Classification of Postal Distribution Centres

Sandra M. Aleixo[1,2(✉)], Célia Fernandes[1], Paulo Ramos[1], Margarida Ramires[1], and Rute Milho[3]

[1] Department of Mathematics, ISEL – Instituto Superior de Engenharia de Lisboa, IPL – Instituto Politécnico de Lisboa, Rua Conselheiro Emídio Navarro, 1, 1959-007 Lisboa, Portugal
{sandra.aleixo,celia.fernandes,paulo.ramos}@isel.pt, A50051@alunos.isel.pt

[2] CEAUL – Centro de Estatística e Aplicações da Universidade de Lisboa, Faculdade de Ciências da Universidade de Lisboa, Bloco C6, Piso 4, 1749-016 Lisboa, Portugal

[3] CTT-Correios de Portugal, S.A., Strategy and Operations Development, Modelling, Networks and Optimisation Directorate, Av. Marechal Gomes da Costa, 13, 1849-001 Lisboa, Portugal
rute.m.milho@ctt.pt

Abstract. This work was conducted in the context of industrial mathematics, in partnership with CTT-Correios de Portugal, S.A., highlighting the strong link between academia and the business world. It addresses the classification of Postal Distribution Centres using multivariate statistical methods combined in a less common way. First, a correlation analysis was conducted on the variables selected for the study, revealing several statistically significant associations. To mitigate collinearity, Principal Component Analysis was applied to derive uncorrelated components, which were then used as inputs to Cluster Analysis. Both hierarchical and partitional clustering methods were applied, and solution quality was assessed using graphical inspection, silhouette coefficients, and within- and between-cluster sums of squares. Ward's method, k-means, and k-medoids produced consistent and similar groupings; in this paper, the k-means results are considered. Linear Discriminant Analysis and Multinomial Logistic Regression were used to assign new Postal Distribution Centres to the previously identified groups. The second classifier was used to corroborate the classification results obtained by LDA, since some LDA assumptions are violated and one of the clusters is small. The classification results from both methods showed a high degree of agreement. To avoid inflated accuracy caused by using the same data throughout all stages of the pipeline (standardisation, PCA, CA, LDA/MLR), the leave-one-out cross-validation method was applied across the entire pipeline, rather than only at the classification stage. As a result, the accuracy value obtained is a more realistic estimate of the model's performance. All analyses were conducted in R.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16759, pp. 224–241, 2027.
https://doi.org/10.1007/978-3-032-30518-3_15

Keywords: Industrial mathematics · Postal distribution centres · Principal component analysis · Cluster analysis · Linear discriminant analysis · Multinomial logistic regression · Cross-validation

1 Introduction

This work builds on a study conducted at CTT-Correios de Portugal, S.A., the Portuguese postal operator founded in 1520 and one of the oldest postal institutions worldwide. CTT operates multiple Postal Distribution Centres (PDCs), where postal items are sorted and prepared for last-mile delivery. From each PDC, mail carriers depart on delivery routes ("Giros"), which may be fixed (predefined routes) or flexible (routes that vary according to daily traffic conditions). Traffic refers to the volume of items handled and can be tracked (e.g., express or registered mail) or untracked (e.g., non-priority mail). Each route comprises an active path, corresponding to distances within areas containing distribution points, and a dead path, corresponding to the remaining travel distance (e.g., from the PDC to the start of the route and from the end of the route back to the PDC). Distribution points (DPs) denote the delivery locations. These operational concepts motivate the variables considered in this study [7].

The objectives are twofold: (i) to classify the existing PDCs into homogeneous groups and (ii) to use models that enable the assignment of a new PDC to one of the identified groups. Several multivariate statistical techniques were applied using R (version 4.5.2) [23].

First, the variables provided by the company were examined through a correlation analysis, and several statistically significant associations were found. To reduce dimensionality and mitigate collinearity, Principal Component Analysis (PCA) was applied. After Varimax rotation, a set of uncorrelated principal components (PCs) was retained, explaining a large proportion of the original variance. These components were then used as inputs to Cluster Analysis (CA) to form groups of PDCs with similar characteristics, in line with the company's operational perspective. Linear Discriminant Analysis (LDA) and Multinomial Logistic Regression (MLR) were used as classifiers for assigning new PDCs to the previously defined clusters. Finally, a fully consistent leave-one-out cross-validation procedure was applied to the entire modelling pipeline: standardisation, PCA, CA, and LDA/MLR. Related work in industrial mathematics and other real-world data applications has addressed similar problems; see, for example, [1,5,6,9].

The remainder of the paper is organised as follows. Section 2 summarises the main methodological concepts. Section 3 presents the empirical analysis and results. Section 4 concludes with the main findings and practical implications.

2 Methodology

In this study, PDCs were classified using a multivariate framework: PCA, CA, and LDA/MLR. PCA was applied to obtain a set of uncorrelated components

that served as inputs to CA, which identifies groups of PDCs with similar operational profiles. Finally, LDA and MLR were fitted to the resulting cluster labels to evaluate cluster separability and to assign new PDCs to the predefined groups, with the aim of achieving high expected classification accuracy.

2.1 Principal Component Analysis

PCA was used to reduce the dimensionality of a dataset by transforming an initial set of correlated variables into a smaller set of uncorrelated variables (PCs) that retain most of the information in the original data. Each PC is a linear combination of the original variables. Components are ordered by decreasing explained variance: the first accounts for the maximum possible variance, the second accounts for the maximum remaining variance subject to being orthogonal to the first, and so on. In this study, PCA was used in three steps: (i) assessing data adequacy, (ii) selecting the number of components to retain, and (iii) deriving variables for subsequent CA.

Assessment of Data Adequacy. Before applying PCA, data suitability was assessed using the Kaiser-Meyer-Olkin (KMO) statistic and Bartlett's test of sphericity. The KMO statistic measures the proportion of variance that may be attributed to common factors. Values range from 0 to 1, with higher values indicating that PCA is appropriate. As a rule of thumb, values between 0.7 and 0.8 are considered acceptable, values above 0.8 are considered good, and values below 0.5 suggest that PCA is not suitable for the data. Bartlett's test of sphericity tests the null hypothesis that the correlation matrix is the identity matrix. Rejecting the null hypothesis indicates that at least some variables are significantly correlated, which supports the use of PCA.

Determining the Number of Principal Components. If the adequacy criteria are met, PCA can be fitted and the number of components to retain can be selected using standard criteria. The following criteria were considered: (i) the cumulative percentage of variance explained (retaining enough components to explain at least 70% of the total variance) and (ii) the Kaiser criterion (retaining components with eigenvalues greater than or equal to one).

Obtaining Variables for CA Using PCA. To improve interpretability, Varimax rotation was applied to the retained components. The Varimax orthogonal rotation method aims to ensure that, for each principal component, only a few loadings are substantial, with all others close to zero; that is, it maximises the variation among the loadings within each principal component. Differences in the variables' contributions are amplified by increasing the loadings of those that contribute most to the component and decreasing the loadings of those that contribute least. Loadings with an absolute value of 0.5 or greater are considered relevant. The resulting scores of the rotated, uncorrelated components

were then used as inputs to the subsequent CA. For further details on PCA, see, for example, [11,13,14,19].

2.2 Cluster Analysis

CA partitions a set of objects into groups (clusters) based on measurements of p variables, with the aim of producing clusters that are internally cohesive and well separated from each other. In practice, CA involves two key steps: selecting an appropriate clustering solution (including the number of clusters) and assessing its quality.

Clustering Methods. Objects are grouped so that members of the same cluster are more similar to each other than to members of different clusters. Clustering can be performed using hierarchical or partitional (non-hierarchical) methods. Hierarchical methods rely on a pairwise distance matrix, while partitional methods operate directly on the data matrix. In this study, multicollinearity was mitigated by using uncorrelated component scores as inputs. Euclidean distance, computed on standardised variables, was used as the dissimilarity measure. Among hierarchical approaches, single linkage, complete linkage, average linkage, and Ward's method were considered; only the methods that yielded the most coherent solutions were reported. For partitional clustering, the k-means and k-medoids algorithms were applied. The **k-means** method seeks to partition the observations into k clusters by minimising the within-cluster sum of squares. The algorithm alternates between (i) assigning each observation to the cluster with the nearest centroid and (ii) updating centroids as the mean of the assigned observations, until convergence. To visualise the clustering results, a two-dimensional graphical representation based on the first two principal components was used.

Cluster Validation. Cluster quality was assessed using the **silhouette coefficient** for both hierarchical and partitional solutions. For observation i, the silhouette value S_i compares its average dissimilarity to objects within its assigned cluster with its average dissimilarity to objects in the nearest alternative cluster. Values of S_i close to 1 indicate well-clustered observations; values near 0 suggest borderline assignments between two clusters; and negative values indicate potential misclassification. For further details on CA, see, for example, [8,12,17,18].

2.3 Linear Discriminant Analysis

LDA is used to assess whether the variables considered can reliably separate the groups previously obtained by CA and to assign new observations to those groups. LDA constructs up to $m = \min\{k-1, p\}$ linear discriminant functions for k groups and p discriminant variables, aiming to maximise between-group separation while controlling within-group variability.

Assumption Assessment. Classical LDA relies on approximately multivariate normal distributions within each group and equal within-group covariance matrices across groups. These conditions were assessed using standard diagnostic procedures, including the Royston test for multivariate normality [20–22,24,26] and Box's M test for homogeneity of covariance matrices [3,4,25]. In practice, LDA is often reasonably robust to moderate departures from these assumptions when group sizes are not too small and the separation between groups is adequate. In addition, the discriminant variables are not linear combinations of one another; LDA assumes that the predictor variables are not highly multicollinear. This concern is mitigated when using uncorrelated component scores as inputs.

Model Estimation. LDA constructs m discriminant functions as linear combinations of the independent variables. The coefficients are chosen to maximise between-group separation relative to within-group dispersion, yielding uncorrelated discriminant scores. Discriminant-score plots provide a convenient way to visualise group separation and identify potential overlap or outliers.

Classification Rule. Classifying observations–either those used to fit the discriminant functions or new cases–into predefined groups is one of the main objectives of discriminant analysis. Classification can be performed using classification functions. If group sizes are unequal or prior probabilities differ, these functions can be adjusted by adding a term that reflects group priors. For each observation, a classification score is computed for each group by evaluating the corresponding classification function, and the observation is assigned to the group with the highest score. Classification functions can be used both to assess the classification performance of the discriminant model and to assign new observations to the predefined groups.

Evaluation of Classification Performance. To assess classification performance, the rates of correct and incorrect classification were estimated. One common approach is the leave-one-out (jackknife) method, where one observation is removed, the discriminant functions are re-estimated using the remaining data, and the omitted observation is then classified. Repeating this procedure for all observations provides a nearly unbiased estimate of the misclassification rate and is often preferred for small to moderate sample sizes. See, for example [16] for further details.

2.4 Multinomial Logistic Regression

MLR is a supervised classification method used when the response variable has more than two nominal categories. In this study, it is used to assess whether the variables considered can reproduce the predefined group structure and to assign new observations to those groups. Unlike LDA, MLR does not require the explanatory variables to follow a multivariate normal distribution within each

group, nor does it require equality of covariance matrices across groups. For this reason, it provides a less restrictive framework for multiclass classification problems.

Assumption Assessment. MLR relies on fewer distributional assumptions than LDA. In particular, it does not assume multivariate normality of the explanatory variables within groups or homogeneity of covariance matrices across groups. However, some conditions should still be considered. The response variable must consist of mutually exclusive nominal categories, the observations should be independent, and severe multicollinearity among the explanatory variables should be avoided. In addition, each category should contain a sufficient number of observations to allow stable estimation of the model parameters. When component scores are used as explanatory variables, the risk of multicollinearity is reduced because these scores are uncorrelated by construction.

Model Estimation. MLR models the relationship between a set of explanatory variables and a categorical response variable with more than two possible outcomes. One category is taken as the reference category, and the model estimates the relative effect of the explanatory variables on the probability of belonging to each of the remaining categories. The estimated model therefore provides, for each observation, a set of predicted probabilities, one for each possible group. These probabilities summarise the evidence in favour of assigning the observation to each category.

Assessment of the Fitted Model. The adequacy of the fitted MLR model can be assessed using standard model-based criteria and diagnostic measures. These may include likelihood-based tests, information criteria, pseudo-coefficients of determination, and the statistical significance of the estimated coefficients. Such measures provide information on the overall contribution of the explanatory variables to group separation. However, when the response categories correspond to groups obtained from a previous unsupervised procedure, these results should be interpreted as evidence of reproducibility of the predefined group structure rather than as independent validation of that structure.

Classification Rule. Classification is based on the predicted probabilities obtained from the fitted MLR model. For each observation, the model estimates the probability of membership in each predefined group. The observation is then assigned to the group with the highest predicted probability. This rule can be applied both to the observations used to fit the model and to new observations, provided that the same explanatory variables are available and processed in a consistent way.

Evaluation of Classification Performance. To assess classification performance, the predicted group labels are compared with the original group memberships. This comparison is usually summarised through a confusion matrix, from which the rates of correct and incorrect classification can be obtained. A robust internal assessment can be obtained using a leave-one-out (jackknife) procedure. In this approach, one observation is removed from the dataset, the MLR model is fitted using the remaining observations, and the omitted observation is then classified according to the highest predicted probability. This procedure is repeated until each observation has been omitted once. The global confusion matrix is then constructed by aggregating the out-of-sample predictions obtained across all iterations. This provides a less optimistic estimate of the misclassification rate and is particularly useful when an independent validation sample is not available. For further details on MLR, see, for example, [10].

3 Practical Application to CTT Data

This work is part of a study conducted at CTT-Correios de Portugal, S.A., which operates multiple PDCs. The main goals are to (i) group existing PDCs into homogeneous, well-separated clusters and (ii) use models that enable the assignment of a new PDC to one of the identified groups. These goals were addressed using multivariate statistical techniques.

3.1 Variables

The dataset provided by CTT includes the following variables:

- PDC identifier (`pdc`);
- Average number of postmen per kilometre travelled (`postmen`);
- Population density, inhabitants per km^2 (`pop_density`);
- Fast transport, by car or moped, per kilometre travelled (`fast_transp`);
- Slow transport, on foot or by bicycle, per kilometre travelled (`slow_transp`);
- Distribution points per kilometre travelled (`dist_points`);
- Traffic volume, number of objects, per kilometre travelled (`traffic_vol`);
- Total distance travelled, km (`route`).

The average number of postmen per kilometre travelled is defined as the average number of postmen working on a given route divided by the length of that route. This variable captures distribution efficiency in relation to the human effort required to cover the route: a high number of postmen per kilometre may indicate lower efficiency, whereas a low number may indicate excessive workload.

Population density per km^2 is the number of inhabitants per unit area (km^2) in the region considered. It serves as a proxy for demand for postal services, since more densely populated areas typically require more deliveries. It also supports route planning and resource allocation by helping to adjust the number of postmen and vehicles to the underlying population.

Fast transport (car or moped) per kilometre travelled reflects the distance covered using faster modes of transport. This variable assesses the capacity for rapid coverage and delivery efficiency, and supports route planning by indicating when motorised transport is prioritised, for example, on longer routes or in less walkable areas.

Slow transport (on foot or by bicycle) per kilometre travelled reflects the distance covered on foot or by bicycle. It helps identify areas where delivery speed is constrained–often dense urban centres or narrow streets–and supports optimisation in locations where vehicles have limited access.

Distribution points per kilometre travelled represents the number of delivery points (residences, businesses, post office boxes) relative to the distance travelled. It measures delivery concentration: high values indicate dense delivery patterns, whereas low values suggest dispersed areas, which can reduce efficiency.

Traffic volume (number of objects) per kilometre travelled is the average number of letters or parcels delivered per kilometre travelled. This variable quantifies workload relative to distance and is useful for scaling resources, predicting delivery times, and planning routes. High values may indicate more productive routes, whereas low values may reflect more dispersed delivery patterns.

The distance travelled (km) represents the total distance travelled by the routes associated with each PDC. It enables classification of distribution centres according to the size and complexity of their routes and supports strategic planning, such as grouping similar PDCs for logistics optimisation.

3.2 Correlation Analysis

Since CA is sensitive to strong correlations among inputs, which can influence the results, pairwise associations among the original variables were first examined. Figure 1 shows the correlogram computed using Spearman's rank correlation, which is appropriate when normality cannot be assumed for the variables under study.

The correlogram presented in Fig. 1 indicates that most variables are strongly correlated. Therefore, applying CA directly to the original variables could bias the clustering solution. To mitigate this issue, PCA was used to derive uncorrelated component scores, which were then used as inputs.

3.3 Principal Component Analysis

The suitability of the data for PCA was assessed using the KMO statistic and Bartlett's test of sphericity. The KMO value was 0.84, and Bartlett's test strongly rejected the null hypothesis ($p < 0.001$), indicating that PCA is appropriate given the presence of significant correlations among the original variables.

PCA was then performed on standardised variables, since they are measured on different scales. Table 1 summarises component importance.

Only the first two components have eigenvalues greater than one (Kaiser criterion), and together they explain approximately 83% of the total variance.

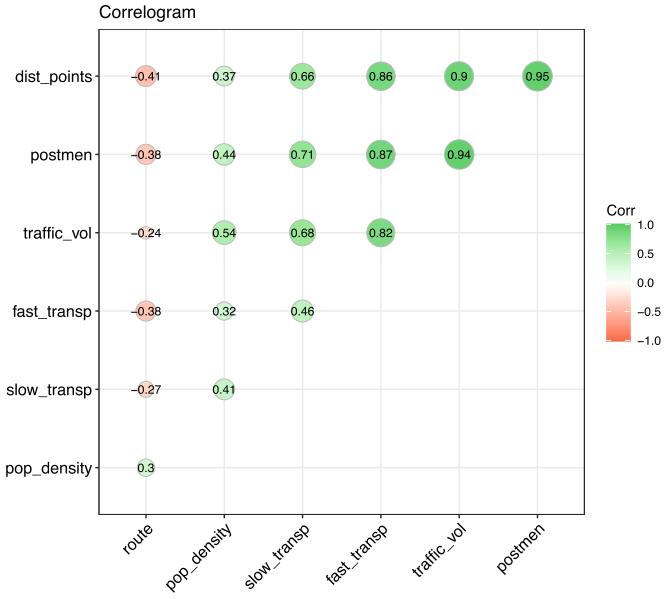

Fig. 1. Spearman rank correlations for the original variables.

Table 1. PCA component importance.

Importance of components	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	2.162	1.059	0.808	0.614	0.316	0.215	0.162
Proportion of variance	0.668	0.160	0.093	0.054	0.014	0.007	0.004
Cumulative proportion	0.668	0.828	0.922	0.975	0.990	0.996	1.000

Therefore, the first two components were retained, and an orthogonal Varimax rotation was applied to improve interpretability.

Table 2 reports the loadings of the rotated principal components (RPCs). The percentage of variance explained by the first rotated principal component is approximately 65%, and the percentage of total variance explained by the first two rotated principal components is approximately 83%.

All variables have relevant weights (absolute loading ≥ 0.5) on at least one rotated component. The rotated components were interpreted as follows: RC1 has high negative loadings on all variables except `route`, and thus captures operational intensity and service density (more postmen, higher traffic, more distribution points, higher population density, and more slow-transport activity); RC2 has a strong positive loading on `route` and a moderate positive loading on `pop_density`, contrasting long/extensive routes with more compact delivery environments.

Table 2. Loadings after Varimax rotation.

	postmen	pop_density	fast_transp	slow_transp	dist_points	traffic_vol	route
RC1	-0.964	-0.753	-0.673	-0.889	-0.930	-0.964	0.196
RC2	-0.180	0.447	-0.289	0.000	-0.218	-0.108	0.918

3.4 Clustering PDCs Using PCA Scores

CA was performed on the scores of the two rotated components. Both hierarchical and partitional methods were considered, for different numbers of clusters, and the results were analysed to identify the best method.

A good cluster analysis should produce groups that are internally cohesive, well separated from each other, appropriately defined in number, and meaningful and useful for practical decision-making. Therefore, for each method, a graphical analysis of the cluster representation was carried out to assess separation between groups; the silhouette plot was also analysed to verify whether the number of negative silhouette values was reasonable; the average silhouette width was calculated to determine whether it was acceptable; and, finally, the practical meaningfulness of the clusters obtained was assessed.

Comparing the results, three approaches that produced coherent and operationally meaningful solutions were identified: Ward's method, k-means, and k-medoids. Across these methods, the most consistent solutions were obtained with seven clusters. The k-medoids solution differed from k-means in only five PDC assignments.

In what follows, only the k-means solution is reported, as it provided the most stable and interpretable clustering in this study. Table 3 reports cluster sizes and centroids in the rotated-component space.

Table 3. Number of PDCs per cluster and cluster centroids (rotated PC space).

Clusters	1	2	3	4	5	6	7
Sizes	9	29	64	5	32	61	13
RC1 means	0.547	-0.870	0.276	-2.279	0.528	0.543	-2.769
RC2 means	2.839	-0.471	-0.816	2.669	0.851	-0.085	0.377

Based on the rotated loadings (Tab. 2) and the cluster centroids (Tab. 3), the clusters can be characterised as follows:

– Clusters 4 and 7 have similarly high negative centroid values on RC1, but they differ in their centroid values on RC2. PDCs in both clusters have many postmen, high traffic, and many distribution points; they are located in areas with higher population density and rely more on slow transport. However, PDCs in cluster 4 have longer routes, whereas those in cluster 7 have more compact routes.

- Clusters 3, 5, and 6 have centroid values on RC1 that are positive and similar in magnitude, with smaller absolute values than those in clusters 4 and 7; they differ mainly in their centroid values on RC2. PDCs in these three clusters have fewer postmen, lower traffic, are located in areas with lower population density, and make less use of slow transport. They differ primarily in route length: routes are long in cluster 5, medium in cluster 6, and short in cluster 3.
- In cluster 1, the centroid value on RC1 is small, whereas the centroid value on RC2 is positive and the highest among all clusters. This corresponds to PDCs with very long routes located in areas with average population density.
- In cluster 2, the centroid values for the two RCs are both negative and similar in magnitude. This corresponds to PDCs with intermediate values in terms of the number of postmen, traffic volume, population density, and transport modes, but with more compact routes.

For confidentiality reasons, the real identifiers of the PDCs are not disclosed, and no geographical analysis of the cluster composition is presented. According to CTT, the results of this analysis are interesting and consistent with what was expected.

Figure 2 visualises the seven clusters and Fig. 3 shows the corresponding silhouette plot. The average silhouette width is approximately 0.41, with few observations showing negative silhouette values, suggesting that most PDCs are well assigned to their clusters.

The k-means objective diagnostics were also evaluated. It was verified that the within-cluster sums of squares are relatively small, and the ratio of between-cluster to total sum of squares is high (0.86), indicating substantial separation.

Overall, the seven-cluster k-means solution is supported by standard validity diagnostics and provides a coherent operational segmentation of PDCs, according to CTT's assessment.

3.5 Linear Discriminant Analysis and Multinomial Logistic Regression Classifiers

LDA was fitted using the rotated PC scores as discriminant variables and the k-means cluster labels as groups. The aim was to use this analysis to enable the classification of new PDCs into the predefined clusters. Since this method is based on assumptions, the initial step was to verify those assumptions. In this sense, multivariate normality within clusters and homogeneity of covariance matrices across clusters were assessed. Multivariate normality was supported in five of the seven clusters, whereas equality of covariance matrices was not. Despite these limitations, it was decided to proceed with the classification using LDA. The high classification performance reported below indicates that the violation of the assumptions had no major practical consequences.

However, MLR has advantages over LDA in contexts where its assumptions are potentially violated. In particular, it does not require the assumption of multivariate normality of the predictors or the equality of covariance matrices

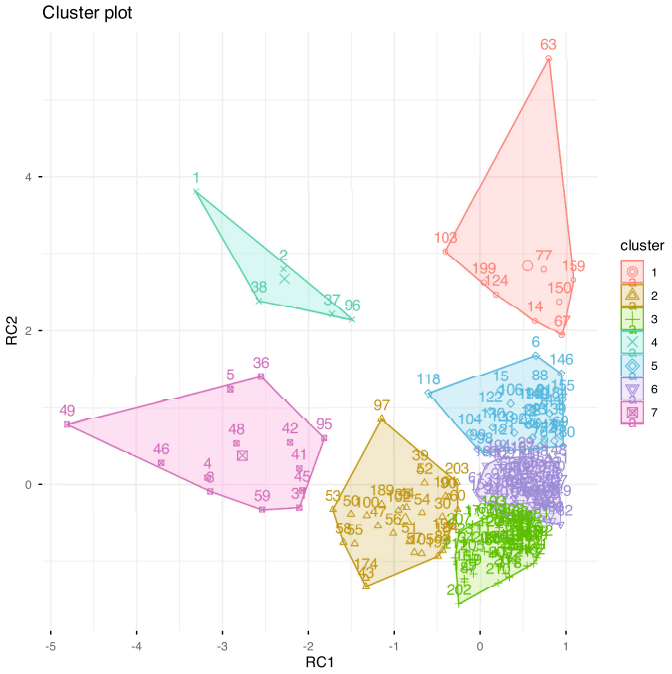

Fig. 2. Rotated PCA (RC1–RC2) visualisation of the k-means clusters.

between groups, making it more flexible in modelling real data. Additionally, because it does not depend on the estimation of group-specific covariance matrices, it tends to exhibit greater stability in situations where some clusters are small, as observed in this study. In this sense, the complementary use of this approach helps mitigate the limitations of LDA and reinforces the robustness of the proposed classification rule.

To evaluate classification accuracy, it is important to use an adequate cross-validation method. In this study, it is critical to mention that the data analysed correspond to all PDCs of the organisation, and therefore do not constitute a sample in the classical statistical sense. Given that each PDC has specific operational characteristics, the creation of independent training/test subsets could compromise the representativeness of the analysis. Therefore, the leave-one-out (jackknife) cross-validation method was adopted, which maximises the use of available information and allows direct evaluation of the stability of the assignment rule. Thus, classification accuracy was evaluated using this cross-validation procedure for both methods. The leave-one-out confusion matrices are presented in Table 4.

Inspection of the LDA confusion matrix shows that clusters 4 and 7 have no misclassified PDCs. The correct classification rates for clusters 1, 2, 3, 5, and 6 are 0.8889, 0.8276, 0.9844, 0.9375, and 0.9836, respectively. The overall correct classification rate is approximately 0.9531, indicating that the derived rule is

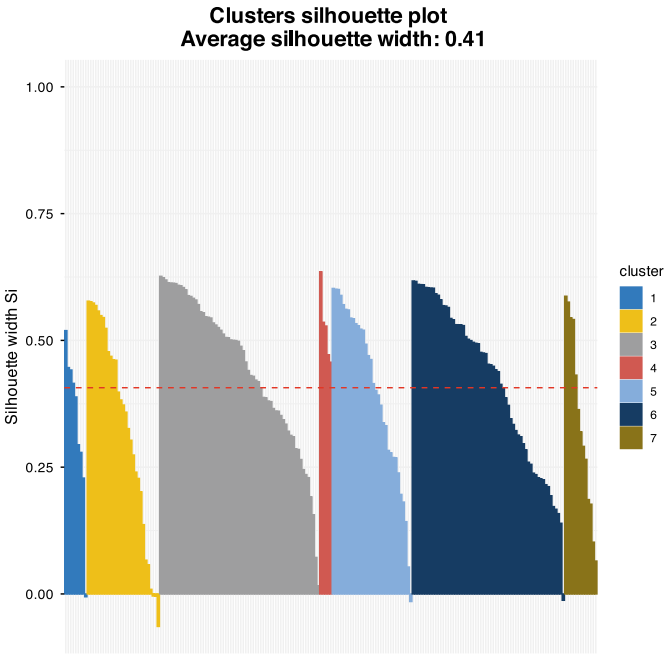

Fig. 3. Silhouette plot.

effective for assigning PDCs to clusters and can be used to classify new PDCs with low expected error.

Inspection of the MLR confusion matrix shows that the overall correct classification rate is approximately 0.9624. Clusters 1 and 4 have no misclassified PDCs. The correct classification rates for clusters 2, 3, 5, 6, and 7 are 0.9310, 0.9531, 0.9688, 0.9836, and 0.9231, respectively.

It can be noted that there are 5 PDCs that are in the same cluster for both classifiers.

To classify a new PDC using the LDA classifier, the RPC scores were computed by standardising the original variables and then applying the rotated loadings. The seven LDA classification functions were then evaluated, and each PDC was assigned to the cluster with the highest score.

To classify a new PDC using the MLR classifier, the RPC scores were first obtained by standardising the original variables and applying the rotated loadings. The fitted MLR model was then used to compute the predicted membership probabilities for the seven clusters, and each PDC was assigned to the cluster with the largest predicted probability.

Table 4. LOOCV (jackknife) confusion matrices for LDA and MLR classifications.

Prediction	LDA							MLR						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	8	0	0	0	1	0	0	9	0	0	0	0	0	0
2	0	24	4	0	0	1	0	0	27	0	1	0	0	1
3	0	0	63	0	0	1	0	0	1	61	0	0	2	0
4	0	0	0	5	0	0	0	0	0	0	5	0	0	0
5	0	0	0	0	30	2	0	0	0	0	0	31	1	0
6	0	0	1	0	0	60	0	0	0	1	0	0	60	0
7	0	0	0	0	0	0	13	0	1	0	0	0	0	12

3.6 Classification Improvement Using Leave-One-Out Validation of the Overall Pipeline: External Cross-Validation

LDA and MLR classifiers do not provide an independent validation of the clusters because the same dataset was used to derive the models and to assess their classification performance. Therefore, the high apparent accuracy should not be interpreted as a low expected error for new PDCs, since leave-one-out was applied only to the classifier, while standardisation, PCA, rotation, and clustering were kept fixed and were fitted using the full dataset.

The originally reported evaluation could lead to an optimistic estimate of predictive performance. It should be noted that only the final LDA and MLR classifications were subjected to leave-one-out validation, while earlier stages of the pipeline (standardisation, PCA, Varimax rotation, and clustering) were estimated using the full dataset. To address this concern, a fully consistent leave-one-out cross-validation procedure was applied to the entire modelling pipeline. In this framework, each iteration of the cross-validation loop re-estimates all data-dependent steps exclusively on the training subset, including standardisation, PCA, Varimax rotation, k-means clustering, LDA, and MLR. Note that cluster solutions are made comparable across resamples via label-switching resolution (i.e., label matching/alignment by permuting cluster labels to maximise agreement). This overall validation ensures that every observation is evaluated strictly out of sample with respect to the full modelling process, rather than only the final classification stage. For further details on the global pipeline, see, for example, [2,15,27].

Using the procedure presented above, the leave-one-out confusion matrices corresponding to the k-means clustering and to both classifiers are presented in Table 5. It is important to note that the cluster assignments obtained under the full leave-one-out cross-validation framework are not necessarily identical to those derived in the earlier analysis, where validation was restricted to the classification stage. This is because all components of the pipeline, including clustering, are re-estimated at each iteration. In fact, differences in cluster assignments across leave-one-out iterations are expected when the full pipeline is re-

estimated. All stages of the pipeline procedure are data-dependent, so small perturbations in the sample propagate through the feature space and may alter cluster centroids and memberships. Additionally, cluster labels are inherently subject to permutation invariance (label switching) and are therefore aligned across iterations using an optimal matching procedure. Consequently, observed differences reflect the intrinsic variability of the data-driven pipeline rather than methodological inconsistency.

Table 5. LOOCV (jackknife) confusion matrices for k-means, LDA and MLR classifications.

Prediction	K-1	2	M 3	E 4	A 5	N 6	S 7	L 1	D 2	A 3	4	5	6	7	M 1	L 2	R 3	4	5	6	7
1	11	0	0	0	0	0	0	10	0	0	0	1	0	0	8	0	0	0	3	0	0
2	0	76	0	0	0	0	0	0	76	0	0	0	0	0	0	76	0	0	0	0	0
3	0	0	21	0	0	0	0	0	0	21	0	0	0	0	1	0	19	0	0	1	0
4	0	0	0	42	0	0	0	0	10	0	32	0	0	0	0	0	0	39	0	1	2
5	2	0	0	0	5	0	0	4	0	0	0	3	0	0	2	0	0	0	5	0	0
6	0	2	0	0	0	45	0	0	10	0	0	0	37	0	0	2	0	1	0	44	0
7	0	0	0	0	0	0	9	0	0	0	1	0	0	8	0	0	0	0	0	0	9

Under full leave-one-out cross-validation of the global pipeline, the confusion matrix for the k-means baseline yields 209 correctly classified observations out of 213, corresponding to an overall accuracy of 98.1%, indicating strong stability of the clustering structure under small perturbations of the data. Based on the confusion matrices of both classifiers, the calculated accuracies were 87.8% for LDA and 93.9% for MLR. The gap between the accuracies of LDA and MLR suggests that the separation between clusters is not fully linear in the transformed feature space, with the more flexible multinomial model better capturing the underlying structure.

The percentages of correct classification for the global pipeline validation process, based on the matrices in Table 5, are lower than those reported in the analysis of Table 4, but still indicate a high level of agreement. This reduction confirms that the earlier performance estimates were indeed optimistic due to partial reuse of information in the preprocessing and clustering stages.

Importantly, the revised results provide a more realistic estimate of the expected generalisation performance of the global pipeline for new PDC observations.

4 Conclusions

This study addressed the classification of CTT PDCs using several multivariate statistical techniques, with the outputs of each step used as inputs to the next.

First, the aim was to group existing PDCs into homogeneous clusters based on the operational variables under study. Since the original variables exhibited strong correlations, PCA was applied to derive a small set of uncorrelated variables with minimal loss of information. The RPC scores were then used as inputs to CA, yielding a seven-cluster solution with good internal cohesion and satisfactory separation. This solution was supported by standard validity diagnostics and was operationally meaningful for the company.

The obtained clusters can be characterised as follows: two clusters consist primarily of PDCs with many postmen, high traffic, and many distribution points, located in areas with higher population density and relying more on slow transport; three clusters comprise PDCs with fewer postmen and lower traffic, located in areas with lower population density, and making less use of slow transport; one cluster consists of PDCs with very long routes in areas with average population density; and one cluster comprises PDCs with intermediate values for the number of postmen, traffic volume, population density, and transport modes, but with more compact routes.

LDA results indicate that the rotated PCs provide strong discrimination among the seven clusters, with a high overall correct classification rate and a low expected misclassification rate. Therefore, the derived LDA classification functions constitute a practical tool for assigning new PDCs to the established groups.

To assess the robustness of the classification rule obtained, the LDA results were compared with those of an alternative MLR model. This approach was considered particularly suitable given the violation of some LDA assumptions, namely homogeneity of covariance matrices across groups and multivariate normality, as well as the small size of some clusters. The results obtained through MLR were globally consistent with those obtained using LDA, showing a high degree of agreement in the assignment of PDCs to the different groups. This convergence between methodologies, based on distinct assumptions, reinforces the robustness of the identified segmentation and suggests that the captured cluster structure is stable.

To prevent overestimating accuracy due to reusing the same data across all stages of the global pipeline (standardisation, PCA, CA, LDA/MLR), leave-one-out cross-validation was implemented throughout the entire process rather than only during the classification step. This approach yields a more reliable estimate of the model's true performance.

Overall, the work illustrates how established multivariate methods can be combined into an effective decision-support framework for operational segmentation in an industrial setting, reinforcing the value of collaboration between academic modelling and business expertise.

Acknowledgments. This work was funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under CEAUL Research Unit, UID/00006/2025, DOI: https://doi.org/10.54499/UID/00006/2025, and by the European Union – NextGenerationEU through the project UID/PRR/00006/2025, DOI: https://doi.org/10.54499/UID/PRR/00006/2025. The authors thank CTT – Correios de Portugal,

S.A., for participating in the study and for authorising the publication of the information used in this paper.

Disclosure of Interest. The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this article. All analyses were conducted independently, and no external funding sources affected the study design, data collection, or interpretation of results. During the preparation of this manuscript, the authors used ChatGPT-5 (OpenAI) for the purpose of improving the clarity and readability of the manuscript. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

References

1. Aleixo, S.M., Pereira, B., Fernandes, T., Rodrigues, J., Pina, M.: Cluster analysis for two HoReCa channels of a large grocery retail company. In: Gervasi, O., Murgante, B., Garau, C., Taniar, D., Rocha, A., Faginas Lago, M. (eds.) Computational Science and Its Applications - ICCSA 2024, LNCS, vol. 14815, pp. 337–356. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-65154-0_21
2. Ambroise, C., McLachlan, G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. **99**(10), 6562–6566 (2002)
3. Box, G.E.P.: A general distribution theory for a class of likelihood criteria. Biometrika **36**, 317–346 (1949)
4. Box, G.E.P.: Non-normality and tests on variances. Biometrika **40**, 318–335 (1953)
5. Brentari, E., Dancelli, L., Manisera, M.: Clustering ranking data in market segmentation: a case study on the Italian McDonald's customers' preferences. J. Appl. Stat. **43**(11), 1959–1976 (2016)
6. Bulut, H., Oner, Y.: The evaluation of socio-economic development of development agency regions in Turkey using classical and robust principal component analyses. J. Appl. Stat. **44**(16), 2936–2948 (2017)
7. CTT – Correios de Portugal, S.A (2025). https://www.ctt.pt/
8. Everitt, B.S., Landau, S., Leese, M., Stahl, D.: Cluster Analysis. John Wiley & Sons, United Kingdom (2011)
9. Gardner-Lubbe, S.: Linear discriminant analysis for multiple functional data analysis. J. Appl. Stat. **48**(11), 1917–1933 (2021)
10. Gelman, A., Hill, J.: Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press (2006)
11. Hair, J.F., Jr., Black, W.C., Babin, B.J., Anderson, R.E.: Multivariate Data Analysis. Pearson Education Limited, London (2013)
12. Hennig, C., Meila, M., Murtagh, F., Rocci, R.: Handbook of Cluster Analysis. Chapman and Hall/CRC, Boca Raton, Florida (2016)
13. Johnson, R.A., Wichern, D.W.: Applied Multivariate Statistical Analysis. Pearson Prentice Hall, Hoboken, New Jersey (2007)
14. Jolliffe, I.T.: Principal Component Analysis. Springer-Verlag, New York (2002)
15. Kuhn, M., Johnson, K.: Applied Predictive Modeling. Springer (2013)
16. McLachlan, G.J.: Discriminant Analysis and Statistical Pattern Recognition. John Wiley & Sons, Hoboken, New Jersey (2005)
17. Rai, P., Singh, S.: A survey of clustering techniques. Int. J. Comput. Appl. **7**, 1–5 (2010)

18. Reddy, C.K., Aggarwal, C.C.: Data Clustering: Algorithms and Applications. Chapman and Hall/CRC, Boca Raton, Florida (2014)
19. Rencher, A.C., Christensen, W.F.: Methods of Multivariate Analysis. John Wiley & Sons, Hoboken, New Jersey (2012)
20. Royston, P.: Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. Appl. Stat. **32**, 121–133 (1983)
21. Royston, P.: Approximating the Shapiro-Wilk W-test for non-normality. Stat. Comput. **2**, 117–119 (1992)
22. Royston, P.: A toolkit for testing for non-normality in complete and censored samples. The Statistician **42**(1), 37–43 (1993)
23. The R Project for Statistical Computing (2025). https://www.r-project.org/
24. Shapiro, S.S., Wilk, M.B.: An analysis of variance test for normality (complete samples). Biometrika **52**(3/4), 591–611 (1965)
25. Srivastava, M.S.: Some tests concerning covariance matrices in high-dimensional data. J. Jpn. Stat. Soc. **35**(2), 251–272 (2005)
26. Thode, H.: Testing for Normality. CRC Press, New York (2002)
27. Varma, S., Simon, R.: Bias in error estimation when using cross-validation for model selection. BMC Bioinform. **7**(91), 1–8 (2006)