Studying Road Transportation Demand in the Spanish Industrial Sector through *k*-means Clustering

Carlos Alonso de Armiño[1[0000-0002-9228-4914]], Miguel Ángel Manzanedo[1[0000-0002-9046-8306]], Álvaro Herrero[1[0000-0002-2444-5384]]

[1] Departamento de Ingeniería Civil, Escuela Politécnica Superior, Universidad de Burgos, Av. Cantabria s/n, 09006, Burgos,
{caap, mmanz, ahcosio}@ubu.es

Abstract. Transportation is the economic activity that is the most tightly coupled with the other ones. As a result, knowledge about transportation in general, and market demand in particular, is key for an economic analyisis of a sector. In present paper, the official data about the industrial sector, coming from the Ministry of Public Works and Transport in Spain, is analysed. In order to do that, k-means clustering technique is applied to find groupings or patterns in the dataset that contains data from a whole year (2015). Samples allocation to clusters and silhouette values are used to characterize the demand of the industrial transportation. Useful insights into the analysed sector are obtained by means of the clustering technique, that has been applied with 4 different criteria.

Keywords. Clustering, *k*-means, road transportation, logistics, industrial sector.

1 Introduction

The concept of economic activity is an aggregate of various sector functions that are developed in a system, and that are largely interrelated. A common frame of aggregation of such activities are the geographical territories and the periods of time that we understand as natural years.

Transportation is the economic activity that is the most tightly coupled with the other ones, given that it is in those relationships in which its essence is based. This may have motivated that this activity has historically been controlled and supervised by the governments, for which it has constituted, in a large part of its history. Nowadays, being a logically liberalized activity, transportation continues to be a subject closely supervised and controlled by governments in its different territories. Under the current geopolitical framework, the European Union has been demonstrating for decades its interest in transport activities that unite their territories through their exchanges of people and goods.

Scant attention has been devoted until now to apply clustering techniques in order to analyse transportation data. In [1] authors apply clustering techniques in order to identify areas of high porosity, or high permeability, for pedestrians along the border region of some countries using terrain, land use, and road data along with geocomputational methods. Obtained results could be potentially useful for decision making processes for tourism development and road transportation management in that region.

Although many previous studies have been focused in vehicle route optimization, few of them applied clustering algorithms. That is the case of [2] where *k*-means is proposed to improve the computational performance of a new algorithm aimed at solving vehicle routing problems. Authors claim that the new extended formulation where the clustering algorithm is applied, captures truckload and travel distance, supporting the development of systems that respond fast, possibly online, to changes in the real problem situations.

The analysis of transport activity is normally developed under two aggregation frameworks: 1) the one that distinguishes between travellers and merchandise, and 2) the one that distinguishes the so-called different modes of transport (rail, highway, and navigable roads). Focusing on the transport of goods, and analyzing the different modes in the European Union, the road presents an overwhelming supremacy over the rest of the modes, representing 75.4% of the total amount of tons per kilometres transported within and between EU member states in 2016 (latest available data) [3]. For all the above, the European Union has imposed on its member states control mechanisms of transport activity, especially the transport of goods by road, and has designed a homogenized data collection framework for its member states [4]. In the case of Spain, these processes have been integrated into the National Statistical Plan, and materialized in the issuance and collection of data from the Permanent Survey of Transport of Goods by Road (PSTGR), that is issued, collected and supervised by the General Directorate of Transport integrated in the Spanish Ministry of Public Works and Transport.

The data collected in PSTGR are designed for a regional statistical representativeness, with the observation unit being the vehicle-week, analyzing the transport activity performed by the vehicle sample in that period. It gathers data about the characteristics of the vehicle, the transported goods, the origin, destination and distance of the operation and, when appropriate, the price of the service. Not all these data are publicly available; the prices of services are specially protected data from this agency.

According to the situation above described, present paper focuses on analysing the demand on the transportation of industrial goods by analysing some data from the PSTGR. In order to gain deep knowledge of such real-life dataset, a standard clustering technique (*k*-means) is applied for descriptive tasks.

The rest of this paper is organized as follows: the k-means clustering technique and associated measurements are described in section 2, the setup of experiments and the dataset under analysis are described in section 3. Finally, the obtained results and the conclusions of present study are stated in section 4.

2 Clustering

Cluster analysis [5], [6] consist in the organization of a collection of data items or patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity. Hence, patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster.

Pattern proximity is usually measured by a distance function defined on pairs of patterns. A variety of distance measures are in use in various communities [7], [8]. There are different approaches to data clustering [5], but in general terms, there are two main types of clustering techniques: hierarchical and partitional approaches. Hierarchical methods produce a nested series of partitions (illustrated on a dendrogram which is a tree diagram) based on a similarity for merging or splitting clusters, while partitional methods identify the partition that optimizes (usually locally) a clustering criterion. Hence, obtaining a hierarchy of clusters can provide more flexibility than other methods. A partition of the data can be obtained from a hierarchy by cutting the tree of clusters at certain level. A representative and well-known technique for partitional clustering is applied in present study, namely *k*-means [9], that is described in subsection 2.1.

As similarity is fundamental to the definition of a cluster, a measure of the similarity is essential to most clustering methods and it must be carefully chosen. Present study applies well-known distance criteria used for examples whose features are all continuous when applying the *k*-means algorithm.

Additionally, different techniques have been applied in present study in order to estimate the optimal value of *k* for the *k*-means algorithm. Such techniques evaluate the goodness of clustering results [10] by taking into account a certain criterion. The following criteria have been applied in present work, for comparison purposes:

- Calinski-Harabasz Index [11]: it evaluates the cluster validity based on the average between-and within-cluster sum of squares. Index measures separation based on the maximum distance between cluster centers, and measures compactness based on the sum of distances between objects and their cluster center.
- Davies-Bouldin Index [12]: it computes the sum of the maximum ratios of the intra-cluster distances to the inter-cluster distances for each cluster.
- Gap criterion [13]: it uses the output of any clustering algorithm, comparing the change in within-cluster dispersion with that expected under an appropriate reference null distribution. This index is especially useful on well-separated clusters and when used with a uniform reference distribution in the principal component orientation.
- Silhouette Index [14]: it validates the clustering performance based on the pairwise difference of between and within cluster distances. In addition, the optimal cluster number is determined by maximizing the value of this index.

2.1 *k*-means Clustering Technique

The well-known *k*-means [9]is a partitional clustering technique for grouping data into a given number of clusters. Its application requires two input parameters: the number of clusters (*k*) and their initial centroids, which can be chosen by the user or obtained through some pre-processing. Each data element is assigned to the nearest group centroid, thereby obtaining the initial composition of the groups. Once these groups are obtained, the centroids are recalculated and a further reallocation is made. The process is repeated until there are no further changes in the centroids. Given the heavy reliance of this method on initial parameters, a good measure of the goodness of the grouping is simply the sum of the proximity Sums of Squared Error (SSE) that it attempts to minimize, Where *p()* is the proximity function, *k* is the number of the groups, c_j are the centroids, and *n* the number of rows:

$$SSE = \sum_{j=1}^{k} \sum_{x \in G_j} \frac{p(x_i, c_j)}{n} \tag{1}$$

In the case of Euclidean distance [15], the expression is equivalent to the global mean square error.

K-means technique takes distance into account to cluster the data. Different distance criteria were defined and the distance measures applied in the study are described in this subsection.

An *mx*-by-*n* data matrix *X*, which is treated as *mx* (1-by-*n*) row vectors x_1, x_2, ...,x_{mx}, and *my*-by-*n* data matrix *Y*, which is treated as *my* (1-by-*n*) row vectors y_1, y_2, ...,y_{my}.. are given. Various distances between the vector x_s and y_t are defined as follows:

Seuclidean distance

In Squared Euclidean metrics (Seuclidean), each coordinate difference between rows in *X* is scaled, by dividing it by the corresponding element of the standard deviation:

$$d_{st}^2 = (x_s - y_t)\, V^{-1} (x_s - y_t)' \tag{2}$$

Where *V* is the *n*-by-*n* diagonal matrix the *j*th diagonal element of which is $S(j)^2$, where *S* is the vector of standard deviations.

Cityblock distance

In this case, each centroid is the component-wise median of the points in that cluster.

$$d_{st} = \sum_{j=1}^{n} |x_{sj} - y_{tj}| \tag{3}$$

Cosine Distance

This distance is defined as one minus the cosine of the included angle between points (treated as vectors). Each centroid is the mean of the points in that cluster, after normalizing those points to unitary Euclidean lengths:

$$d_{st} = 1 - \frac{x_s y_t'}{\sqrt{(x_s x_s')(y_t y_t')}} \tag{4}$$

Correlation Distance

In this case, each centroid is the component-wise mean of the points in that cluster, after centering and normalizing those points to a zero mean and a unit standard deviation.

$$d_{st} = 1 - \frac{\left(x_s - \bar{x}_s\right)\left(y_t - \bar{y}_t\right)'}{\sqrt{\left(x_s - \bar{x}_s\right)\left(x_s - \bar{x}_s\right)'}\sqrt{\left(y_t - \bar{y}_t\right)\left(y_t - \bar{y}_t\right)'}} \tag{5}$$

3 Real-life Case Study on Industrial Road Transportation

As previously stated, in present study data about Road Transportation in the Industrial sector have been analysed through *k*-means. Data were obtained from the 2015 PSTGR, picking those transportation services related to the industrial sector (raw material for industry, mainly), according to the NST 2007 standard goods classification for transport statistics [16]. Among all the services in 2015 PSTGR, only those associated to the following types of goods were included in the dataset:

- 01 - Products of agriculture, hunting, and forestry; fish and other fishing products.
- 02 - Coal and lignite; crude petroleum and natural gas.
- 03 - Metal ores and other mining and quarrying products; peat; uranium and thorium.
- 06 - Wood and products of wood and cork (except furniture); articles of straw and plaiting materials; pulp, paper and paper products; printed matter & recorded media.
- 08 - Chemicals, chemical products, and man-made fibers; rubber and plastic products; nuclear fuel.
- 09 - Other non-metallic mineral products.
- 10 - Basic metals; fabricated metal products, except machinery and equipment
- 11 - Machinery and equipment n.e.c.; office machinery and computers; electrical machinery and apparatus n.e.c.; radio, television and communication equipment and apparatus; medical, precision and optical instruments; watches and clocks.
- 12 - Transport equipment.
- 16 - Equipment and material utilized in the transport of goods.

For the included services, the following features (from 3 different areas) were gathered:

- Time:
  - Day of the week: the day of the week in which the goods were transported (from Monday to Sunday).
- Vehicle:
  - Combination: vehicle combination used in the operation. It takes one of the following values: Lorry, Cab + Body, Cab or Body.
  - Number of axes: it takes values from 1 to 8.
  - Load capacity (in hundreds of kilos).
  - Permissible Maximum Weight (in hundreds of kilos): it includes the vehicle and the load.
- Service:
  - Distance of the operation (in kilometers).
  - Type of goods. It takes one of the values from [16] above listed.

- Type of packaging. It takes one of the following values: liquid in bulk, solid in bulk, big container, other containers, freight package, pre-slinged, motor vehicles and live animals, other mobile units, other kinds of cargos, no load.
- Type of service: It takes one of the following values: normal, delivery and/or pickup, repetitive, normal with no load.
- Stratum: whether the service is public/private and load capacity.
- Contracting party. It takes one of the following values: 1 (contract) or 2 (on its own).
- Geographical range. It takes one of the following values: 0 (within a city), 1 (from one city to another, in the same region), 2 (from one region to another one), 3 (import), 4 (export) or 5 (between other countries, what is called cabotage).
- Weight (in tons): transported weight multiplied by the coefficient of elevation of the statistical representativeness of the sample element. It is an estimation of the real quantity of merchandise transported by the whole market represented by this statistical item.
- Number of operations: number of operations multiplied by the coefficient of elevation of the statistical representativeness of the sample element. It is an estimation of the real quantity of merchandise transported by the whole market represented by this statistical item.

All in all, 4586 samples were gathered, which extrapolates to 3086000 realized transportation services in that year based on their statistical representativeness.

4 Clustering Results

The technique and parameters described in Section 2 were applied to the case study presented in Section 3 and the results are discussed below. Table 2 shows the information on the k estimation for the whole normalized dataset, performed by applying the four different measures. In this table, column ‘Estimated Optimal *k*’ represents the optimum number of clusters calculated by each one of the measures from the initial range (from 1 to 10).

Table 1. k-estimation for the whole dataset.

Measure	***Estimated Optimal k***
Calinski-Harabasz	2
Davies-Bouldin	4
Gap	9
Silhouette	4

Once the cluster evaluation was performed, values of 2 and 4 for *k* paramenter were selected for subsequent experiments.

Table 4 shows the results obtained for the *k*-means, with k-means++ [17] algorithm for cluster center initialization and different distance criteria and values of *k* equal to 2 and 4 (as stated before). In this table, '*k*' is the value of such parameter, 'Distance' is the applied distance criteria, 'Sum Distance' is the within-cluster sums of point-to-centroid distances and 'Mean Silhouette' is the mean of silhouette values for all the points [-1, 1] according to the same distance criteria used by *k*-means.

Table 4. *k*-means clustering results from different experiments.

#	*k*	Distance	Sum Distances	Mean Silhouette
1	2	Seuclidean	$[3.9146 \quad 3.5818] \cdot 10^4$	0.3149
2	2	Cityblock	$[2.2486 \quad 1.7686] \cdot 10^4$	0.3423
3	2	Cosine	$[1.2920 \quad 1.1683] \cdot 10^3$	**0.3436**
4	2	Correlation	$[1.3448 \quad 1.2621] \cdot 10^3$	0.3030
5	4	Seuclidean	$[3.2812 \quad 3.3197 \quad 0.1074 \quad 0.2121] \cdot 10^4$	**0.3314**
6	4	Cityblock	$[0.7196 \quad 1.3503 \quad 1.0884 \quad 0.3393] \cdot 10^4$	0.2943
7	4	Cosine	[509.5208 382.8418 204.9995 855.7570]	0.2879
8	4	Correlation	[819.4047 585.7896 475.0979 195.2167]	0.2521

From the results in table 4, the following experiments were selected for subsequent analysis:

- #3: it is selected as the representative experiments with *k* parameter equal to 2, as it obtained the highest Mean Silhouette value. Samples assigned to each one of the clusters in this experiment: 50.8% to cluster 1 and 49.2% to cluster 2.
- #5: it is selected as the representative experiments with *k* parameter equal to 4, as it obtained the highest Mean Silhouette value. Samples assigned to each one of the clusters in this experiment: 46.1% to cluster, 49.3% to cluster 2, 1.9% to cluster 3, and 2.7% to cluster 4.

In order to visually check the goodness of the clustering, the silhouette graphs of such experiments are shown in Figs. 1 and 2. These graphs depict the silhouette index [14] (horizontal axis) for each one of the data assigned to each one of the clusters (vertical axis).

From Fig. 1 it can be clearly seen that some of the samples assigned to clusters 1 and 2 (those at the top of each one of the clusters in the graph) are clearly different to samples from the other cluster. Although some of the samples from both clusters have a low silhouette value, there is not any sample with a negative silhouette value.

From a thorough analysis of the samples assigned to each one of the two clusters, it has been noticed that samples in cluster 1 are mainly those from lorry combination and low distance of the operation (mean value within the cluster: 53.31 Km). Additionally, the average transported weight (multiplied by the coefficient of elevation) of this cluster is 4.040 tons. On the other hand, samples allocated in cluster 2 are mainly those from cab+body combination and high distance of the operation (mean value within the cluster: 271.37 Km). The average transported weight (multiplied by the coefficient of elevation) of this cluster is 12.946 tons.

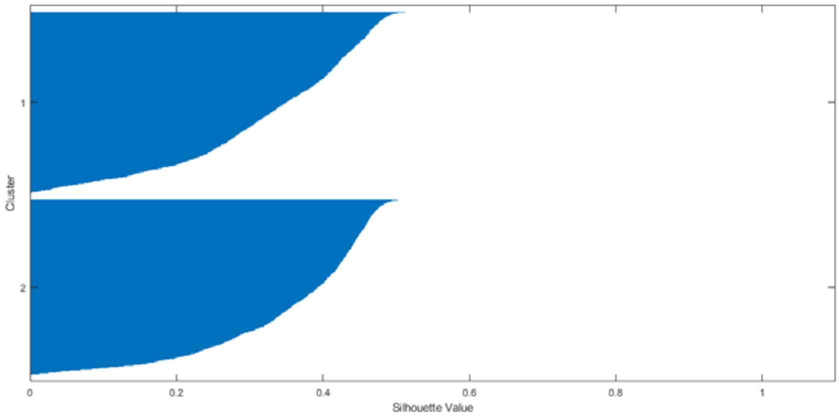

Fig. 1. Silhouette graph generated from clustering results of experiment #3.

From the silhouette graph based on results from experiment #5, it can be said that only very few samples (all of them from cluster 1) have negative silhouette, while most of them have positive values. More precisely, a reduced number of samples were assigned to clusters 3 and 4 but those have the highest values (greater than 0.6 in some case). This graph confirms that 4 is an appropriate number of clusters as higher silhouette values were obtained when compared with those in experiment #3.

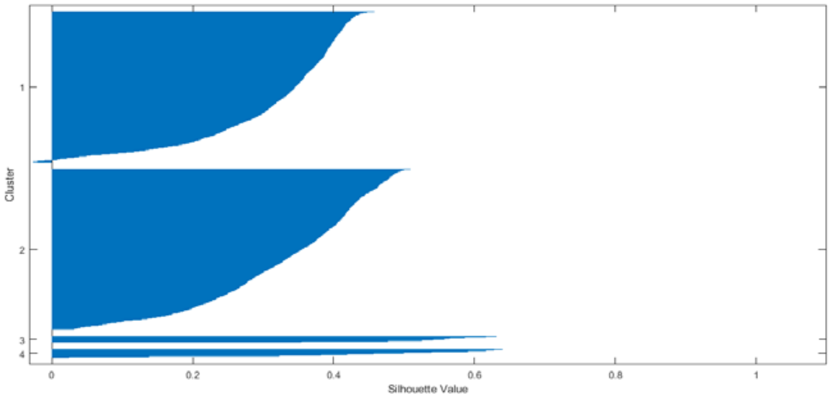

Fig. 2. Silhouette graph generated from clustering results of experiment #5.

As in the case of experiment #3, a thorough analysis of the samples assigned to each one of the two clusters, let us characterize them in the following way:

- Cluster 1: it gathers samples that could be defined as transportation from one province to the other one with heavy loads. Most of them are cab+body combination and medium/low distance of the operation (mean value within the cluster: 183.98 Km). The accumulated transported weight (multiplied by the coefficient of elevation) of this cluster is 29,696,675 tons; that is 76.89% of the total transported weight (whole dataset).
- Cluster 2: it gathers samples that could be defined as transportation within the same province with medium loads. Most of them are lorry combination and low distance of the operation (mean value within the cluster: 51.45 Km). The accumulated transported weight (multiplied by the coefficient of elevation) of this cluster is 7,718,800 tons; that is 19.99% of the total transported weight (whole dataset).
- Cluster 3: it gathers samples that could be defined as non-freelance. Most of them are body combination and medium/low distance of the operation (mean value within the cluster: 209.31 Km). The accumulated transported weight (multiplied by the coefficient of elevation) of this cluster is 258,597 tons; that is 0,67% of the total transported weight (whole dataset).
- Cluster 4: it gathers samples that could be defined as international. Most of them are cab+body combination and very high distance of the operation (mean value within the cluster: 1,688.52 Km). The accumulated transported weight (multiplied by the coefficient of elevation) of this cluster is 945,883 tons; that is 2,45% of the total transported weight (whole dataset).

Once results from the two experiments are compared, it can be concluded that clustering from experiment #5 (k = 4) could be more interesting for the transportation analysis. Clusters 3 and 4 from this experiment gather samples with higher silhouette values than any other samples from the two experiments.

5 Conclusions and Future Work

Main conclusions derived from obtained results are that clustering is a useful tool to gain deep knowledge of a previously unknown dataset. By a standard clustering technique such as *k*-means, clear patterns are identified in a road transportation dataset.

Best results have been obtained when applying k-means with the following parameters: initialization = kmeans++, k = 4, distance = Seuclidean.

More precisely, certain trends have been identified in the demand of transportation services in the Spanish industrial sector during 2015. The majority of the transportation services are associated to short distances (within the same province of from one province to another one), while a small amount of them are associated to an international context.

Conclusions obtained from the analysis of the datasets are useful for companies within transportation sector as knowledge about demand is obtained. Non-evident aggregations are obtained, what could also be interesting for States as well in order to gain abstract knowledge about such an important sector, to identify subjacent typologies of services, and to find emerging services increasingly demanded.

Future work will consist of extending proposed analysis to a wider time period, adding additional features ton the dataset and applying some other clustering techniques for a wider comparison.

References

1. Hisakawa, N., Jankowski, P., Paulus, G.: Mapping the Porosity of International Border to Pedestrian Traffic: a Comparative Data Classification Approach to a Study of the Border Region in Austria, Italy, and Slovenia. Cartography and Geographic Information Science 40, 18-27 (2013)
2. Cinar, D., Gakis, K., Pardalos, P.M.: A 2-phase Constructive Algorithm for Cumulative Vehicle Routing Problems with Limited Duration. Expert Systems with Applications 56, 48-58 (2016)
3. Report from the Commission to the European Parliament and the Council: Fifth Report on Monitoring Development of the Rail Market. European Comission (2016)
4. Council Regulation (EC) No 1172/98 of 25 May 1998 on Statistical Returns in Respect of the Carriage of Goods by Road. Council of the European Union (2000)
5. A.K. Jain, M.N.M., P.J. Flynn: Data Clustering: A Review. ACM Computing Surveys 31, (1999)
6. Xu, R., Wunsch, D.C.: Clustering. Wiley (2009)
7. Andreopoulos, B., An, A., Wang, X., Schroeder, M.: A roadmap of clustering algorithms: finding a match for a biomedical application. Briefings in Bioinformatics 10, 297-314 (2009)
8. Zhuang, W.W., Ye, Y.F., Chen, Y., Li, T.: Ensemble Clustering for Internet Security Applications. Ieee Transactions on Systems Man and Cybernetics Part C-Applications and Reviews 42, 1784-1796 (2012)
9. Ding, C., He, X.: K-means Clustering via Principal Component Analysis. In: Proceedings of the 21st International Conference on Machine learning, pp. 29. ACM, (2004)
10. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J.: Understanding of Internal Clustering Validation Measures. In: IEEE 10th International Conference on Data Mining (ICDM), pp. 911-916. IEEE, (2010)
11. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics-theory and Methods 3, 1-27 (1974)
12. Davies, D.L., Bouldin, D.W.: A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 224-227 (1979)
13. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63, 411-423 (2001)
14. Rousseeuw, P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53-65 (1987)
15. Danielsson, P.-E.: Euclidean distance mapping. Computer Graphics and Image Processing 14, 227-248 (1980)
16. Standard goods classification for transport statistics (NST 2007), http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Standard_goods_classification_for_transport_statistics_(NST), last accessed 09/03/2018.
17. Arthur, D., Vassilvitskii, S.: k-means++: The Advantages of Careful Seeding. In: Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027-1035. Society for Industrial and Applied Mathematics, (2007)