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

Technical Report 1 (Version 1)*

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

February 16, 2017

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). The footfall data was supplied by Springboard LTD^{TM}and consists of hourly records broadcast from several hundred counters located in traditional retail centres throughout the UK. Before using the data, we checked its completeness for every counter by identifying any missing hourly data. The computed completeness figure at 96.38 % proved the counters to be very reliable. The identified retail centre types are: *comparison*, *holiday*, *convenience/community* and *speciality*. 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 convenience/community 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

*Updated versions of this report will be issued as appropriate. For example, when more data becomes available, and to cover daily and hourly footfall signatures.

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

pronounced as they are in pure comparison and holiday centres. Beginning with a vast amount of raw hourly data gathered from each counter (8760 readings per year), monthly totals per retail centre were computed, and then further processing was carried out to produce monthly footfall profiles for a "standard" year: one for each retail centre (provided they could supply at least a full year of data). Using K Means clustering techniques we are able to firstly, produce four convincing signature templates to closely match those proposed in the High Street UK2020 study and secondly, classify each of the aforementioned retail centres as one of the four signature types. During this process, we use *Silhouette Analysis* 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 found that setting $K = 2$ produces signatures for comparison and holiday, $K = 3$ produces comparison, holiday and speciality, and $K = 4$ produces all four of the expected signatures types. Furthermore, the Silhouette scores prove to be highest when $K = 2$ indicating that comparison and holiday types produce the strongest profiles. Silhouette scores then drop a little between $K = 2$ and $K = 3$ and then little further for $K = 4$. Setting $K = 5$ picks up some spurious looking peaks that seem to coincide with inconsistent behaviour of the counters, such as footfall missing for whole months of the year (corresponding to periods, for example, that counters were switched off and on again). For this reason we curtailed 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 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 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 form a cluster well separated from all the other retail centre profiles, and thus emphasizes the very strong profile and distinct nature of footfall patterns in holiday towns. The other clusters (comparison, speciality and convenience/community) all showed some degree of overlap with each other. However, by applying Principal Components Analysis (PCA) to the monthly profiles for our retail centres, we were able to separate all the clusters and produce a two dimensional plot without any overlaps at all, clearly demonstrating that the clusters for comparison, speciality, convenience/community 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 this instance a 2D plot of retail centres.

In addition to the "'cleaning up" of the classification clusters, PCA supported 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. March was also identified as an important month, although its usefulness in distinguishing between signature types is less intuitive and warrants further investigation. A final analysis reported here attempts to correlate 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 seems to consist mostly of large city and town centres.

The work outlined in this report demonstrates that four distinct monthly 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.

1 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^{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, speciality centres, holiday towns, and convenience/community centres [9]. Comparison shopping centres are typified by a peak in footfall in the month of December, presumably coinciding with spending coming up to Christmas. On the other hand, convenience/community 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.

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 2 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 3 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 4, where we also outline the next steps planned for our research.

2 Methodology

In this report we analyse monthly footfall counts on a large set of data in an attempt to verify the existence of the four distinct signatures 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 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 2.1), and then we go on to explain in Subsection 2.2 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 2.3), and finally we briefly introduce Principal Components Analysis, which we will return to in Section 3.3.

2.1 Data Files

Footfall data provided by Springboard UK Limited$^{\text{TM}}$consists of hourly footfall counts from 421 counters located in 151 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. In addition, for this study we have excluded retail centres located in and around London which, as the UK's largest conurbation, has a unique retail landscape. The historical data was provided by Springboard as 11 comma separated values (csv) files consisting of records in the format seen in Table 2.1. 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 2.1: *Format of Springboard raw files*

Data Preparation For this study we examine monthly 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 421 counters. For each counter 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 421 cameras is 96.38, 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 separate monthly totals for each retail centre. The second step is to compute a "representative year" for each retail centre, consisting of mean footfall values for each 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 in a similar way.

We compute our monthly footfall totals from the original hourly data in two different ways:

1. Using an average hourly count for ALL counters.
2. Determining and using one (Main Counter) only for each retail centre.

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

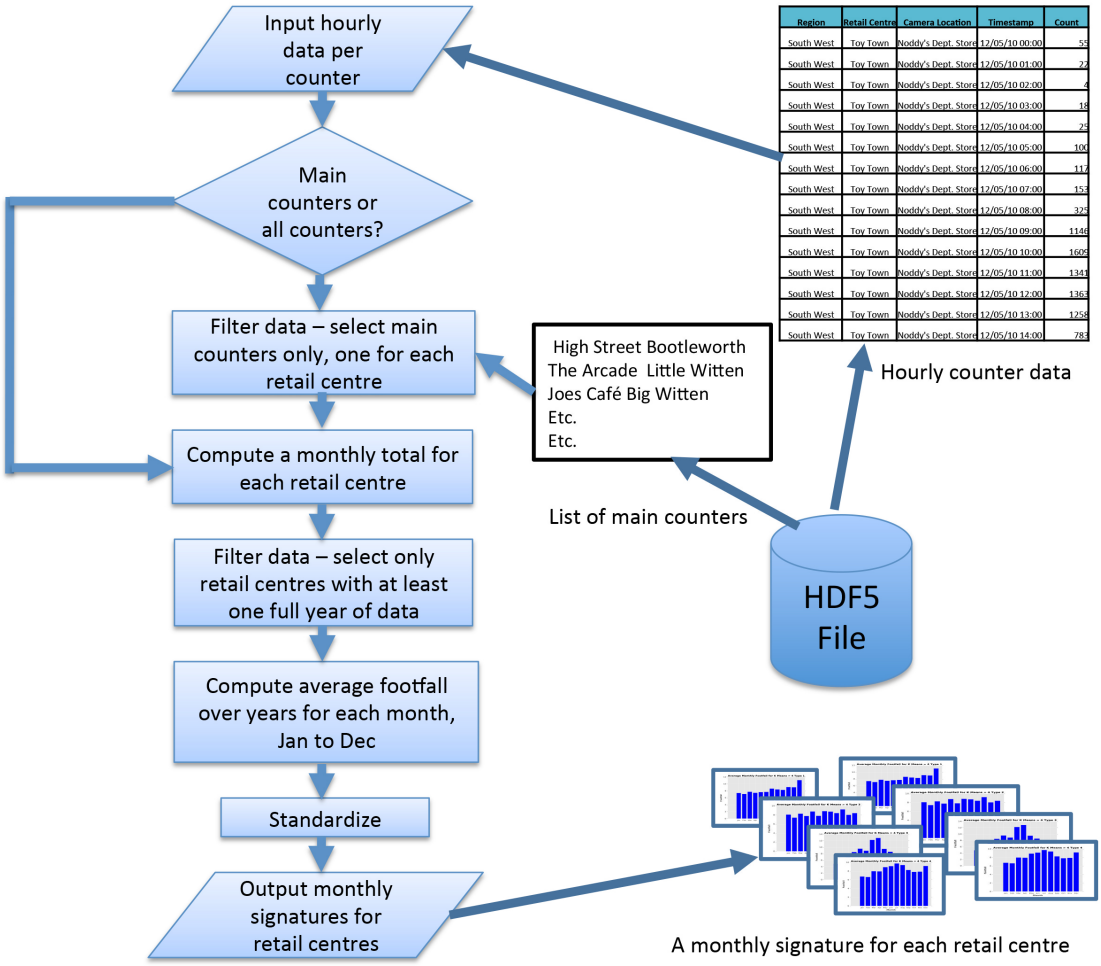


Figure 2.1: *Flowchart to show the data preparation required for our clustering experiments*

For method (1) for each retail centre, the footfall count is averaged over all the counters in that retail centre. For method (2), we select the counter recording the highest average annual footfall to be our "main counter" for each retail centre, and use the total monthly footfall counts for that counter. The reason that we prepare two different sets of data is that we are unsure at this stage which approach will produce the more consistent results for the clustering experiments. Using all the counters (method 1) could prove less susceptible to issues with individual counters or temporary local road or pavement closures etc. On the other hand, the counters located in the busiest places should give more reliable figures. 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 %. Details of our data preparation can be seen in the flowchart in Figure 2.1. For both method 1) and method 2), for any given year, we exclude counters that have been active for only part of that year: i.e., they were newly installed part way through a year.

2.2 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.

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 $C = \{C_1, C_2, \ldots, C_K\}$ 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:

$$A(x_i) = arg\min_{C} \sum_{j=1}^{K} \sum_{x_p \in C_j} d(x_i, x_p) \tag{2.1}$$

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.

2.3 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. Ground truth labels are available, for example, if 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 this study 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. However, 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, 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."*

2.3.1 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. 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 3. 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]:

$$a(x_i) = \frac{1}{m_A - 1} \sum_{j=1}^{m_A} d(x_i, x_j), \quad \forall x_j \in A \ \ \textit{such that } x_j \neq x_i \tag{2.2}$$

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:

$$d(x_i, C) = \frac{1}{m_C} \sum_{j=1}^{m_C} d(x_i, x_j), \quad \forall x_j \in C \tag{2.3}$$

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.

$$b(x_i) = \min_{C \neq A} d(x_i, C) \tag{2.4}$$

We now define a silhouette:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max\{a(x_i), b(x_i)\}} \tag{2.5}$$

From Equation 2.5 we can easily see that:

$$1 \leq s(x_i) \leq 1 \tag{2.6}$$

2.3.2 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 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 3.3.

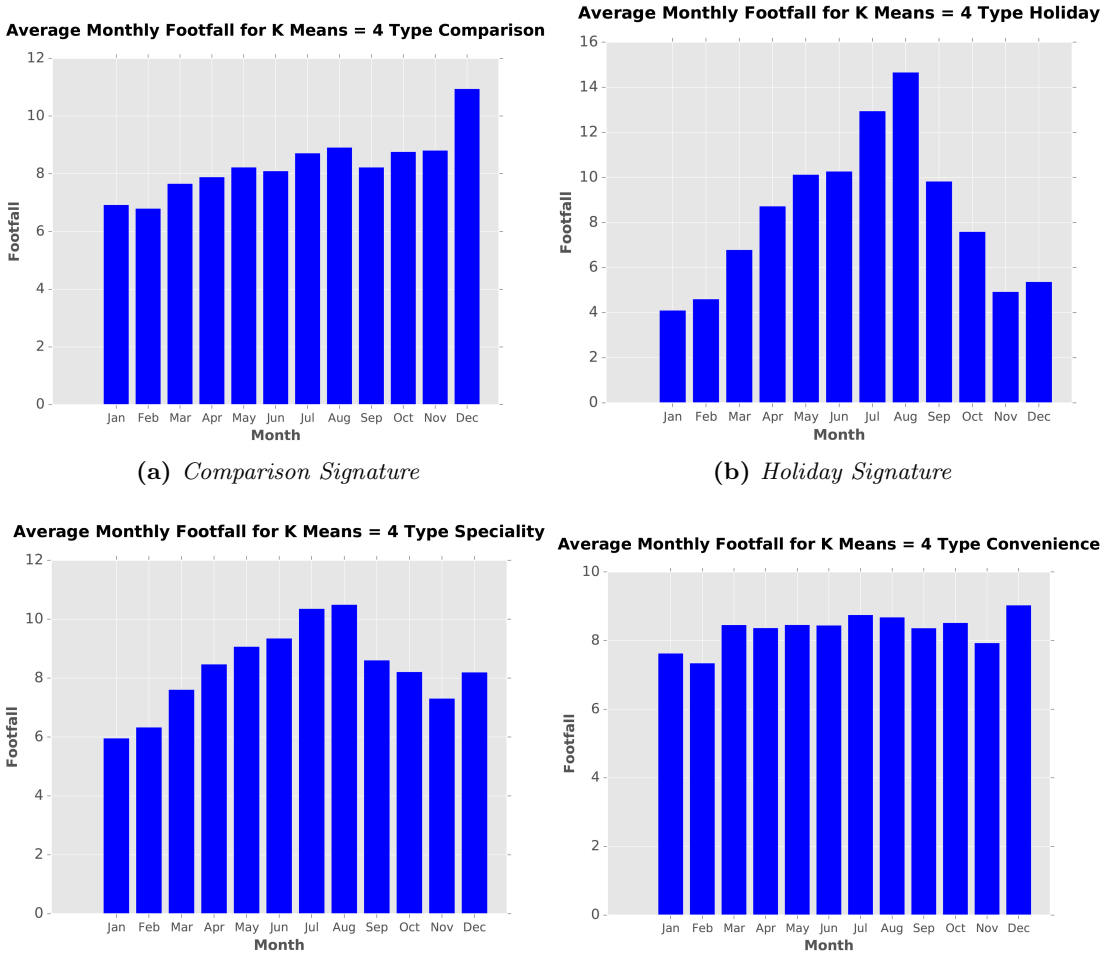
3 Results

3.1 K-Means Signatures

Standardized monthly footfall profiles are produced for ninety-nine UK retail centres, as described in Section 2.1. Each profile distributes the 100 % annual footfall over the constituent months and is stored in a .csv file. As mentioned previously, we produce two versions of this data: 1) averaging hourly for all counters in a retail centre, and 2) recording the hourly footfall for one main counter identified for each centre. K-Means clustering is then applied separately to the two sets of ninety-nine 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 convenience/community. These four signature types are illustrated in Figure 3.1.

Figure 3.2 shows Silhouette coefficients for our K-Means experiments with $K = 1 \ldots 4$. The left hand column gives the results for all counters and the right hand column for the main counter experiments. We have not displayed results for $K = 5$ because the fifth signature clearly picked up spurious defects in footfall, which were probably due to temporary closures of pedestrian areas due to utility works etc.

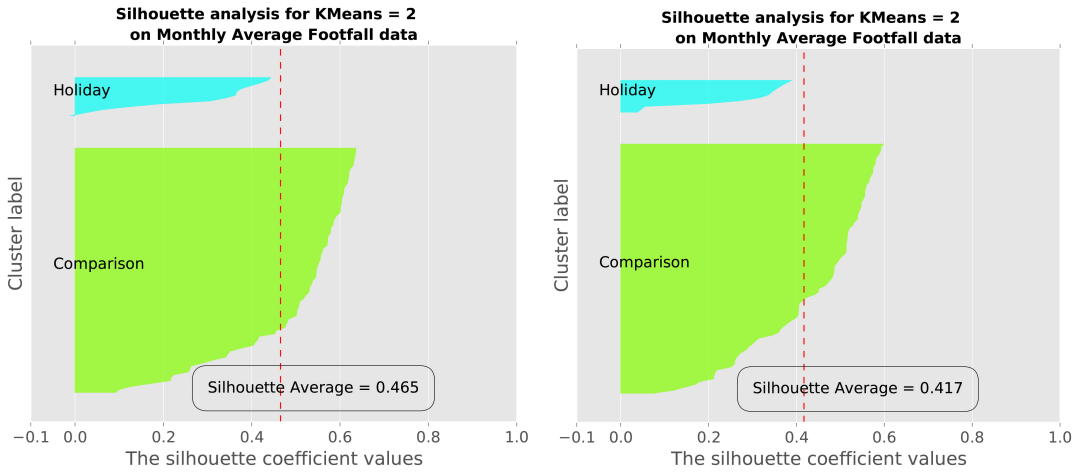
Each colour-coded cluster in the diagrams represent histograms of individual retail centres, showing their silhouette values. Thus the vertical height of each cluster on the page represents its size (i.e., the number of retail centres classified as “comparison”, or



(a) *Comparison Signature* **(b)** *Holiday Signature*

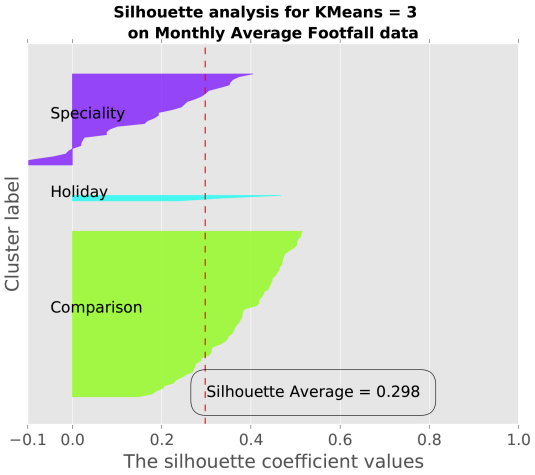
(c) *Speciality Signature* **(d)** *Convenience/community Signature*

Figure 3.1: *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 ninety-nine centres using the data for average footfall from all counters in each retail centre. The pictured signatures are the centroids.*

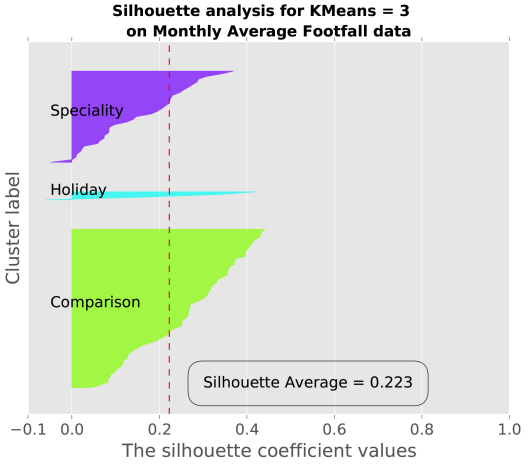


(a) *$K = 2$ all counters*

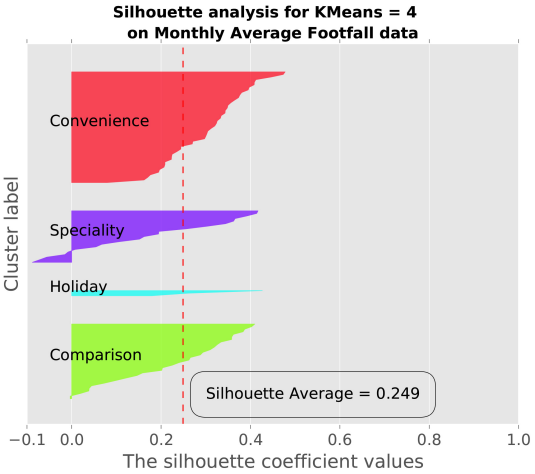
(b) *$K = 2$ main counters*



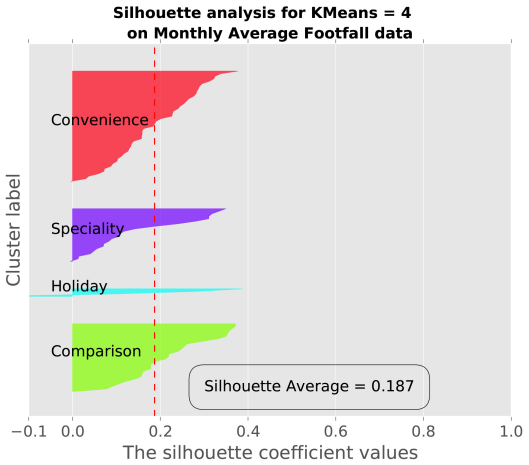
(c) *$K = 3$ all counters*



(d) *$K = 3$ main counters*



(e) *$K = 4$ all counters*



(f) *$K = 4$ main counters*

Figure 3.2: *Silhouette Values for signature classes*

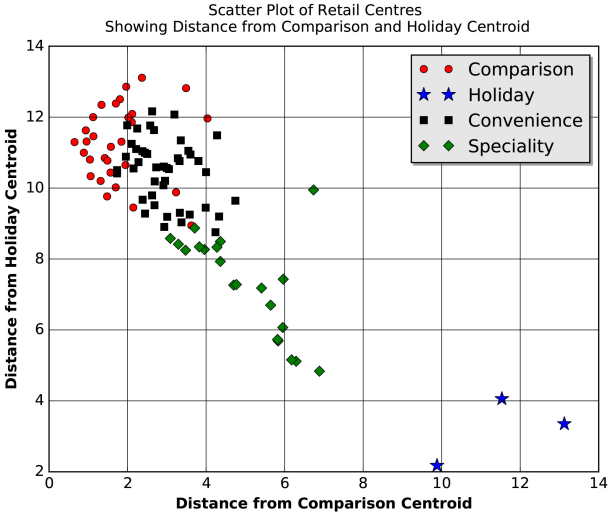
"holiday" etc.), and its width dimension shows the individual silhouette values for the retail centres that belong to that cluster. The vertical red dashed line on each diagram denotes the average silhouette value for all the retail centres (also recorded in the rectangular box at the bottom of each 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 2.3.1), 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 very small proportion of the whole, for all values of K tried. The average silhouette values for the different clustering experiments appear to slightly favour the results for all counters. However, the results for the main counters produce very similar profiles. 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.

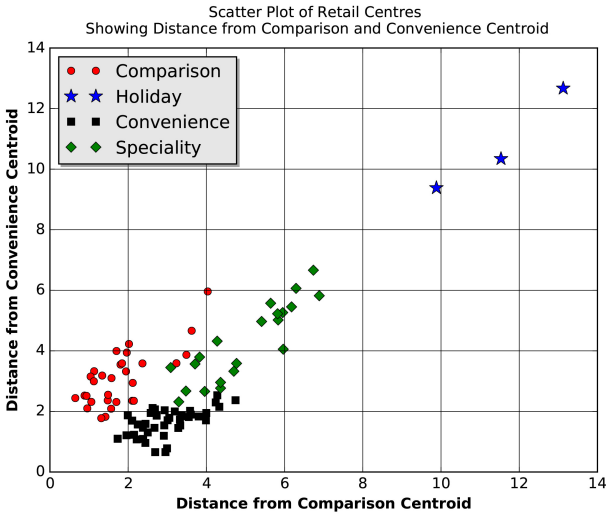
Observing the "standard year" profiles for individual retail centres, it is clear that some centres match one of the four standard templates very closely indeed, whilst others produce monthly profiles that are much more difficult to assign to one of the four classes. As well as sharing characteristics with more than one type, some centres demonstrate patterns unique to themselves. To visualize aspects of the great variability between centres, we compute distance values for every retail centre from each of the centroids form $K = 4$, to give measures of how closely each centre resembles the template signatures. The subplots in Figure 3.3 illustrate scatter diagrams of the distances from three of the four centroids all our ninety-nine retail centres. (We omit the centroid for speciality, because it is clearly a hybrid between comparison and holiday, given its dual peaks in the summer and December). 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 form a cluster well separated from all the other retail centre profiles, and thus emphasizes the very strong profile and 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 3.3, (comparison, speciality and convenience/community), all show some degree of overlap with each other.

3.2 Signature versus total footfall

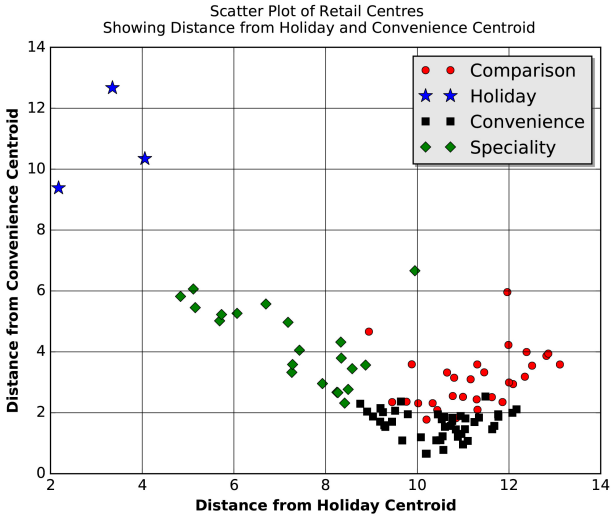
For our next task we carry out an analysis of variance (ANOVA) to test whether there is any relationship between signature type and total footfall for particular retail centres, i.e., to ascertain if some types of town are busier than others. As the ANOVA results demonstrate a significant difference between mean annual footfall for retail centres depending on their signatures, we next carry out a multi-comparison Tukey test, to find



(a) *Comparison vesus holiday*



(b) *Comparison versus convenience/community*



(c) *Holiday versus convenience/community*

Figure 3.3: *Scatter plots to show distance of each retail centre from various centroids*

out exactly which pairs of values show that significant differences exist between them. The results are illustrated in Figure 3.4, and show that comparison towns are, on average, busier than speciality and convenience retail centres. Holiday towns are a small group with insufficient data to establish any significant findings to relate total footfall to its signature type.

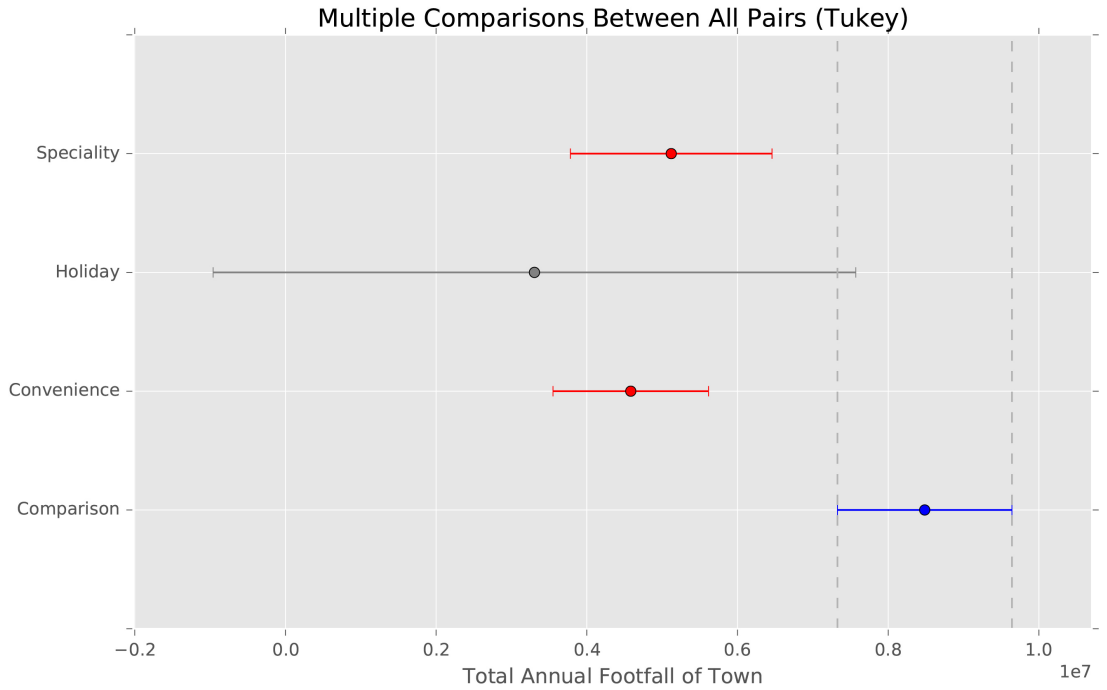


Figure 3.4: *Comparing total footfall with retail centre signature. Comparison centres are significantly busier than speciality or convenience/community centres.*

3.3 Principal Components Analysis to identify distinguishing features in the monthly signatures

Now that we have completed the clustering experiments and verified the four footfall signature profiles as: comparison, holiday, speciality and convenience/community, 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 3.1, 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. PCA 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:

$$Principal\ Component\ ij = L_1^i F_{Jan}^j + L_2^i F_{Feb}^j + \cdots + L_{12}^i F_{Dec}^j \tag{3.1}$$

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 %). We begin by computing twelve principal components, to coincide with the number of variables. Figure 3.5 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 3.1 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 0.399, 0.514 and -0.460 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 and December. December is clearly the peak month for comparison shopping centres. The predominance of March is something of a surprise. Looking at the relatively high value for footfall in

March for the convenience/community signature in figure 3.1, we hypothesise that this could possibly be a key month for identifying convenience/community centres.

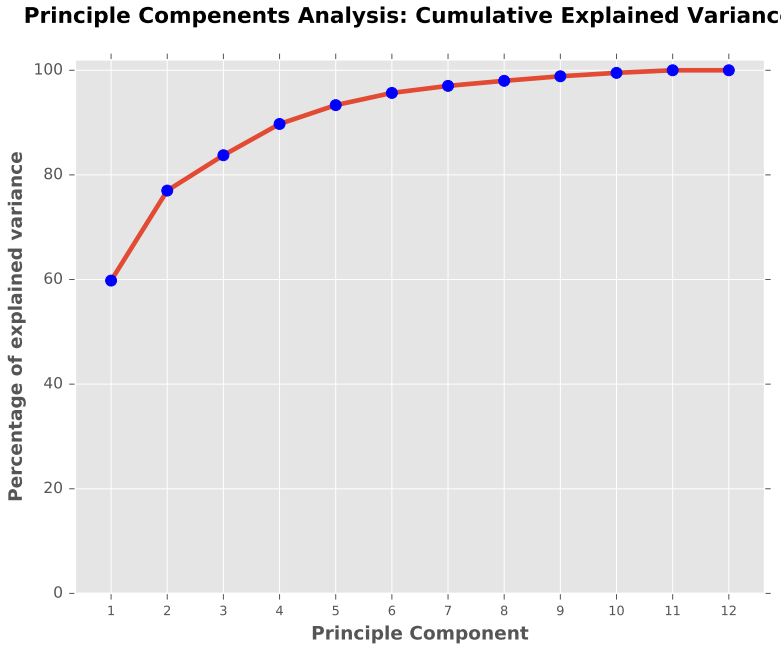


Figure 3.5

Table 3.1: *Loadings for Principal Components Analysis*

PC	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	-0.294	-0.216	-0.110	0.094	0.176	0.216	0.399	0.514	0.114	-0.104	-0.330	-0.460
2	-0.311	-0.277	-0.414	-0.259	-0.081	-0.057	0.121	0.344	0.001	0.073	0.246	0.615

Thus the two PC equations are as follows:

$$PC1_j = (-0.294 \times F_{Jan}) + (-0.216 \times F_{Feb}) + \cdots + (-0.460 \times F_{Dec}) \tag{3.2}$$

$$PC2_j = (-0.311 \times F_{Jan}) + (-0.277 \times F_{Feb}) + \cdots + (+0.615 \times F_{Dec}) \tag{3.3}$$

Finally, in Figure 3.6 we plot values of PC1 and PC2 for all of our sample 99 retail centres, and label them with the signature classification obtained from our K-Means study. From Figure 3.6 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).

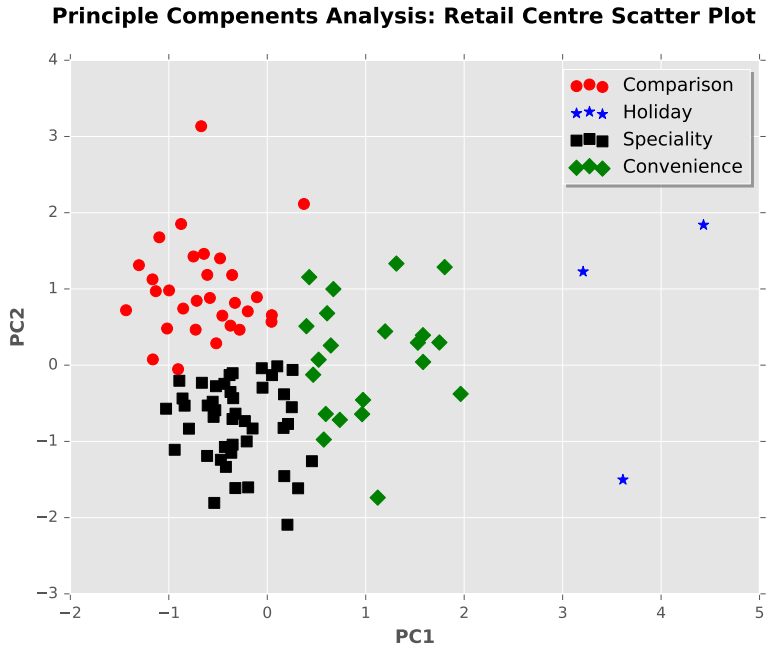


Figure 3.6

4 Conclusions and next steps

In this report we have demonstrated the following:

- Four clear footfall signatures exist, distinguishing different types of retail centre we have named comparison, holiday, convenience/community and speciality.
- Some centres have a clearer "offer" 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 signature types.

We have demonstrated that the four distinct signatures exist, but the crucial question to answer 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 collective offer?

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.