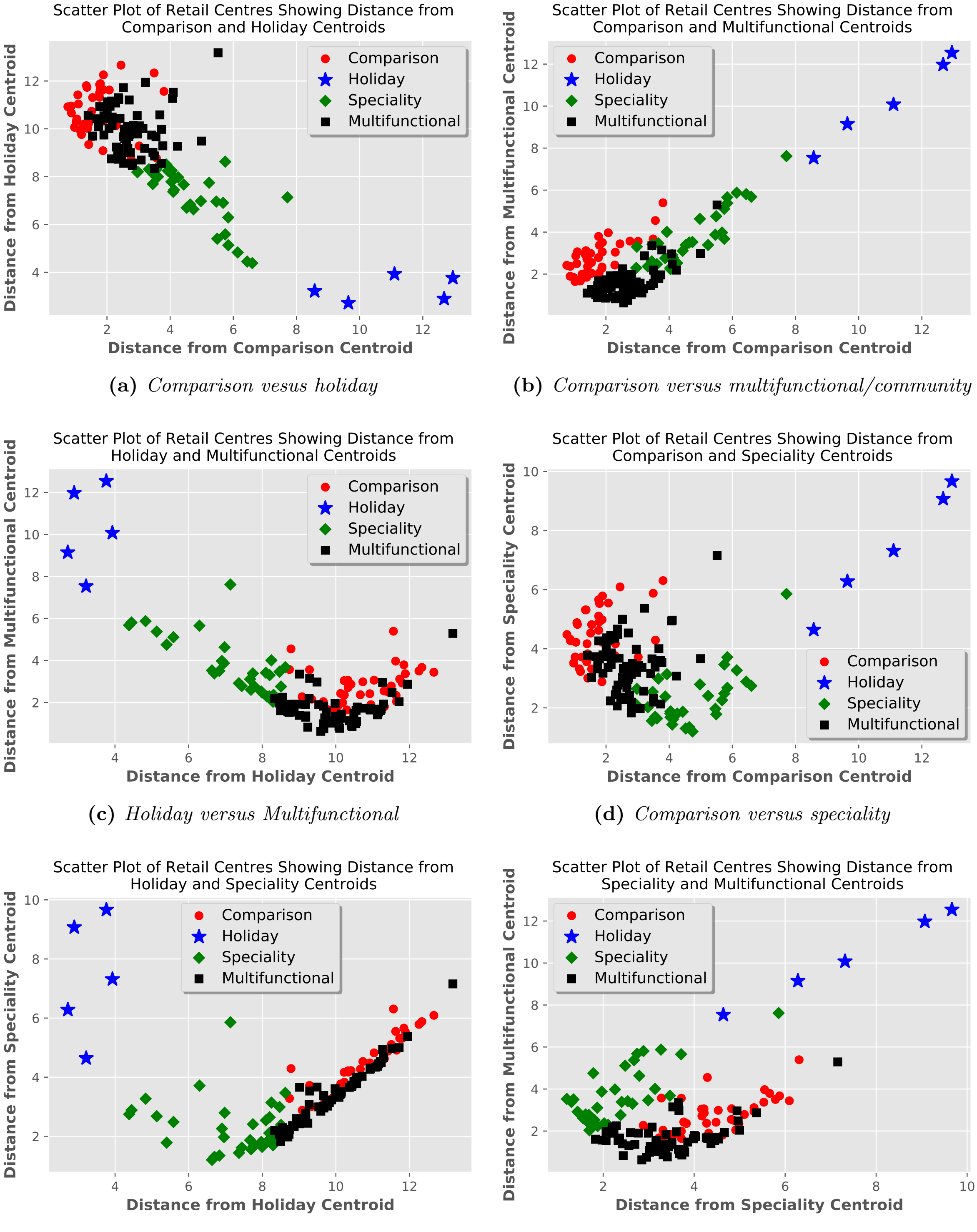
[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[13] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.

[14] A.D. Singleton, L. Dolega, D. Riddlesden, and P.A. Longley. Measuring the spatial vulnerability of retail centres to online consumption through a framework of e-resilience. *Geoforum*, 69:5 – 18, 2016.

[15] Digital High Street Advisory Board (Chair:John C Walden). Digital high street 2020 report. http://thegreatbritishhighstreet.co.uk/pdf/Digital_High_Street_Report/The-Digital-High-Street-Report-2020.pdf, March 2015. (Accessed: 2016-11-30).

[16] Neil Wrigley, Dionysia Lambiri, G Astbury, L Dolega, C Hart, C Reeves, M Thurstain-Goodwin, and SM Wood. British high streets: from crisis to recovery? a comprehensive review of the evidence. 2015.



(a) *Comparison vesus holiday*

(b) *Comparison versus multifunctional/community*

(c) *Holiday versus Multifunctional*

(d) *Comparison versus speciality*

(e) *Holiday versus speciality*

(f) *Specialaity versus multifunctional*

Figure 5.3: *Scatter plots to show distance of each retail centre from various centroids*

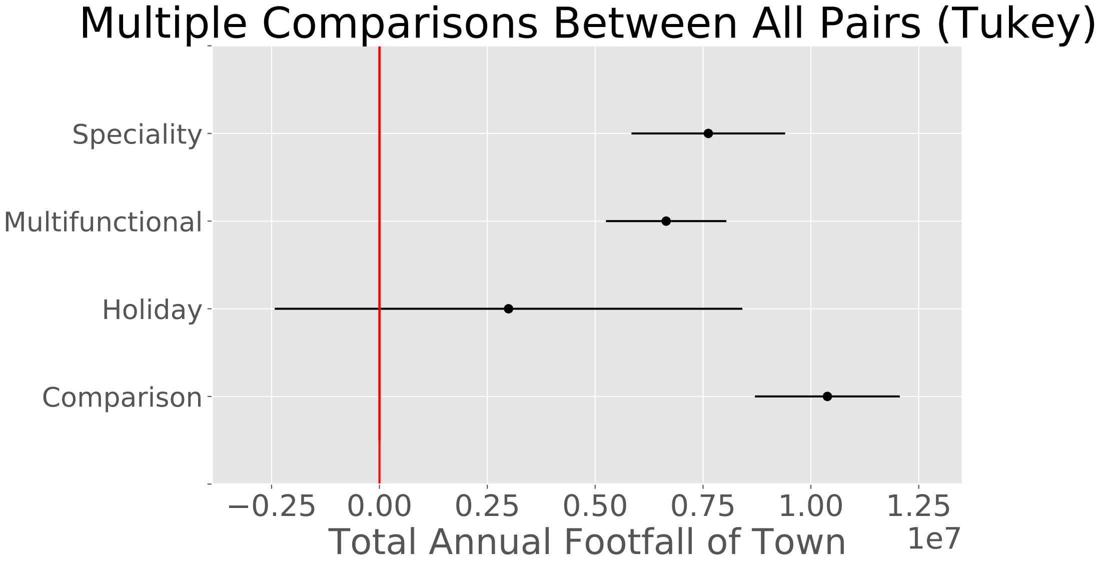


Figure 5.4: *Tukey test showing 95 % confidence intervals: Comparison centres are significantly busier than holiday, speciality or multifunctional centres.*

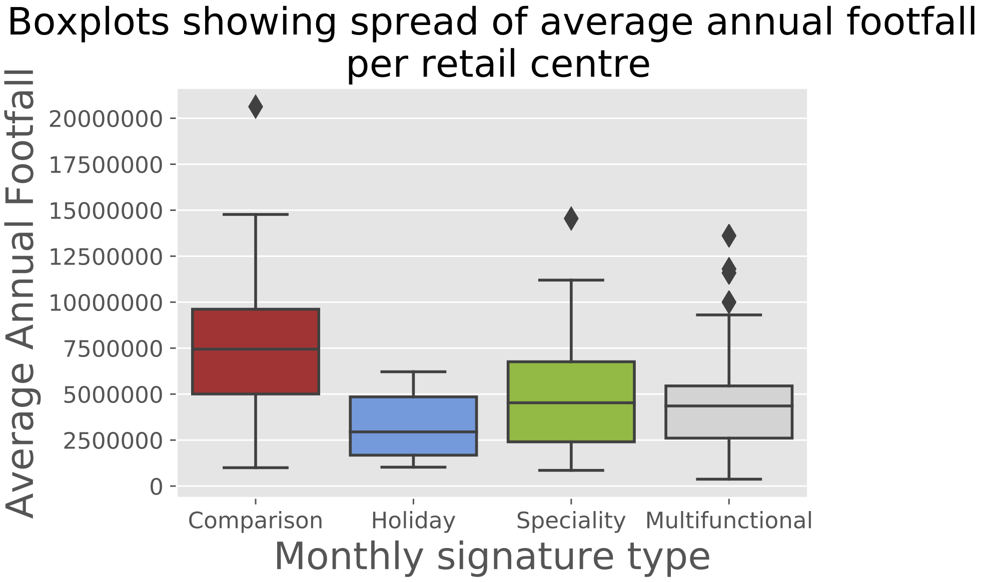


Figure 5.5: *Box plots illustrating the distribution of raw data for retail centres with the four signature types.*

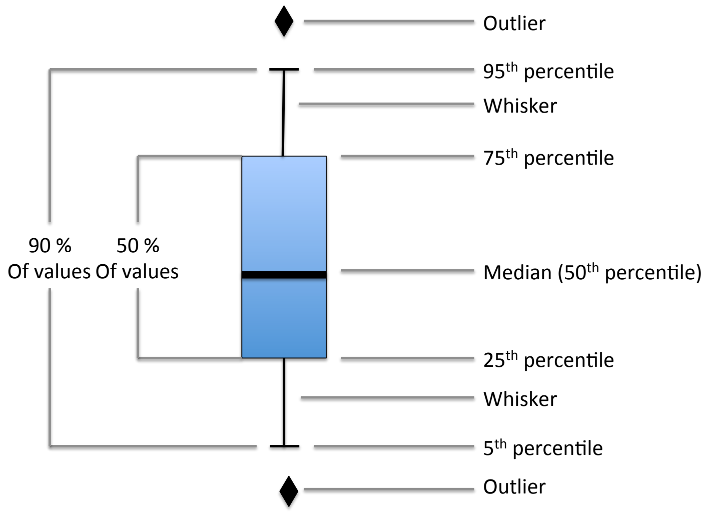


Figure 5.6: *Box plots components.*

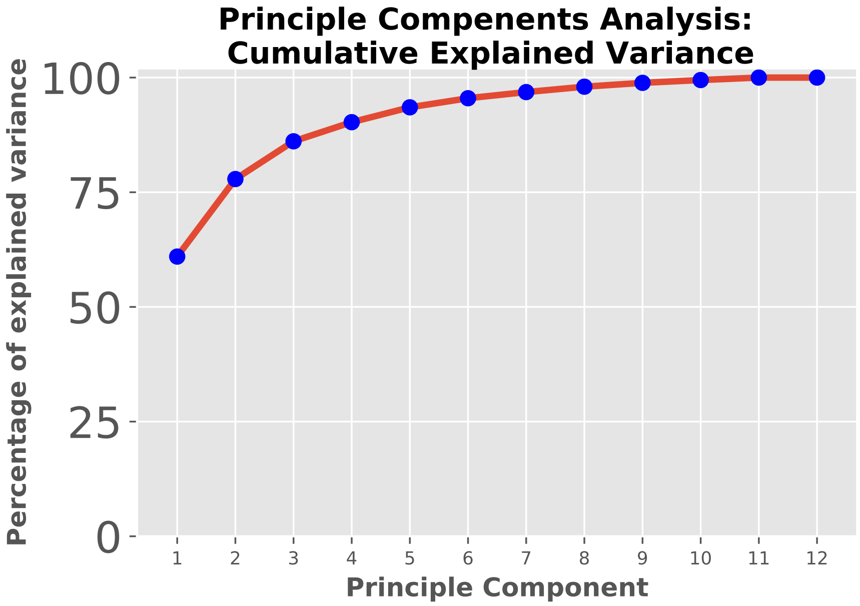


Figure 5.7

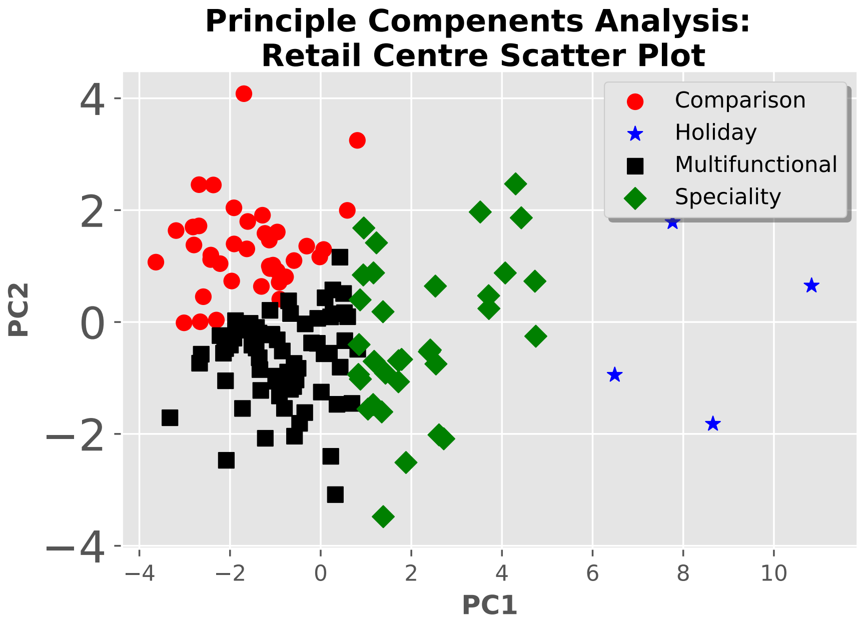
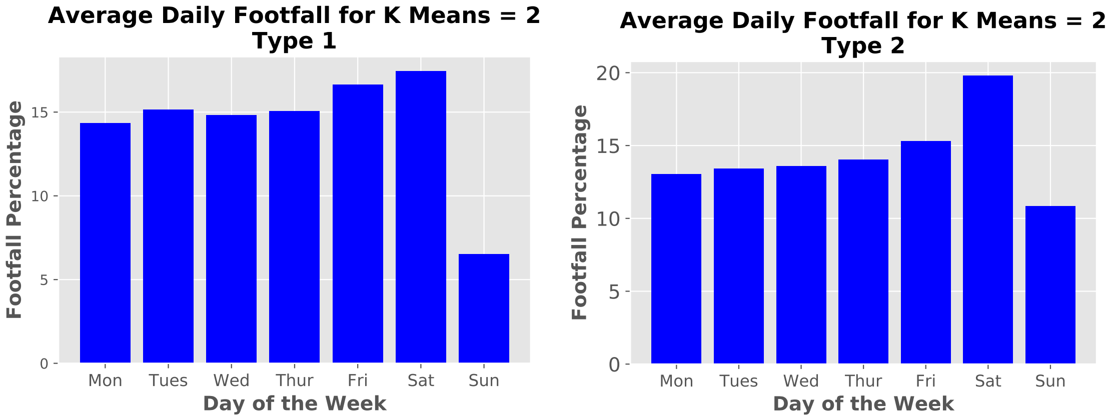


Figure 5.8



(a) *Small Saturday Peak*

(b) *Large Saturday Peak*

Figure 5.9: *The two distinct signatures that emerge from our K Means clustering study on daily footfall signatures for our 135 retail centres. The pictured signatures are the centroids.*

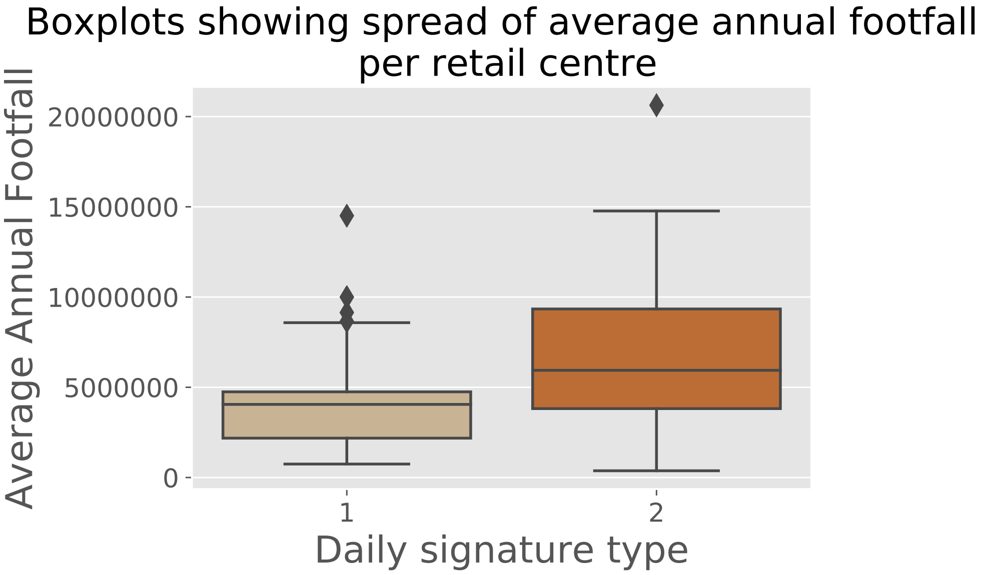


Figure 5.10

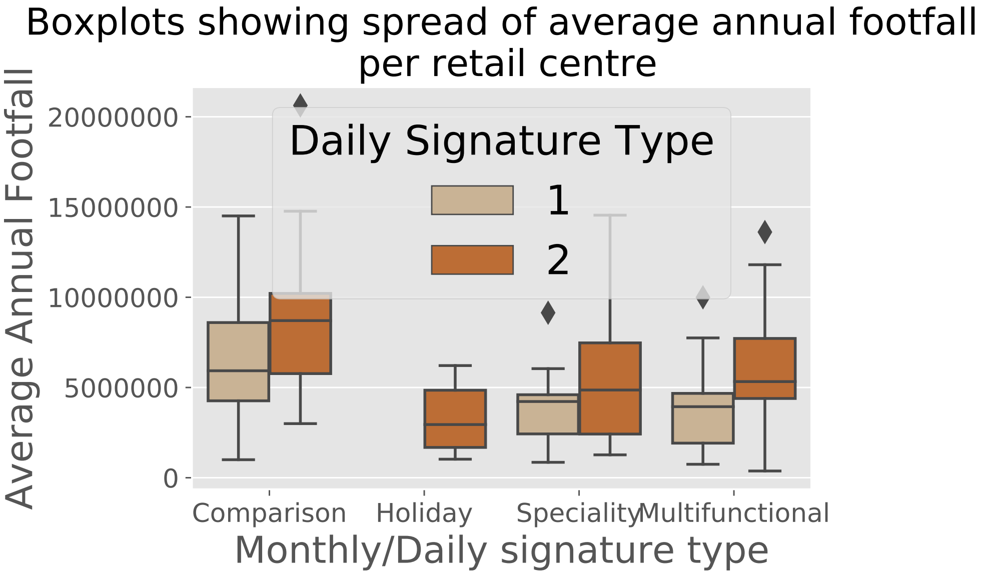
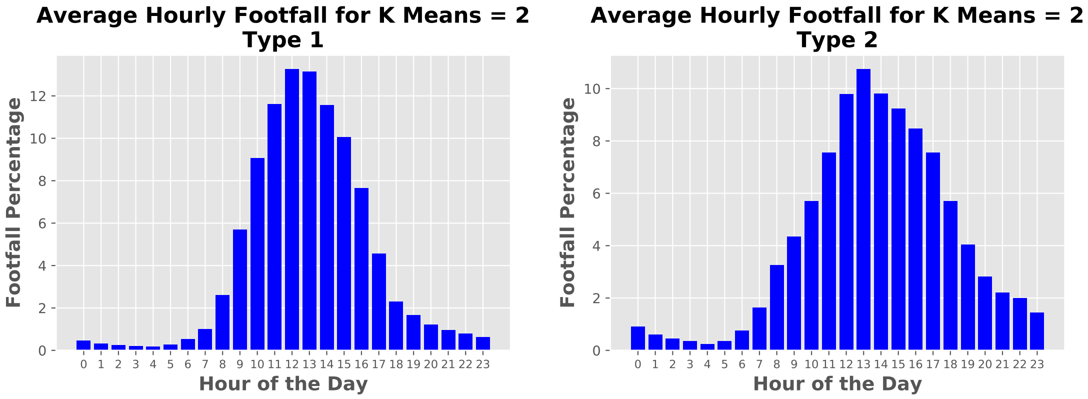


Figure 5.11



(a) *Quiet at night*

(b) *Busier at night*

Figure 5.12: *The two signatures that emerge from our K Means clustering study on daily footfall signatures for 154 retail centres. The pictured signatures are the centroids.*