applied sciences

MDPI

Article

# Analysis of Madrid Metro Network: From Structural to HJ-Biplot Perspective

**E. Frutos Bernal [1,*], A. Martín del Rey [2] and P. Galindo Villardón [1]**

[1] Department of Statistics, University of Salamanca, 37007 Salamanca, Spain; pgalindo@usal.es
[2] Department of Applied Mathematics, University of Salamanca, Institute of Fundamental Physics and Mathematics, 37008 Salamanca, Spain; delrey@usal.es
* Correspondence: efb@usal.es

Received: 23 July 2020; Accepted: 14 August 2020; Published: 17 August 2020


**Abstract:** With the growth of cities, urban traffic has increased and traffic congestion has become a serious problem. Due to their characteristics, metro systems are one of the most used public transportation networks in big cities. So, optimization and planning of metro networks are challenges which governments must focus on. The objective of this study was to analyze Madrid metro network using graph theory. Through complex network theory, the main structural and topological properties of the network as well as robustness characteristics were obtained. Furthermore, to inspect these results, multivariate analysis techniques were employed, specifically HJ-Biplot. This analysis tool allowed us to explore relationships between centrality measures and to classify stations according to their centrality. Therefore, it is a multidisciplinary study that includes network analysis and multivariate analysis. The study found that closeness and eccentricity were strongly negatively correlated. In addition, the most central stations were those located in the city center, that is, there is a relationship between centrality and geographic location. In terms of robustness, a highly agglomerated community structure was found.

**Keywords:** subway networks; complex network analysis; HJ-Biplot; cluster analysis; multivariate statistical analysis; madrid metro network

## 1. Introduction

The role that public transport plays in the growth of cities is vital in ensuring their sustainable development. It provides a effective way to reduce congestion, air pollution and improve the quality of life of citizens. To give service to the ever-increasing ridership demand, governments and researchers focus on services' optimization and infrastructures [1].

To achieve this goal, it seems very important to determine central stations to improve their connectivity and strengthen their security against attacks or disruptions. In order to evaluate the importance of stations in transit networks, complex network analysis has been used in several works. More specifically, different indicators such as centrality measures, network diameter and connectivity are employed to analyze transport networks such as bus [2–5], metro [4,6,7], rail [8] and air transport networks [9]. Centrality measures such as node degree, betweenness centrality, closeness centrality and eigenvector centrality are employed to analyze the structural importance of stations in the network. To evaluate network's connectivity, the network indicators more frequently used are clustering coefficient and average path length. In addition, there are two desirable properties in transport networks which are small-word and scale-free properties. Networks which exhibit these properties are able to manage congestion and are robust to random attacks.

Moreover, complex network analysis has been applied in other study areas, such as biology and epidemiology [10,11], social networks [12], power grids [13] and vehicular sensing networks (VSN) in smart cities [14].

In recent years, the study of transport networks has become a field of interest for researchers. Some of them focus on the study of resilience of transport networks: for example, in [15] an analysis of resilience of the London metro system is developed using clustering coefficient, path length, passenger strength, modularity and assortativity indicators. In [16] the robustness of two public transport networks which exhibit different properties is analyzed. In other cases, some specific indicators are studied; this is the case of [17] where the influence of network size on centrality measures is analyzed. Furthermore, correlation between centrality measures has been studied in different types of networks, such as scientific collaboration networks, airline networks and internet routing networks (see, for example [18–20]) finding that the correlations between centrality measures are different from one type of network to another.

This paper analyzes Madrid metro network using the most important centrality measures, some structural coefficients and robustness indicators. Once the results of the network analysis were obtained, they were inspected using multivariate analysis techniques. Specifically, the HJ-Biplot [21] was employed to analyze correlation between centrality measures and also to classify stations according to their centrality.

The main advantage of using HJ-Biplot is that we can interpret simultaneously the position of the variables (centrality measures) represented by vectors and the individuals (stations) represented by points and, also, the relationships between them. Furthermore, the coordinates obtained in the HJ-Biplot were used to calculate cluster of stations.

In Biplot methods, no parametrical assumptions are considered, also they have the advantage of being a specific statistical tool to display multivariate data. This technique has been applied to analyze data from different knowledge fields (see, for example [22–25]) but never to analyze transport networks. So, as far as we know, there is no study with these characteristics (multidisciplinary) of the Madrid metro network.

In summary, the principal purpose of this study is to analyze the most important centrality measures, structural coefficients and robustness indicators of Madrid metro network using multivariate techniques, which will allows us to analyze relationships between centrality measures, to evaluate stations according to their values in centrality measures, to classify the stations into disjoint cluster and finally to study the relationship between the clusters obtained and the passenger flow.

The rest of the paper is organized in five sections. Section 2 deals with the basics of complex network analysis. Section 3 describes the study area and data. In Section 4 the methodology used in this work is introduced. Section 5 describes a statistical analysis of the Madrid metro. Finally, in Section 6 the conclusions and further work are presented.

## 2. Complex Network Analysis Applied to Subway Networks

In this study we use graph theory to represent Madrid metro network. Specifically, we use L-space representation of the network, thus the nodes of the graph correspond to the stations of the metro and the edges of the graph are the tracks which connect two consecutive stations. Therefore, we have an undirected graph $G = (V, E)$, where $V = \{v_1, v_2, ..., v_N\}$ is the set of nodes, and $E = \{e_{ij} = (v_i, v_j), v_i, v_j \in V\} \subset V \times V$ is the set of edges. The adjacency matrix $A = (a_{ij})$ is an $N \times N$ symmetric and non negative matrix whose element $a_{ij}$ is defined as follows:

$$a_{ij} = \begin{cases} 1, & \text{if } (v_i, v_j) \in V \\ 0, & \text{if } (v_i, v_j) \notin V \end{cases} \quad (1)$$

In Figure 1 the graph corresponding to Madrid metro network is shown. Note that, for the sake of simplicity and because this is not relevant for the structural and topological analysis of the metro network, the real spatial location of the stations is not considered.

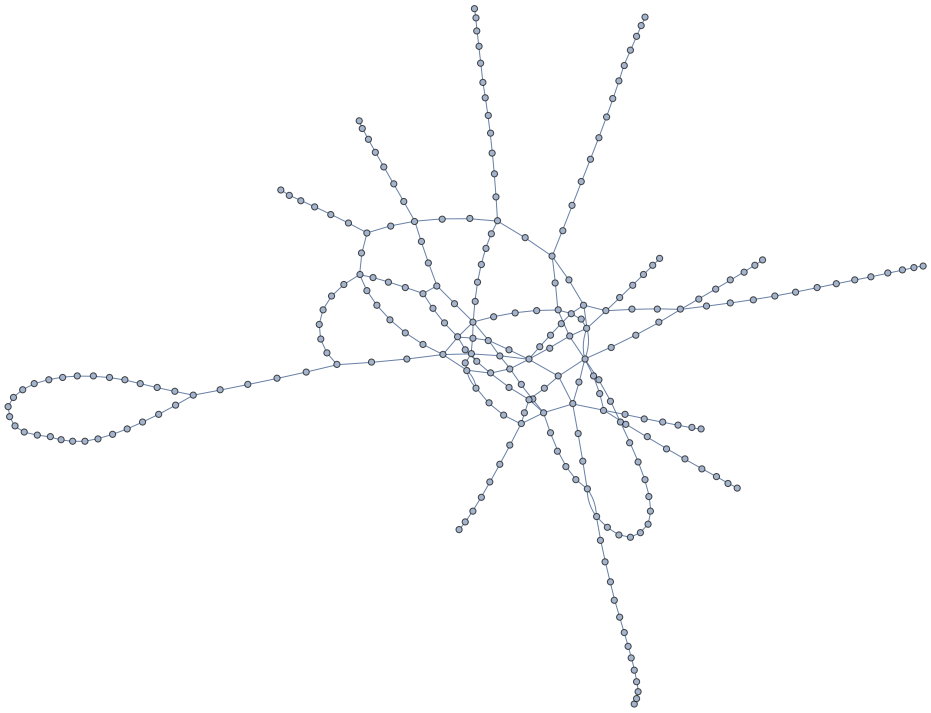

**Figure 1.** Madrid metro network computed using Mathematica.

### *2.1. Network Parameters*

Network parameters consist of a set of metrics which capture information on network connectivity and accessibility. These parameters can be defined as structural coefficients since they are invariant under graph isomorphisms; that is, if the network $G$ is isomorphic to the network $H$ (there exists a bijection $\phi$ between their respective node sets such that $(v_i, v_j)$ is an edge of $G$ if and only if $(\phi(v_i), \phi(v_j))$ is an edge of $H$) then the node structural coefficient $C\colon V \to \mathbb{R}$ satisfies the following: $C(v) = C(\phi(v))$. Moreover, node centrality measures are structural coefficients that take values in $[0, 1]$ whose aim is to quantify the relative importance or influence of each node within $G$. Note that all this parameters are local (they are associated to nodes and, in some cases, to edges) and the most important are degree centrality, closeness centrality, betweenness centrality, eccentricity, clustering coefficient and eigenvector centrality. Nevertheless, also global coefficients (in the sense that they are associated to the whole network) can be computed: average path length, diameter, etc. (see [26]).

In this subsection the definition of these topological parameters is given.

The *degree* of a node $v_i$ is one of the most important parameters in network analysis. It is defined as the total number of edges incident to $v_i$: $k_i = \sum_{j=1}^{N} a_{ij}$. Two local structural coefficients can be easily determined from it: (1) the *degree centrality* that measures the average number of nearest neighbors that a node $v_i$ is connected to, that is:

$$C_D(v_i) = \frac{k_i}{N-1}, \quad 0 \leq C_D(v_i) \leq 1; \tag{2}$$

and (2) the *importance* of $v_i$:

$$I(v_i) = \frac{k_i}{\sum_{i=1}^{N} k_i}, \quad 0 \leq I(v_i) \leq 1, \tag{3}$$

that is used to measure the level of "disorder" of the network. Moreover, also global structural coefficients can be defined. For example, the *average degree* of the network is given by the following formula:

$$\langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{1}{N}\sum_{i,j=1}^{N} a_{ij} = \frac{N-1}{N}\sum_{i=1}^{N} C_D\left(v_i\right). \tag{4}$$

Note that $I\left(v_i\right) = \frac{k_i}{N\langle k \rangle}$ for each $i$. Moreover, the *standard entropy* of the network can be defined as follows:

$$E = -\sum_{i=1}^{N} I\left(v_i\right)\log\left(I\left(v_i\right)\right). \tag{5}$$

As is shown, degree centrality "computes" in some way the "importance" of a node by simply considering the number of neighbors. Nevertheless, such characteristic (the 'status', 'prestige', or 'importance') can be measured in a more sophisticated way taking into account not only a quantitative perspective (the total number of connections a node has) but also a qualitative perspective (the number of neighbor nodes with highest central coefficients). This is precisely measured by the *eigenvector centrality*. The idea is that a node is important not only because it connects with many nodes but also because the nodes it connects to are also important [27]. Mathematically it is defined as the principal eigenvector, $C_G = \left(C_G\left(v_1\right), \ldots, C_G\left(v_N\right)\right)$, of the adjacency matrix. That is, the eigenvector associated to the largest eigenvalue of $A$: $\lambda$. The $i$-th component of $C_G$ stands for the relative eigen centrality score of node $v_i$ and the following holds:

$$C_G\left(v_i\right) = \lambda^{-1}\sum_{j=1}^{N} a_{ij} C_G\left(v_j\right). \tag{6}$$

As a metro network can be considered as connected graph, then its adjacency matrix is irreducible and its largest eigenvalue is positive (by applying the Perron–Frobenius Theorem). As a consequence $0 \leq C_G\left(v_i\right) \leq 1$.

The *distance* between two nodes $v_i, v_j \in V$, $d\left(v_i, v_j\right)$, is defined as the total number of edges that link them through the shortest path. This notion is invariant under isomorphisms and leads to both local and global structural indices. In this sense the *network diameter* $D$ is the longest distance between any pair of nodes in the network:

$$D = \max\{d\left(v_i, v_j\right), 1 \leq i < j \leq N\}. \tag{7}$$

The *eccentricity* of the node $v_i$ is defined as the maximum distance from $v_i$ to any other node in the network:

$$C_e\left(v_i\right) = \frac{1}{D}\max\{d\left(v_i, v_j\right), 1 \leq j \leq N, j \neq i\}, \quad 0 \leq C_e\left(v_i\right) \leq 1, \tag{8}$$

and the *radius* of the network, $R$, is defined as the minimal eccentricity. Furthermore, the *average path length* of the network is defined as the average distance between every pair of nodes.

$$L = \frac{1}{N\left(N-1\right)}\sum_{1 \leq i < j \leq N} d\left(v_i, v_j\right). \tag{9}$$

The *closeness centrality* of node $v_i$ is defined as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. Specifically it is as follows:

$$C_{CL}\left(v_i\right) = \frac{1}{\sum_{l=1, l \neq i}^{N} d\left(v_i, v_l\right)}, \quad 0 \leq C_{CL}\left(v_i\right) \leq 1. \tag{10}$$

This local structural coefficient measures the mean distance from the node to the rest of nodes of the network. Note that, the more central a node is, the greater the value of closeness centrality is.

The *betweenness centrality* of the node $v_i$ measures the "presence" of the node in the shortest paths between another two nodes. Mathematically it is defined as follows:

$$C_B(v_i) = \sum_{\substack{1 \le r < s \le N \\ r \ne i, s \ne i}} \frac{\ell_{rs}(v_i)}{\ell_{rs}}, \quad 0 \le C_B(v_i) \le 1, \tag{11}$$

where $\ell_{rs}$ is the total number of shortest paths from $v_r$ to $v_s$, and $\ell_{rs}(v_i)$ is the number of those paths that pass through $v_i$.

The *clustering coefficient* of node $v_i$ measures how well the neighbors of a node are connected to each other. It is depicted as follows:

$$C_{CLU}(v_i) = \frac{2\epsilon_i}{k_i(k_i - 1)}, \quad 0 \le C_{CLU}(v_i) \le 1, \tag{12}$$

where $\epsilon_i$ is the number of links between the neighbors of the node $v_i$. Note that the larger the clustering coefficient is, the better the local connectivity around $v_i$ is. A global structural coefficient called *average clustering coefficient* can be obtained by averaging local clustering index over all the nodes in the network:

$$C_{CLU} = \frac{1}{N} \sum_{i=1}^{N} C_{CLU}(v_i). \tag{13}$$

### 2.2. Robustness

Robustness or resilience of a network can be defined as its ability to resist random failures or deliberate attacks and its capacity to solve them and return to a state of equilibrium [15,28]. There are several metrics to evaluate the robustness of a network: robustness indicator, effective resistance, effective conductance, average efficiency, algebraic connectivity, etc. (see [29]). Of special interest in this type of studies are the modularity and the assortativity correlation that are described as follows.

It is important to analyze the community structure of the networks in order to identify cascading impacts resulting from disruptions [4]. A *community* is a subgraph of a network in such a way that within the nodes there is high probability of being connected to each other but between communities there are fewer links [30]. The computation of communities involves processes with a high computational load; several detection and partitioning schemes have been proposed: CNM scheme, BGLL scheme, etc. [31] A commonly used metric for evaluating the quality of the network partition into communities is given by a numerical index called *modularity* [30,32,33].

Suppose the network is partitioned into $g$ communities, and let $\epsilon_{ij}$ be the fraction of edges in the original network that connect vertices between $i$-th community and $j$-th community. The *modularity index* is defined as:

$$Q = \sum_{1 \le i < j \le g} \epsilon_{ij} - \sum_{1 \le i < j < k \le g} \epsilon_{ij}\epsilon_{ki}, \quad 0 \le Q \le 1. \tag{14}$$

The greater $Q$ the stronger the connection within the communities is.

The *assortativity correlation* studies to what extent nodes link to nodes of similar or dissimilar nodal degree. Depending on the degree correlation, a network can be *assortative*, when nodes with height (resp. low-degree) connect with high (resp. low-degree) nodes, or *dissassortive*, when connections

happen between high and low degree nodes. The assortativity correlation can be measured by means of the Pearson correlation coefficient [34] as follows:

$$r = \frac{M^{-1}\sum_{e_{ij}\in E} k_i k_j - \left[M^{-1}\sum_{e_{ij}\in E}\frac{k_i+k_j}{2}\right]^2}{M^{-1}\sum_{e_{ij}\in E}\frac{k_i^2+k_j^2}{2} - \left[M^{-1}\sum_{e_{ij}\in E}\frac{k_i+k_j}{2}\right]^2}, \quad (15)$$

where $k_i$ and $k_j$ are the degrees of the nodes connected by the edge $e_{ij}$ and $M$ is the total number of edges. This coefficient ranges from $-1$ (disassortative network) to 1 (assortative network). It is important to note that in assortative networks failures of high-degree stations will have little impact since other high-degree stations are still connected to other high-degree stations [35].

## 3. Study Area and Data

Madrid metro is a metropolitan railway network which gives service to the Spanish city of Madrid and its metropolitan area (see Figure 2). Madrid metro consists of 13 lines (not counting Light Rail 2 and 3) and 302 stations. It is the fifth largest network in the world in number of stations, behind the London, New York, Shanghai and Paris metro. In addition, it has a 294 km of network, which makes Madrid metro the eighth largest network behind the subway in New York, London, Beijing, Guangzhou, Seoul, Shanghai and Moscow [36]. In 2019, more than 677.4 million people used Madrid metro [36].

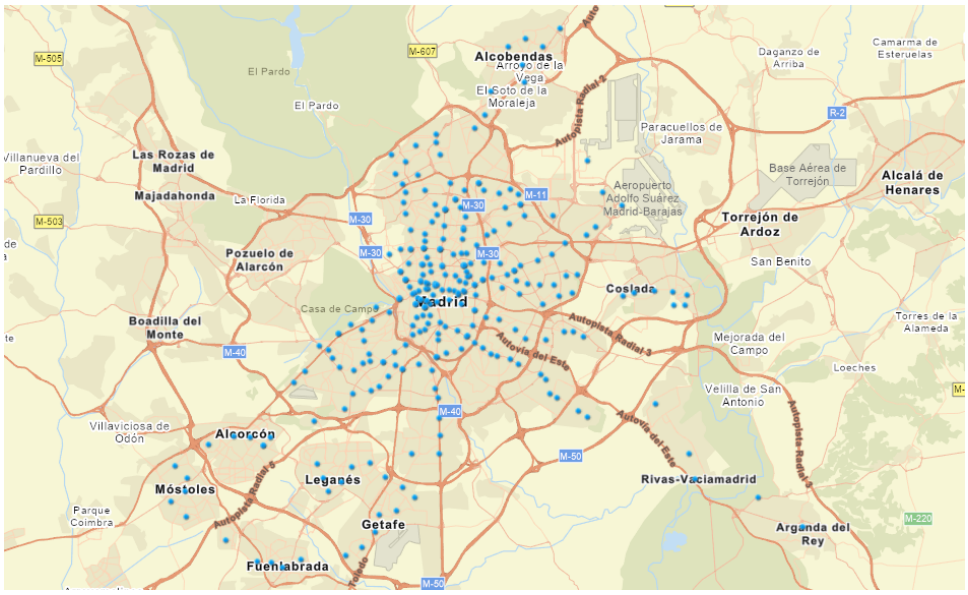


**Figure 2.** Spatial distribution of Madrid metro stations in 2020 (from http://opendata.esri.es/).

In this work, the L-space topology was used to represent the metro so that stations correspond to the nodes of the graph and the links between stations correspond to the edges of the graph.

Firstly, the most usual coefficients used in the Complex Network Analysis were computed [29]. Furthermore, 2019 boarding data were taken from the Madrid subway website [37].

So that, for the statistical analysis, a matrix with 243 rows which correspond to the Madrid subway stations considered in this study and 7 columns is developed. These columns correspond to the six centrality measures used (degree centrality, betweenness centrality, eccentricity, closeness, clustering coefficient and the eigenvector centrality) and the 2019 boarding data.

## 4. Methodology

Once network parameters were calculated, a statistical analysis of them was performed. The methodology used in this work consists of three parts (see Figure 3). First, the associations between the centrality measures were searched and analyzed. For that, we used the HJ-Biplot technique which allows one to interpret simultaneously the position of rows and columns of a data matrix in a low-dimensional subspace. After that, a cluster analysis was performed to classify metro stations into homogeneous groups. The Biplot coordinates were used to calculate clusters (K-means method, Euclidean distance). Lastly, the relationship between these groups and passenger flow, was identified. For that, an analysis of variance (ANOVA) was used.

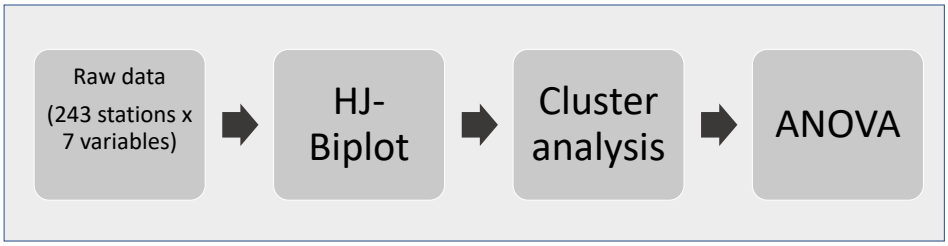


**Figure 3.** Flowchart of the approach.

### 4.1. Hj-Biplot

To find the relationship between the centrality measures, the HJ-Biplot technique has been used. The HJ-Biplot is an exploratory data analysis method thus, no parametric assumptions are considered.

A Biplot is a joint representation in a low dimensional Euclidean space (usually $\mathbb{R}^2$) of the rows and columns of a matrix $X = \left(x_{ij}\right)_{1\leq i\leq n, 1\leq j\leq p} \in \mathcal{M}_{n\times p}\left(\mathbb{R}\right)$, using vectors as points called markers, $a_1, \ldots, a_n$ for its rows, and vectors called markers, $b_1, \ldots, b_p$ for its columns, so that the scalar product $a_i \bullet b_j$ defines the element $x_{ij}$ of the matrix $X$ [38,39]. These markers are obtained from the singular value decomposition (SVD) of the data matrix $X = UDV^T$, where the column vectors of $U$ stand for the eigenvectors of $X \cdot X^T$, the column vectors of $V$ are the eigenvectors of $X^T \cdot X$, and $D$ is a diagonal matrix defined by the singular values.

The HJ-Biplot is an extension of the classical Biplots introduced by Gabriel [38]: the JK-Biplot and the GH-Biplot. The JK-Biplot is appropriate for visualizing the similarity/dissimilarity among row factors because it preserves row metric, while the GH-Biplot is appropriate for analyzing the relationships among column factors because it is column-metric preserving. In the HJ-Biplot the coordinates for columns coincide with the column markers in the GH-Biplot and the coordinates for the rows coincide with the row markers in the JK-Biplot, so that, both markers can be represented in the same reference system, achieving a good quality of representation for both rows and columns.

The HJ biplot is conceptually similar to the correspondence analysis, but it applies to continuous data rather than to categorical data.

The rules for the interpretation of the HJ-Biplot are a combination of those used in other multidimensional scaling techniques such as classical Biplots, correspondence analysis and factor analysis. Specifically:

(a) The distances among row markers are interpreted as an inverse function of similarities, in a such a way that closer markers (stations) are more similar; this property is used to identify the clusters of stations with similar profiles.
(b) The lengths of the column markers (vectors) approximate the SD of the centrality measures.
(c) The cosines of the angles between the column vectors approximate the correlations among the centrality measures. So that, acute angles indicate high positive correlation, obtuse angles indicate negative correlation and finally right angles indicate that the variables are not correlated.

(d) The order of the orthogonal projections of the row markers (points) onto a column marker (vector) approximates the order of the row elements (values) in that column so, the projection of a point (station) away from the center of gravity (average coordinate point), the value that this station takes on the variable is far from its mean (see Figure 4).

Data were analyzed using the software MultBiplot [40].

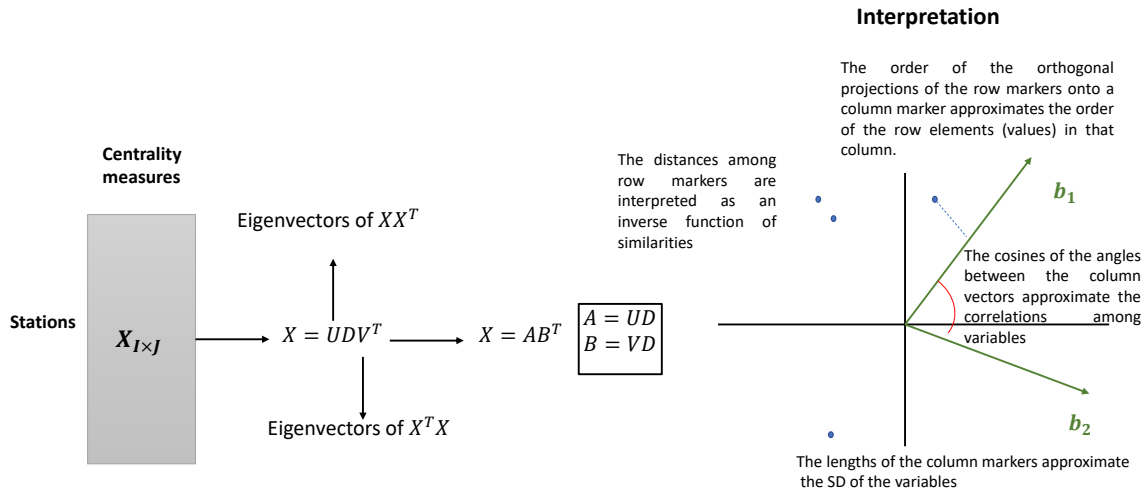


**Figure 4.** Phases of HJ-Biplot process.

### *4.2. Cluster Analysis*

A cluster analysis has been employed to identify group membership for each station in this study. Cluster analysis is a technique to group similar individuals into a number of cluster based on the observed values of some variables for each individual. In our case, the clusters have been calculated through the Biplot-coordinates (K-means method, Eucliden distance). K-means [41,42] is a clustering algorithm very popular for cluster analysis in data mining. K-means is an iterative algorithm that tries to partition the data set into $K$ predefined disjoint clusters in such a way that each observation belongs to only one group. The algorithm consists of the following steps:

(1) To fix the number $K$ of cluster to be used.
(2) To initialize the $K$ centroids.
(3) Each station is assigned to the cluster with the nearest centroid. In our case, the distance between the stations and the centroids is calculated using the Euclidean distance.
(4) To recalculate the centroids for each cluster.
(5) To repeat the process until the assignments no longer change.

### *4.3. ANOVA*

Analysis of variance (ANOVA) is a statistical technique which is used to check if the means of two or more groups are significantly different from each other. To do this, it uses the F-distribution, so that the null hypothesis is that the means are equal. Therefore, a significant result means that the means are unequal. In our case, we were interested in knowing if there are significant differences in passenger flow between the different clusters.

## 5. Results

### *5.1. Topological Analysis of Madrid Metro Network*

Firstly, network parameters introduced in Section 2.1 were computed. Madrid metro network consists of 243 nodes connected by 280 edges. The diameter of the network is $D = 44$, therefore the longest distance between any par of nodes is 44, and the average path length is $L_{avg} = 14.682$, which means people can reach their destination by on average traveling 14–15 stations. The radius of the network is 22 and the centre is Príncipe Pío station, that means that the maximum distance from this station to any other on the network is minimal. Moreover, the average clustering coefficient is $\widetilde{C}_{CLU} = 0.008$. The most important global network indicators are summarized in Table 1.

**Table 1.** Summary statistics of network indicators.

| Network Indicator | Value |
|---|---|
| Number of stations, $N$ | 243 |
| Number of links, $M$ | 280 |
| Network diameter, $D$ | 44 |
| Radius, $r(G)$ | 22 |
| Average shortest path, $L$ | 14.682 |
| Clustering coefficient, $\widetilde{C}_{CLU}$ | 0.008 |
| Entropy, $E$ | 5.431 |
| Modularity, $Q$ | 0.875 |
| Number of communities | 15 |
| Assortativity | 0.296 |

To inspect and analyze network parameters, we applied the HJ-Biplot to our $243 \times 7$ data matrix, using only centrality measures variables. The first three axes of the HJ-Biplot analysis explain 86.27% of data variability, achieving 54.406% in the first principal plane (Table 2). From the results shown in Table 2, it can be deduced that there is a dominant axis (axis 1) that takes 54.406% of the total inertia of the system. This axis is characterized by the variables degree centrality, eccentricity, closeness and eigenvector centrality (Table 3).

**Table 2.** Eigenvalues and explained variance.

| Axis | Eigenvalue | Explained var. | Cummulative var. |
|---|---|---|---|
| Axis 1 | 789.976 | 54.406 | 54.406 |
| Axis 2 | 263.913 | 18.176 | 72.582 |
| Axis 3 | 198.745 | 13.688 | 86.27 |

In Table 3 the contribution of each factor to the element is shown. The variables that characterize axis 1 (variables that receive a strong contribution of axis 1 and low in axis 2 and axis 3) are: degree centrality, eccentricity, closeness and eigenvector centrality. The variable that characterizes axis 2 is clustering and finally the one that characterizes axis 3 is betweenness.

**Table 3.** Relative contribution of the q-th factor to column element j.

| Variables | Axis1 | Axis2 | Axis3 |
|---|---|---|---|
| Degree centrality | 688 | 65 | 48 |
| Eccentricity | 567 | 353 | 44 |
| Clustering | 346 | 317 | 169 |
| Closeness | 718 | 196 | 46 |
| Betweenness | 412 | 0 | 508 |
| Eigenvector centrality | 534 | 159 | 6 |

The factorial graph of the plane defined by axis 1 and 2 is shown in Figure 5. The variables with the greatest variability on the first axis are degree centrality, eccentricity, closeness and eigenvector centrality (its associated vectors have the highest length). Degree centrality, closeness and eigenvector centrality characterize the positive side of axis 1 while eccentricity characterizes the negative side. So, the stations located on the left side of the factorial plane are characterized by having high eccentricity values (they project far away in the direction of the vector eccentricity). For instance, stations Arroyo Culebro, Henares, Las Suertes have the highest values in eccentricity. In contrast, the ones on the right side have very low values in eccentricity , this is the case of stations like Puerta del Ángel and Noviciado.

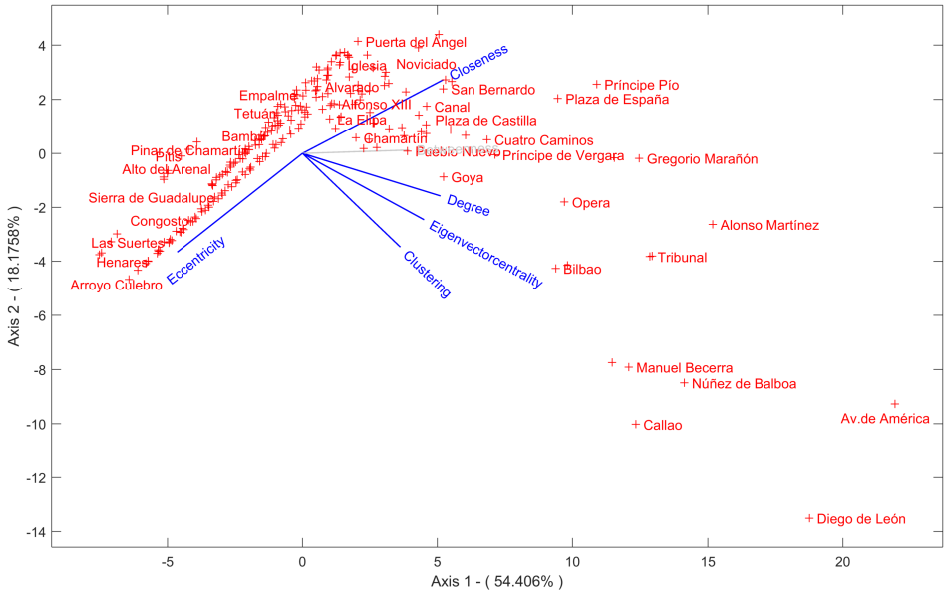


**Figure 5.** HJ Biplot representation on the main first plane.

In addition, those that are located in the first quadrant have high values in closeness because they project far away in the direction of the vector closeness. This is the case of stations like Plaza de España, Príncipe Pío, Noviciado and San Bernardo. Finally, the stations that are located on the right side of the fourth quadrant are those with the highest values in clustering, degree, closeness and eigenvector centrality. Examples of these stations are: Callao, Diego de León, Tribunal, Av. de América and Alonso Martínez.

This analysis also shows that closeness and eccentricity are strongly negatively correlated (flat angle). In addition, there is strong correlation between degree, clustering and eigenvector centrality (acute angles).

It is interesting to analyze the plane defined by axis 1 and 3 to identify the stations with the highest betweenness value (see Figure 6), since in the plane 3 the representation of betweenness is optimal (Table 3). Thus, stations with the highest values in betweenness are: Av. de América, Alonso Martínez, Príncipe Pío and Gregorio Marañon.

To perform the cluster analysis, the coordinates obtained from the HJ-Biplot were used. The algorithm used was the K-means and to define similarity among different objects the Euclidean metric was chosen. Three clusters were formed with the different stations (see Figure 7).

These clusters can be characterized as follows:

- Cluster 1 (C1): It consist of 121 stations. Most of them with degree centrality equal to 2 and some of them with greater degree centrality. They have medium values of eccentricity, closeness, betweenness and eigenvector centrality. The clustering coefficient for all of them is 0.

- Cluster 2 (C2): It is formed by 106 stations. They have degree 1 or 2, the clustering coefficient is 0 and also they have the lowest of betweenness, closeness and eigenvector centrality of the subway network. However, the stations with the highest values of eccentricity are in Cluster 2.
- Cluster 3 (C3): It is formed by 16 stations, all of them belonging to the city center. These stations have the highest values of degree, clustering, closeness, betweenness and eigenvector centrality. In addition, they have the lowest values for eccentricity.

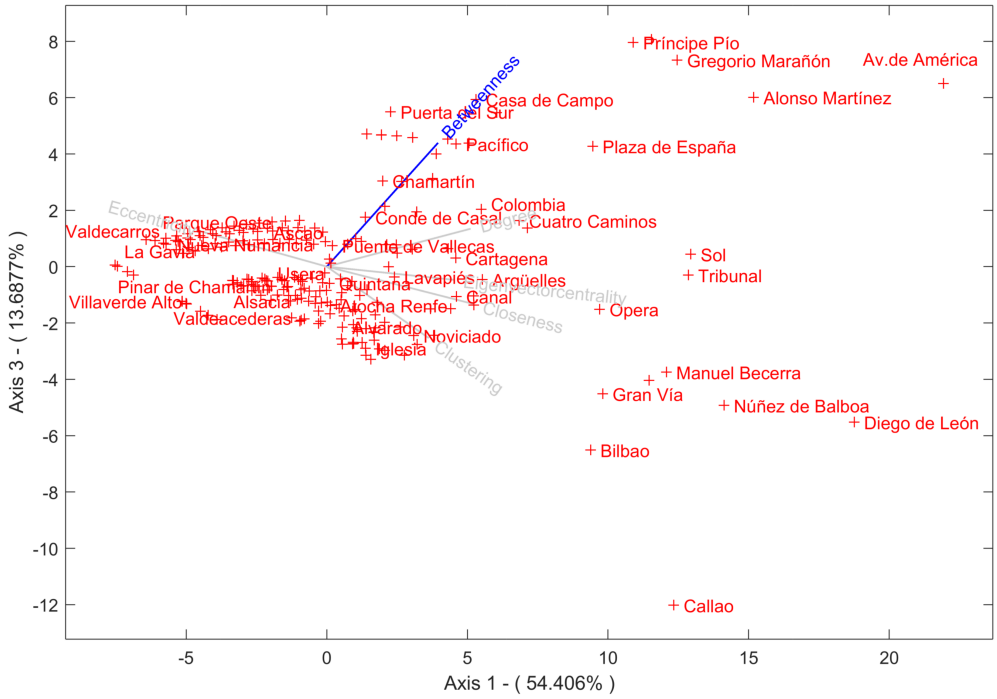


**Figure 6.** HJ Biplot representation on the plane 1–3.

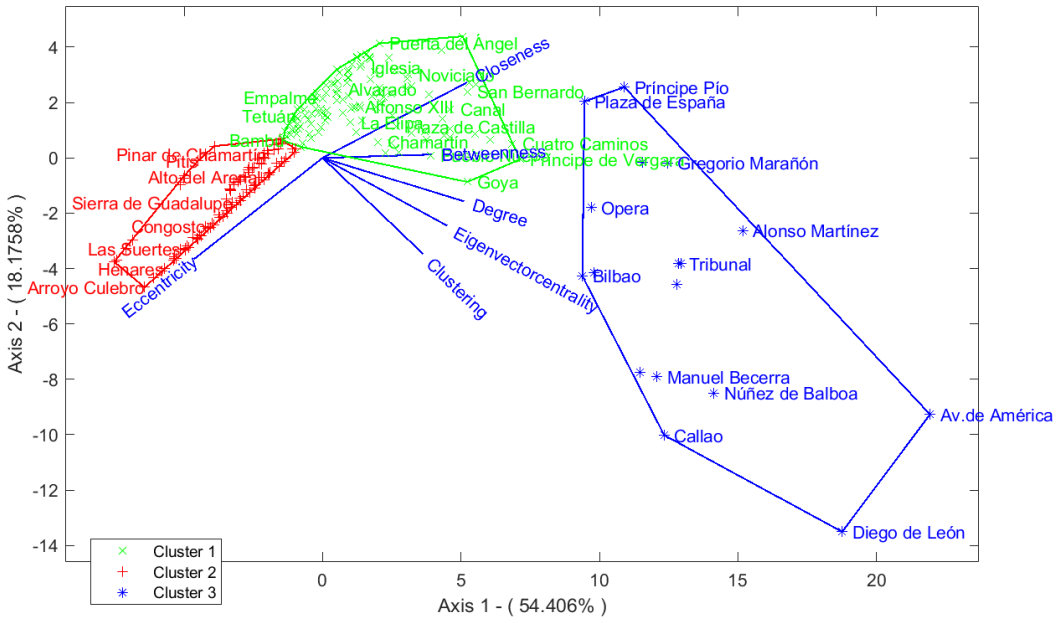


**Figure 7.** Factorial representation of the HJ-Biplot for clusters, plane 1–2.

Figure 8 shows different bloxplots for each centrality measure and for each cluster. It can be seen how the stations in cluster 3 have the highest values of degree, clustering, closeness, betweeeness and eigenvector centrality while stations in cluster 2 have the highest values of eccentricity.

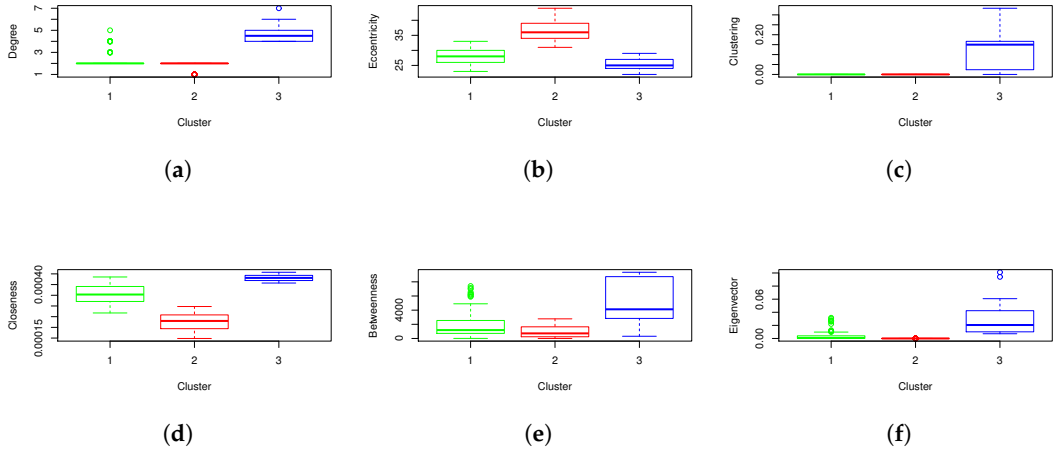


**Figure 8.** Centrality measures boxplots. (**a**) Degree; (**b**) Eccentricity; (**c**) Clustering; (**d**) Closeness; (**e**) Betweenness; (**f**) Eigenvector Centrality.

The spatial distribution of clusters is shown in Figure 9. Specifically, C3 contains the main central district, C1 is located beside C3 area and C2 is located on the edge of the city.

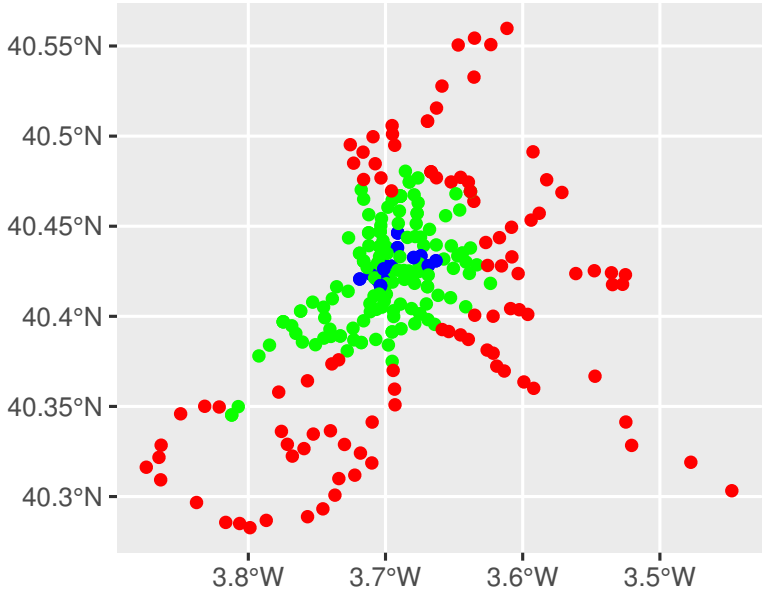


**Figure 9.** Spatial distribution of clusters. In red the Cluster 2 stations, in green the Cluster 1 stations and in blue the Cluster 3 stations.

One way ANOVA was implemented to assess whether there are significant differences in the passenger flow between the different clusters. The results referring to the ANOVA analysis are presented in Table 4.

**Table 4.** ANOVA analysis of the flow passengers.

| | Mean | Standard Deviation |
|---|---|---|
| Global | 2,882,126.11 | 3,080,341.64 |
| Cluster 1 | 3,299,972.89 | 2,715,313.06 |
| Cluster 2 | 1,518,573.90 | 790,653.57 |
| Cluster 3 | 8,981,036.46 | 6,606,708.17 |

The global mean of passengers per station in 2019 is 2, 882, 126.11. There are significant differences between the clusters ($F = 52.19$, $p = 0.00$). In this way, the highest average boarding count is obtained in Cluster 3 (8,981,036.46 passengers) followed by Cluster 1 (3,299,972.89 passengers) and finally Cluster 2 (1,518,573.90 passengers). This implies that the stations in the city center have on average a higher number of passengers than those that surround them, and these ones have on average more passengers than the suburbs.

To analyze which centrality measures are more related to the number of passengers, the Pearson's correlation is calculated. Pearson's correlation coefficients of the centrality measures and boarding count are shown in Figure 10. From the figure, it can be deduced that there is a positive correlation between degree centrality and boarding count. For the rest of the centrality measures the correlation is low, and in the case of eccentricity is negative. This figure also confirms the relationships found between the centrality measures using the HJ-Biplot.

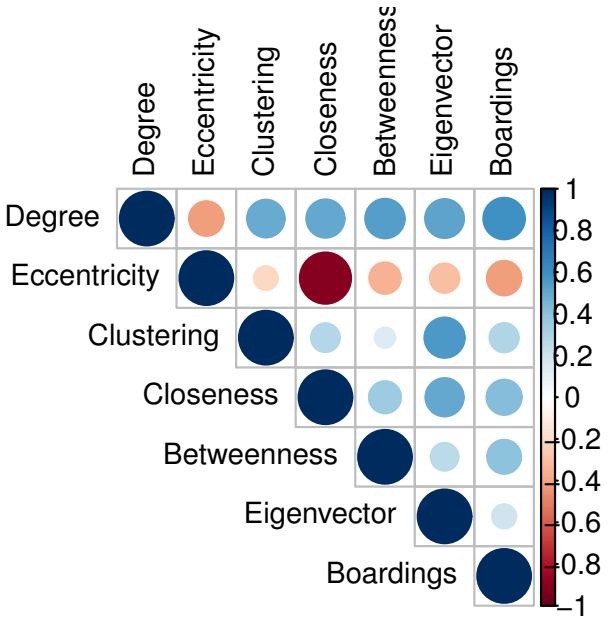


**Figure 10.** Pearson's correlation coefficients of the centrality measures and boarding count.

### *5.2. Analysis of Vulnerability*

A basic study on the robustness of Madrid metro network was presented in [29]. In addition, in this work we analyzed the communities that exist within Madrid metro network using the concept of modularity. Transit operation zones in transport networks can be identified through communities [5]. In transport networks with high modularity, stations in the same community have dense connections but sparse connections between stations in different communities. This implies that flow of passengers is more efficient within each community than between communities.

As many as 15 communities were identified in Madrid metro network (see Figure 11). The modularity is high ($Q = 0.875$); it results in Madrid metro network being robust against breakdowns.

In Table 5, the distribution of stations per community and line is shown. The excessive size of some lines in Madrid metro network leads to their fragmentation in various communities. For example,

note that Lines 9 and 12 are clearly divided between two communities, and stations of Line 5 are approximately equally distributed between four communities. This potentially can lead to some robustness problems in the network, specifically in those most peripheral stations.

On the other hand, Madrid metro network is slightly degree assortative since $r = 0.296$ (note that, this coefficient ranges from $-1$ (disassortative network) to 1 (assortative network))

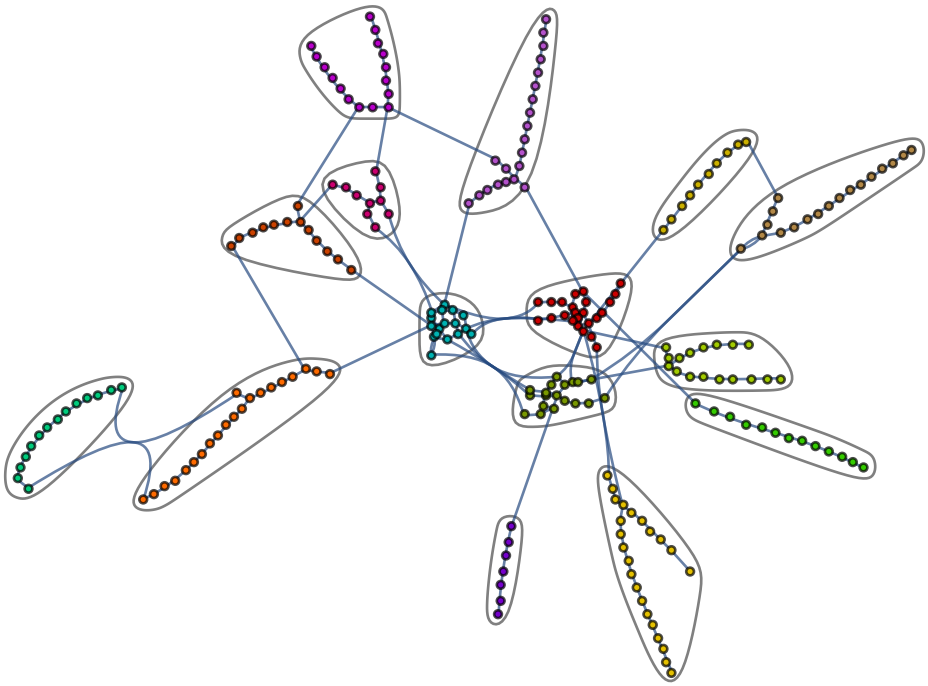

**Figure 11.** Communities in Madrid metro network.

**Table 5.** Distribution of stations per community and line.

| | L1 | L2 | L3 | L4 | L5 | L6 | L7 | L8 | L9 | L10 | L11 | L12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Community 1 | 0 | 12 | 0 | 7 | 5 | 5 | 1 | 0 | 5 | 0 | 0 | 0 |
| Community 2 | 0 | 0 | 0 | 0 | 7 | 0 | 17 | 0 | 0 | 0 | 0 | 0 |
| Community 3 | 18 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| Community 4 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 8 | 0 | 14 |
| Community 5 | 7 | 3 | 0 | 0 | 0 | 6 | 5 | 1 | 0 | 4 | 0 | 0 |
| Community 6 | 4 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 15 | 0 | 0 |
| Community 7 | 4 | 5 | 6 | 4 | 6 | 3 | 0 | 0 | 0 | 4 | 0 | 0 |
| Community 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 11 | 0 | 0 | 0 |
| Community 9 | 0 | 0 | 8 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 7 | 0 |
| Community 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 14 |
| Community 11 | 0 | 0 | 0 | 0 | 7 | 7 | 0 | 0 | 0 | 0 | 0 | 0 |
| Community 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13 | 0 | 0 | 0 |
| Community 13 | 0 | 0 | 4 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Community 14 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Community 15 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 |

## 6. Conclusions and Future Work

In this paper, Madrid metro network has been analyzed using not only complex network analysis but also multivariate statistical methods. Through complex network theory, the main structural and topological properties of Madrid metro network has been obtained as well as its robustness characteristics. Furthermore, statistical methods have allowed us to analyze the relationships between the centrality measures used in this work and also to evaluate stations according to their

centrality. Specifically, the HJ-Biplot method has been used because of its advantages when inspecting multivariate data matrix. Using the coordinates obtained from the HJ-Biplot, a classification of the stations according to their centrality measures has been carried out. As a result of HJ-Biplot analysis, positive correlation has been found between degree and clustering, degree and closeness, degree and betweenness and degree and eigenvector centrality. Instead, closeness and eccentricity are strongly negatively correlated. This implies that to classify the stations it would not be necessary to use all centrality measures. From the Cluster analysis we have obtained a classification of the stations according to their centrality measures in accordance with their remoteness to the city center. Three clusters have been found so that, the stations farthest from the city center have the lowest centrality measures and belong to a cluster, those located around the city center have medium values and belong to another and those from the city center have the highest values and belong to the other one. Finally, an ANOVA analysis was performed and the relationship between the centrality measures of a station and the flow of passengers of the station has been demonstrated. From the Pearson's correlation analysis it was found that the most central stations are those with the highest passenger flow.

Regarding the robustness analysis, Madrid metro network has a highly agglomerated community structure. On the other hand, the excessive length of some lines of Madrid metro network leads to their division between some communities. Although this is an usual characteristic of these transportation networks, it can result in some robustness problems that affect most peripheral stations.

In this work, the importance of Complex Network Analysis and Statistics has been highlighted when offering sustainable solutions. There is no doubt about the vital role that transport networks play in the development of smart cities and how the above techniques can help to achieve this. However, they are not only useful in this area, but they can also be applied to other areas of science in the context of Applied Sciences.

In future works, the robustness of Madrid metro network will be analyzed in more detail. For this, the passenger flow which would be affected by a possible failure in the network will be taken into account. In addition, other factors such as socio-demographic and land-use factors, will be introduced in the study, so that the relationships between them and centrality measures can be analyzed.

**Author Contributions:** E.F.B. conceived and designed the study, the paper has been written and edited by E.F.B. and A.M.d.R. and it has been revised by A.M.d.R. and P.G.V. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Ministerio de Ciencia, Innovación y Universidades (MCIU, Spain), Agencia Estatal de Investigación (AEI, Spain), and Fondo Europeo de Desarrollo Regional (FEDER, UE) under project with reference TIN2017-84844-C2-2-R (MAGERAN) and the project with reference SA054G18 supported by Consejería de Educación (Junta de Castilla y León, Spain).

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## References

1. Sutton, J.C. GIS applications in transit planning and operations: A review of current practice, effective applications and challenges in the USA. *Transp. Plan. Technol.* **2005**, *28*, 237–250. [CrossRef]
2. Chen, Y.Z.; Li, N.; He, D.R. A study on some urban bus transport networks. *Physica A* **2007**, *376*, 747–754. [CrossRef]
3. Shanmukhappa, T.; Ho, I.W.H.; Tse, C.K. Spatial analysis of bus transport networks using network theory. *Physica A* **2018**, *502*, 295–314. [CrossRef]
4. Hong, J.; Tamakloe, R.; Lee, S.; Park, D. Exploring the topological characteristics of complex public transportation networks: Focus on Variations in both single and integrated systems in the Seoul metropolitan area. *Sustainability* **2019**, *11*, 5404. [CrossRef]
5. Chatterjee, A.; Manohar, M.; Ramadurai, G. Statistical analysis of bus networks in India. *PLoS ONE* **2016**, *11*, e0168478. [CrossRef]

6. Yu,W.; Chen, J.; Yan, X. Space-time evolution analysis of the Nanjing metro network based on a complex network. *Sustainability* **2019**, *11*, 523. [CrossRef]
7. Zhang, J.; Wang, S.; Wang, X. Comparison analysis on vulnerability of metro networks based on complex network. *Physica A* **2018**, *496*, 72–78. [CrossRef]
8. Wang, W.; Cai, K.; Du, W.; Wu, X.; Tong, L.; Zhu, X.; Cao, X. Analysis of the Chinese railway system as a complex network. *Chaos Solitons Fractals* **2020**, *130*, 109408. [CrossRef]
9. Wang, J.E.; Mo, H.H.; Wang, F.H.; Jin F.J. Exploring the network structure and nodal centrality of China's air transport network: A complex network approach. *J. Transp. Geogr.* **2011**, *19*, 712–721. [CrossRef]
10. de Silva, E.; Stumpf, M.P. Complex networks and simple models in biology. *J. R. Soc. Interface* **2005**, *2*, 419–430. [CrossRef]
11. Mei,W.; Mohagheghi, S.; Zampieri, S.; Bullo, F. On the dynamics of deterministic epidemic propagation over networks. *Annu. Rev. Control* **2017**, *44*, 116–128. [CrossRef]
12. Gómez, D.; Figueira, J.R.; Eusébio, A. Modeling centrality measures in social network analysis using bi-criteria network flow optimization problems. *Eur. J. Oper. Res.* **2013**, *226*, 354–365. [CrossRef]
13. Pagani, G.A.; Aiello, M. Power grid complex network evolutions for the smart grid. *Physica A* **2014**, *396*, 248–266. [CrossRef]
14. Wang, J.; Jiang, C.; Gao, L.; Yu, S.; Han, Z.; Ren, Y. Complex network theoretical analysis on information dissemination over vehicular networks. In Proceedings of the 2016 IEEE International Conference on Communications (ICC), Kuala Lumpur, Malaysia, 22–27 May 2016; pp. 1–6.
15. Chopra, S.S.; Dillon, T.; Bilec, M.M.; Khanna, V. A network-based framework for assessing infrastructure resilience: A case study of the London metro system. *J. R. Soc. Interface* **2016**, *13*, 20160113. [CrossRef]
16. Cats, O.; Krishnakumari, P.; Tundulyasaree, K. Rail network robustness: The role of rapid development and a polycentric structure in withstanding random and targeted attacks. In Proceedings of the Transportation Research Board 98th Annual Meeting, Washington, DC, USA, 13–17 January 2019.
17. Derrible, S. Network Centrality of Metro Systems. *PLoS ONE* **2012**, *7*, e40575. [CrossRef]
18. Oldham, S.; Fulcher, B.; Parkes, L.; Arnatkeviciute, A.; Suo, C.; Fornito, A. Consistency and differences between centrality measures across distinct classes of networks. *PLoS ONE* **2019**, *14* e0220061. [CrossRef]
19. Li, C.; Li, Q.; Van Mieghem, P.; Stanley, H.E.; Wang, H. Correlation between centrality metrics and their application to the opinion model. *Eur. Phys. J. B* **2015**, *88*, 1–13. [CrossRef]
20. Ronqui, J.; Travieso, G. Analyzing complex networks through correlations in centrality measurements. *J. Stat. Mech. Theory Exp.* **2015**, *9*, P05030. [CrossRef]
21. Galindo, M.P. Una alternativa de representación simultánea: HJ-Biplot (An alternative of simultaneous representation: HJ-Biplot). *Questíio* **1988**, *10*, 13–23.
22. Amor-Esteban, V.; Galindo-Villardón, M.P.; García-Sánchez, I.M. Industry mimetic isomorphism and sustainable development based on the X-STATIS and HJ-Biplot methods. *Environ. Sci. Pollut. Res.* **2018**, *25*, 26192–26208. [CrossRef]
23. Carrasco, G.; Molina, J.L.; Patino-Alonso, M.C.; Castillo, M.; Vicente-Galindo, M.P.; Galindo-Villardón, M.P. Water quality evaluation through a multivariate statistical HJ-Biplot approach. *J. Hydrol.* **2019**, *577*, 123993. [CrossRef]
24. Gallego-Álvarez, I.; Rodríguez-Domínguez, L.; García-Rubio, R. Analysis of environmental issues worldwide: a study from the biplot perspective. *J. Clean. Prod.* **2013**, *42*, 19–30. [CrossRef]
25. Frutos, E.; Purificación-Galindo, P.; Leiva, V. An interactive biplot implementation in R for modeling genotype-by-environment interaction. *Stoch. Environ. Res. Risk Assess.* **2013**, *28*, 1629–1641. [CrossRef]
26. Kolaczyk, E.D. *Statistical Analysis of Network Data*; Springer Science + Business Media: Berlin/Heidelberg, Germany, 2009.
27. Soh, H.; Lim, S.; Zhang, T.; Fu, X.; Lee, G.K.K.; Hung, T.G.G.; Di, P.; Prakasam, S.; Wong, L. Weighted complex network analysis of travel routes on the Singapore public transportation system. *Physica A* **2010**, *389*, 5852–5863. [CrossRef]
28. Barabási, A.L. *Network Science*; Cambridge University Press: Cambridge, UK, 2016.
29. Frutos Bernal E.; Martín del Rey A. Study of the Structural and Robustness Characteristics of Madrid Metro Network. *Sustainability* **2019**, *11*, 3486. [CrossRef]
30. Newman, M.E. Detecting community structure in networks. *Eur. Phys. J. B* **2004**, *38*, 321–330. [CrossRef]

31. Chen, G.; Wang, X.; Li, X. *Fundamentals of Complex Networks. Models, Structures and Dynamics*; John Wiley & Sons: Singapore, 2015.
32. Newman, M.E. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 8577–8582. [CrossRef]
33. Newman, M.E.; Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E.* **2004**, *6*, 1–15. [CrossRef]
34. Newman, M.E. Assortative mixing in networks. *Phys. Rev. Lett.* **2002**, *89*, 208701. [CrossRef]
35. Noldus, R.; van Mieghem, P. Assortativity in complex networks. *J. Complex Netw.* **2015**, *3*, 507–542. [CrossRef]
36. Madrid Metro Data. Available online: https://www.metromadrid.es/es/quienes-somos/somos-centenarios (accessed on 12 July 2020).
37. Statistical Data of Madrid Metro. Available online: https://www.metromadrid.es/es/transparencia/informacion-economica-presupuestaria-y-estadistica/datos-estadisticos (accessed on 12 July 2020).
38. Gabriel, K.R. The Biplot graphic display of matrices with applications to principal components analysis. *Biometrika* **1971**, *58*, 453–467. [CrossRef]
39. Gabriel, K.R.; Odoroff, C.L. Biplots in biomedical research. *Stat. Med.* **1990**, *9*, 469–485. [CrossRef] [PubMed]
40. Vicente Villardón, J.L. *MultBiplot: A Package for Multivariate Analysis Using Biplots*; Departamento de Estadística, Universidad de Salamanca, Spain, 2010.
41. Forgy E. Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics* **1965**, *21*, 768–769.
42. MacQueen, J. Some methods for classification and analysis of multivariate observations, Proc. Fifth Berkeley Symp. *Math. Stat. Probab.* **1967**, *1*, 281–296.