# Analysing the Intermeshed Patterns of Road Transportation and Macroeconomic Indicators through Neural and Clustering Techniques

**Abstract:** As is widely acknowledged, the transportation of goods by road can, in one way or another, be linked to a range of macroeconomic indicators. A Hybrid Artificial Intelligence System is proposed in this paper to analyse the interaction between transportation patterns and the economy. The temporal patterns of road transportation and macroeconomic trends are studied, by combining the use of both (supervised and unsupervised) neural networks and clustering techniques. The proposed system is validated, by establishing links between road transportation data from Spain and macroeconomic trends over 6 years (2011 to 2017). The results reveal an interesting inner structure of the data, through data visualizations of intermeshed relations between road transportation patterns and macroeconomic indicators. The same data structure was also visible in the output of the clustering techniques. Furthermore, a number of high-quality predictions were advanced by processing the road transportation data as time series, and forecasting the future values of the main series. These results demonstrated the validity of the proposed linkage between road transportation data and macroeconomic indicators.

**Keywords:** Artificial Intelligence, Exploratory Projection Pursuit, Clustering, Forecast, Economy, Road Transportation.

**Acknowledgments**: We are grateful for the complete datasets of the Permanent Survey of Goods Transport by Road (*Encuesta Permanente de Transporte de Mercancías por Carretera*) facilitated by the General Sub-Directorate of Economic Studies and Statistics of the Ministry of Development (*Subdirección General de Estudios Económicos y Estadísticas del Ministerio de Fomento*) of Spain.

## 1 Introduction & Previous Work

Scientific study has over many years pointed out various correlations between economic activity and the transportation of goods. Two main areas of contemporary study may be mentioned that are driving research in this field: A) studies [1] that begin with an attempt to model the demand for transport on the basis of economic activity, for the economic planning of that activity [2]; and, B) studies that examine the correlation between transport activity and the analysis of input-output flows of GDP [3], for determining the relation between economic activity and road transport measured by tons per km [4]. In general, those studies that have centred on the transport of goods and that consider the inland mode are clearly dominant; in other words, road transport accounts for 76.2 % of goods (tons-per km) transported in 2017 in the EU [5], and 94.7% of goods transported in Spain, for which reason it is converted into the main reference that is generally employed to measure activity in that field.

The close relation between those magnitudes is confirmed up until the trend no longer holds true, leading to the affirmation that the “link has been broken” [6]. What was initially detected in the economy of New Zealand was soon tested in the European Union [7], and its extension was also verified with a view to mitigating the worrying energetic dependency upon that activity [8]. Upon finding that this broken link was not due to the much desired modal shift of activity towards other more sustainable means of transport [9], it was attributed to the conjunction of a three-fold effect: a) the increased productivity of the transport sector; b) the application of specific new policies on its activity; and, c) changes in the economic model which have since been called dematerialisation [10].

The general line of investigation into the input-output models [11] is at present largely centred on the correlation between transport activity and economies within regional and national boundaries [12]. It is likewise centred on the connection and the interrelation of both factors within models of transport flows [13]. In parallel, research is also busy with toxic emissions and sustainability linked to the activity of transport, in what today appears to be one of the central vectors of the scientific world, as a necessary catalyser of economic activity, but with high hidden social costs [14].

There is however an underlying question on the adaptation of the available data and the general approach of those studies, making a clear reference to the magnitude of transport and its variability as a function of its specific activity [15]. In general, too deterministic a line of interpretation appears to be used towards overly simplified data. And, it is all within a general framework that confronts us with the relation between two meta magnitudes, the realities of which are undergoing profound transformation, and between which a complex relation can be seen: the economy and transport activity.

Finally, it must be highlighted that a huge amount of data is available associated to the two areas under study (road transportation and economy). Thus, it is a challenging task to analyse these vast datasets in order to find underlying patterns. This issue is also addressed in present paper, as relevant indicators are selected and the high dimensionality of the datasets is reduced by appropriate methods in order to obtain informative visualizations.

Artificial Intelligence (AI) in general and machine learning in particular have previously been applied to analyse data related to economic and enterprise management issues [16] [17] [18] [19]. Furthermore, Hybrid Artificial Intelligence Systems, comprising a wide variety of paradigms (neural networks, fuzzy logic, expert systems, etc.), have previously been proposed to a wide variety of application fields such as cybersecurity [20], knowledge management [17], and environmental sciences [21], among others. In the case of road transportation, although many AI-driven programs have been proposed to date for monitoring, supervising, and driving the vehicles, scant attention has been devoted by AI researchers to the analysis and management of road transportation from a high-level (economic) perspective. In [22], the effectiveness of different AI models for forecasting road transportation needs was analysed. More precisely, a multiplicative version of Winter's method, harmonic analysis, and harmonic analysis aided by artificial immune systems were applied to predict road transportation demand in Poland. As those investigations pointed out, transport demand has previously been forecast through some other AI models such as neural networks, evolutionary computation and fuzzy logic. The Multilayer Perceptron (MLP) was applied in [23] to predict future transaction costs associated with transport. The data under analysis concerned direct transport-related trading costs at the frontier of the Czech Republic between 2005 and 2014. Similarly, the MLP was used [24] for predicting container train flows in the direction of China – Europe via Kazakhstan, on a weekly basis for 69 weeks.

Unlike those few previous works, the present paper goes a step further, by proposing a HAIS for studying the interplay between the economy and goods transport by road at a national level in Spain. Although neural and clustering methods have been previously applied in isolation or combined, they have been selected for present paper as they fit the requirements. Furthermore, they are applied in a novel way to analyse patterns present in road transportation and macroeconomic data.

The remaining sections of this study will be structured as follows: the proposed HAIS will be described in section 2 together with the AI models that it comprises. Section 3 will introduce the dataset for analysis, the experiments and the results that were obtained. Finally, both conclusions and future work proposals will be discussed in section 4.

## 2 Proposed HAIS

Hybridization [25, 26] has previously been proposed and successfully applied to combine intelligent paradigms. In the present paper, a Hybrid Artificial Intelligence System (HAIS) is proposed to analyse and to model the interplay between the national economic and road transportation statistics. Under the hybridization perspective, both unsupervised and supervised paradigms are combined to study that data. Firstly, both data sources will be studied with unsupervised models; Exploratory Projection Pursuit (EPP) and clustering techniques will be applied to gain initial knowledge on the dataset under analysis and its structure. Different unsupervised neural methods will be compared, namely Principal Component Analysis (PCA) and Cooperative Maximum Likelihood Hebbian Learning (CMLHL), in order to visualize the multi-dimensional data. For a comprehensive exploratory analysis of the data, both *k*-means and agglomerative clustering will also be applied to the same data. Finally, non-linear neural models will also be applied to forecast the evolution of the most interesting features. The above techniques will be described in the following subsections.

### 2.1 Exploratory Projection Pursuit

Principal Component Analysis (PCA) is a widely-used statistical method [27] based on the analysis of information, through the linear mapping of data dimensions (reduction of the number of variables). Its end-purpose is to reduce the number of variables in a dataset with multiple variables, attempting, as far as possible, to minimize information loss for the new data. The new factors (principal components) obtained through that linear mapping will be the result of a linear combination of the original variables that, in turn, will perform as independent variables between each other.

According to [28], PCA may be described as a mapping of vectors, $X_d$, onto an $N$-dimensional space on vectors, $Y_d$, in an $M$-dimensional space, where $M <= N$. While $X$ can be represented as a linear combination of a set of $N$ orthonormal vectors $W_i$:

$$x = \sum_{i=1}^{N} y_i W_i \tag{1}$$

Differentiating from PCA, Cooperative Maximum Likelihood Hebbian Learning (CMLHL) is an unsupervised ANN that is characterized by its capacity to conserve a degree of global order in the datasets. This technique permits a topological ordering of the different neurons, in keeping with a

neighbourhood or similarity rule, permitting a search for exploratory projections (Exploratory Projection Pursuit – EPP) [29]. One neural implementation of EPP is Maximum Likelihood Hebbian Learning (MLHL) [30], [31], each iteration of which consists of the following steps:

Feed Forward Step:

$$y_i = \sum_{j=1}^{N} W_{ij} x_j, \forall i \tag{2}$$

Feedback Step:

$$e_j = x_j - \sum_{i=1}^{M} W_{ij} y_i \tag{3}$$

Updating of weights:

$$\Delta W_{ij} = \eta y_i sign(e_j) \left| e_j \right|^{p-1} \tag{4}$$

Where $\eta$ is the learning rate and $p$ is a parameter associated to the learning rule to be tuned.

With regard to MLHL, the inclusion of lateral connections is proposed [30], [32] that are derived from the Rectified Gaussian Distribution (RGD) and that are based on cooperative distributions [32], generating the CMLHL model. The lateral connections recalculate the output of the network at a given time ($t+1$) by taking into account the previous output ($t$) in accordance with the following equation:

$$y_i(t+1) = \left[ y_i(t) + \tau(b - Ay) \right]^+ \tag{5}$$

Where $\left[ \ \right]^+$ is the necessary rectification, so that the values of $y$ remain in the positive quadrant and the strength of the lateral connections, $\tau$, between the neurons of the output layer. An appropriate value for $\tau$ must be selected, so that the algorithm converges towards a stationary point of the energy function (generally a local minimum).

### 2.2 Clustering Techniques

Cluster analysis [33] is a data organization technique that groups data samples by a given criterion (mainly distance). Two individuals in a valid group will be of far greater similarity than those in different groups. The clustering *k*-means algorithm [34] groups the data samples into a previously given number of groups. Two input parameters are required for its application: the number of clusters ($k$), and their initial centroids. Firstly, each data sample is assigned to the cluster with the nearest centroid. Once the groups are defined, the centroids are recalculated, and a reallocation of samples takes place. Those steps are repeated until no further modifications are made to the centroids. The criterion to measure the quality of the grouping is the sum of the proximity Sum of Squared Errors (SSE). The algorithm attempts to minimize it, by the following expression:

$$SSE = \sum_{j=1}^{k} \sum_{x \in G_j} \frac{p(x_i, c_j)}{n} \tag{6}$$

Where, $k$ is the number of groups, $p$ is the proximity function, $c_j$ is the centroid of group $j$, and $n$ is the number of data samples.

The main distances were selected from among all of the [35] proposed distances for this clustering algorithm [36]. After comparing the results, the Cityblock distance was selected, as a distance measure where each centroid is placed in the component-wise median of all the samples in the group. The distance from $n$-dimensional point $x$ to each of the centroids is calculated as:

$$d = \sum_{k=1}^{n} \left| x_k - c_{jk} \right| \tag{7}$$

Unlike partitional clustering methods, hierarchical ones can be divided into two types:

1. Agglomerative: the process begins with each data item in a different cluster and the clusters are successively merged together until a stopping criterion is met or until a single cluster is obtained.

2. Divisive: the process begins with all data assigned to only one cluster (and its descendants), which is split until a stopping criterion is satisfied or until every data item is assigned to a different cluster.

In the present study, due to the successful results in initial experiments, agglomerative clustering was selected, for comparison with the (*k*-means) partitional approach. In the case of agglomerative clustering, a variety of linking methods can be applied. In the present study, the following methods were tested:

- Single: shortest distance.
- Complete: furthest distance.
- Ward's hierarchical clustering method: inner squared distance (minimum variance algorithm), appropriate for Euclidean distances only.
- Median: weighted centre of mass distance (WPGMC: Weighted Pair Group Method with Centroid Averaging), only appropriate for Euclidean distances.
- Average: unweighted average distance (UPGMA: Unweighted Pair Group Method with Arithmetic Averaging).
- Centroid: centroid distance (UPGMC: Unweighted Pair Group Method with Centroid Averaging), appropriate for Euclidean distances only.
- Weighted: weighted average distance (WPGMA: Weighted Pair Group Method with Arithmetic Averaging).

### 2.3 Neural Models for Time Series Prediction

Neural models can also be used to model Non-linear dynamic systems. In the present paper, they are therefore applied to Time Series Prediction (TSP), associated with some of the transport features that are described in Section 3.1. Non-linear Input-Output (NIO), Non-linear Auto-Regressive (NAR), and Non-linear Auto-Regressive with Exogenous Input (NARX) [37] were applied, in order to build an accurate prediction model. The mathematical formulation for the former is:

$$y(t) = f\big(x(t-1), \dots, x(t-n_x)\big) \qquad (8)$$

where, *y(t)* is the variable to be predicted over time, *t*; *x(t)* is another (exogenous) input used to predict *y(t)* over time *t*; $n_x$ is the maximum number of time delays in this input to be considered in the model; and, *f()* is the function to be approximated by the neural model. When, rather than x(t), the series that is used to make the prediction, (*y(t)*), is to be predicted, the model, known as NARX, is formulated as:

$$y(t) = f\left(y(t-1), \dots, y\big(t-n_y\big)\right) \qquad (9)$$

where, $n_y$ is the maximum number of time delays in the output to be considered in the model, and *f()* is the function to be approximated by the neural model. In the case of NARX, its mathematical formulation can be described as:

$$y(t) = f\big(y(t-1), \dots, y\big(t-n_y\big), x(t-1), \dots, x(t-n_x)\big) + \xi(\mathrm{t}) \qquad (10)$$

where, *x(t)* is the exogenous input; $n_x$ is the maximum number of time delays in the input to be considered by the model; and, $\xi(\mathrm{t})$ is the noise term. As can be seen, in the case of NARX, both *y(t)* and *x(t)* are used to predict future values of *y(t)*.

These neural models are implemented as a feedforward time delay neural network with no feedback loop [38] and have demonstrated that they can outperform linear models [39].

## 3 Experiments & Results

The settings for the experiments with the details on the dataset are presented below in subsection 3.1. The results from the above-defined models are shown and discussed in the subsequent subsections.

### 3.1 Dataset

Data were retrieved from the following government agencies and some Spanish government ministries: *Instituto Nacional de Estadística* (INE) (National Institute of Statistics), *Ministerio de Empleo y Seguridad Social* (MEYSS) (Ministry of Employment and Social Security), *Ministerio de Industria, Comercio y Turismo* (Ministry of Industry, Commerce and Tourism), the Bank of Spain, and the *Ministerio de Fomento* (Ministry of Development) (through its General Sub-Directorate of Economic Studies and Statistics). Data from between 2011 and 2017 were used, given that no previous data were available from all the data sources in use. All the data represented quarterly levels of aggregation, which therefore included each variable in the study, a total of 28 values.

The values of some annual series of some of the most relevant macroeconomic variables associated with the Spanish economy were employed, as described in Table 1.

| Series | Aggregate | Description |
|---|---|---|
| GDP | Total | Gross Domestic Product (GDP) |
| GVA-Primary | Primary sector | Sectorial Gross Value Added. Value added by the different productive sectors. |
| GVA-Industrial | Industrial sector | |
| GVA-Construction | Construction sector | |
| GVA-Services | Services sector | |
| GVA-Taxes | Production taxes | |
| GDP income | Domestic | Rents. Flows of monetary rents between the different socio-economic actors. |
| | Domestic investment | |
| | Demand for Goods Equipment | |
| | Demand for Means of Transport | |
| | Demand - Exportation of Goods | |
| | Demand - Exportation of Services | |
| | Demand – Importation of Goods | |
| | Demand – Importation of services | |
| Workers Employed | People employed (SCN) | Market and Labour performance |
| Employed full-time | Full-time positions | |
| Hours worked | Hours Worked | |
| Registered Workers | Registered workers | |
| Workers EPA | People employed (EPA) | |
| Unemployed EPA | Unemployed | |
| Work | Self-employed | |
| IPC - CPI | Consumer price index | Market price variation index |
| Debt Sect. Non-Financial | Total Debt | Active debt of different macroeconomic actors. |
| | Debt Public Administrations | |
| | Non-financial companies | |
| | Homes and ISFLSH | |
| ICNE – Business Figures Index (base 2015) | ICNE. General | *Índice de cifra de negocios empresarial* (ICNE) [Index of Volume of Business Turnover]; summary index that measures the short-term development of business turnover, jointly and by different sectors, through three surveys prepared by the National Institute of Statistics and data provided by the Spanish Tax Office. |
| | ICNE. Durable goods | |
| | ICNE. Non-durable goods | |
| | ICNE. Equipment Goods | |
| | ICNE. Intermediate Goods | |
| | ICNE. Energy | |
| IPI – Industrial Production Index (base 2015) | General | Development of the productive activity of industrial branches through an ongoing harmonized methodology survey in the EU |
| | Extraction industries | |
| | Manufacturing industries | |
| | Energy supply | |
| ISAE – Survey Export panorama | Index of General Economic Activity | |
| ICPA - *Encuesta coyuntura exportación* (Export scenario survey) | General Index of Order Books | Indices of export activity and perceived export scenario |
| | Perceived price volatility | |
| | Perceived rise in margins | |
| | Perceived export demand | |
| | Perceived price competitiveness | |
| | Perceived quality competitiveness | |

| | | |
|---|---|---|
| | Perceived availability of financing | |

*Table 1: Macroeconomic data series used in the study.*

Moreover, the road transport data associated with Spain was obtained from the European road freight transport survey (ERFT). Conducted by the ERFT, the survey relates to heavy goods vehicles licenced in Spain for the transport of goods. It has a sufficiently high sampling level to be of statistical representativeness for each Autonomous Region, in order to measure its transport operations. To that end, The survey registers the movement of a single class of goods, from a departure point to a destination. It was conducted in accordance with the corresponding EU Regulation [40]. The total number of records included on that database was 1,932,671 that has a sampling representativeness of 1,259,938,252 transport operations.

The ERTF survey variables and some other variables relating to relevant information on the sectoral activity of the transport of goods by road were also used in this study. The variables under consideration were:

- Price of transport (B): indexed on the basis of 100 over the average annual prices for 2000, according to the quarterly studies of the Ministry of Development.
- Transport costs: indexed on the basis of 100 over the average annual prices for 2000, according to the quarterly studies of the Ministry of Development.
- Fuel prices in Spain: quarterly averages weighted in centimes of an €, according to the data collected by the Ministry of Development.
- Fuel prices in the EU: quarterly averages weighted in centimes of an €, according to the data collected by the Ministry of Development.
- Tons transported (A, B, C): weight of goods in transport operations.
- Journeys completed (A, B, C): number of transport operations and empty running.
- Empty running (A, B): kilometres running empty.
- Maximum load (A, B): maximum possible weight for journeys completed in tons.
- Maximum load for empty running (A, B): maximum possible weight for empty running in tons.
- Haulage distance (A, B): distance travelled in kilometres.
- Haulage distance for empty running (A, B): distance travelled in kilometres for empty running.
- Number of vehicles represented (A): quantity of vehicles represented.
- Load capacity represented (A): maximum load of the vehicles represented.
- Tons-kms (A, B, C): product of tons transported and haulage distance of each operation.
- Average age of the vehicle fleet (A, B): average years covered since the registration of the vehicles.
- Average age of the vehicle fleet in empty running (A, B): years since the registration of vehicles involved in empty running.

The series designated with letters were sub-divided as follows:

(A) Type of carriage: A1) All types; A2) Own account; A3) Hire or reward.

(B) Distance class: B1) All distances; B2) < 50 km; B3) 51-100 km; B4) 101-200 km; B5) 201-300 km; B6) >300 km.

(C) Geographic link: C1) All links; C2) Municipal; C3) Regional; C4) National; C5) Importation; C6) Exportation; C7) Cabotage.

In other words, we assembled a total of 113 transport data series, with the values for 28 quarters in each one.

### 3.2 Results

The determination of the start and the end of a period of recession, whenever they may be, is a mere formality. The truth is that economic activity undergoes pronounced fluctuations, passing through zones of minimums, and recovery phases, which an analytical method will be able to highlight. It is much closer to the study of economic cycles and their phases of recession, depression, and expansion was therefore employed. According to that approach, rather than employ the classification of economic periods at a general level, the set of macroeconomic variables detailed in section 3.1. was analysed.

The results of that study showed 3 clearly defined periods; 1) an economic decline that runs from the first quarter of 2011 to the second quarter of 2012, where the economy continued to develop due to the boom of the preceding good scenario; 2) a zone of serious economic depression that runs from the third term of 2012 to the first of 2015, in which the formal inflection point marking the end of the crisis is the fourth quarter of 2013); and, 3) an economic recovery that runs from the second quarter of 2015 and

thereafter. The majority of the variables respond perfectly to this pattern, except Private Sector Debt (PSD) and Home Debt (HD) that continued to fall, depriving the economy of a necessary monetary mass for an acceptable recovery. There again, very modest growth of the CPI may be mentioned that in some sectors is desirable to ward off the threat of deflation.

The results from the techniques presented in section 2 are shown below, in an attempt to validate the similarities between those macroeconomic scenarios and the data associated with road transport.

### 3.2.1 Results of EPP Techniques

An innovative method, Exploratory Projection Pursuit (EPP), was applied, to validate the relations between the macroeconomic data and the transport data through the visualization of the transport data. The results were then analysed in the light of the macroeconomic data. In the projections shown below, each data sample was labelled by year and by quarter. Principal Component Analysis (PCA) was then applied to the transport data. The PCA analysis with two principal components is shown in Fig. 1.

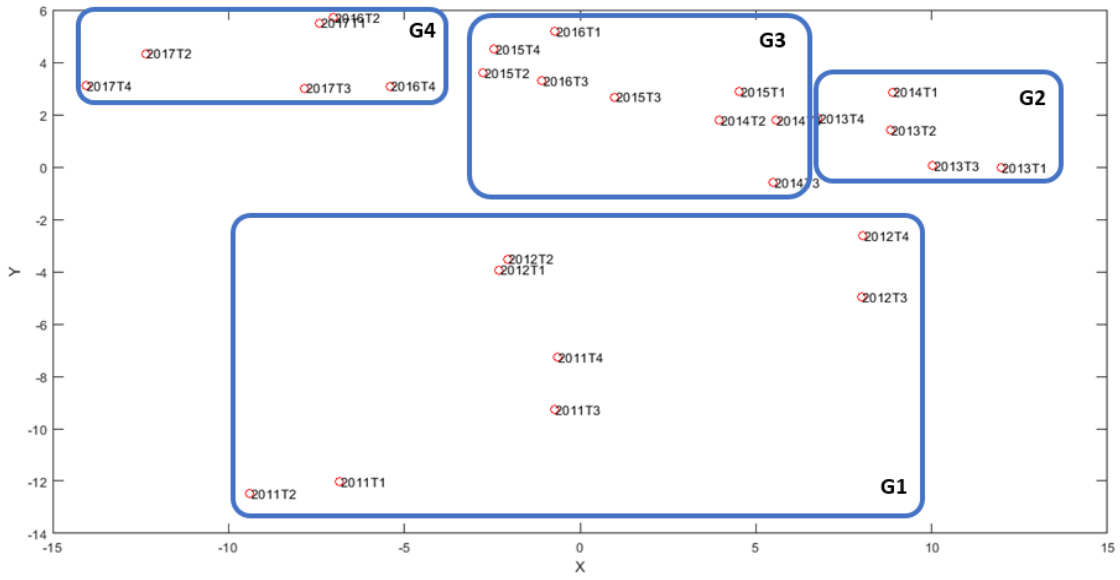


*Figure 1: 2D projection obtained by PCA when applied to transport data extracted from the ERTF survey as described in section 3.1. Data instances are labelled according to the year and quarter and identified groups are labelled for subsequent analysis.*

Data were associated with one of the four main groups to facilitate the analysis, as can easily be seen in Fig. 1 (G1 and G2). The more conventional PCA analysis of the projections defined those four groups as follows: G1) formed of 8 quarters from 2011 and 2012, corresponding to a cooling down of the economic cycle; G2) 2013 and the first quarter of 2014, corresponding to the trough of the economic cycle during the harshest points of the crisis; G3) the rest of 2014, 2015, and the first and third quarter of 2016 (the weakest seasonal quarters); and, G4) the second and the fourth quarters of 2016, with the four quarters of 2017. This analysis of the projection shows a relevant aspect; in the G3 group, the seasonality of the data is highlighted, which respond to the most active quarters in economic terms, which are the second and the fourth quarters, and the least productive quarters, which are the first and the third.

Additional information was added to the PCA projection, as shown in Fig. 2, in order to analyse the PCA projection in a more comprehensive manner. Quarterly data were connected, and the associated yearly trajectories were coloured. Thus, those trajectories are classified by Greek letters: Sigma, Lambda (inverse), and Gamma. Additionally, the two extreme economic situations are identified in the projection, which are also represented by the corresponding labels. The area of economic expansion is situated towards the lower left at the side (labelled expansion), and the most acute point of recession is situated towards the upper right at the side (labelled recession).

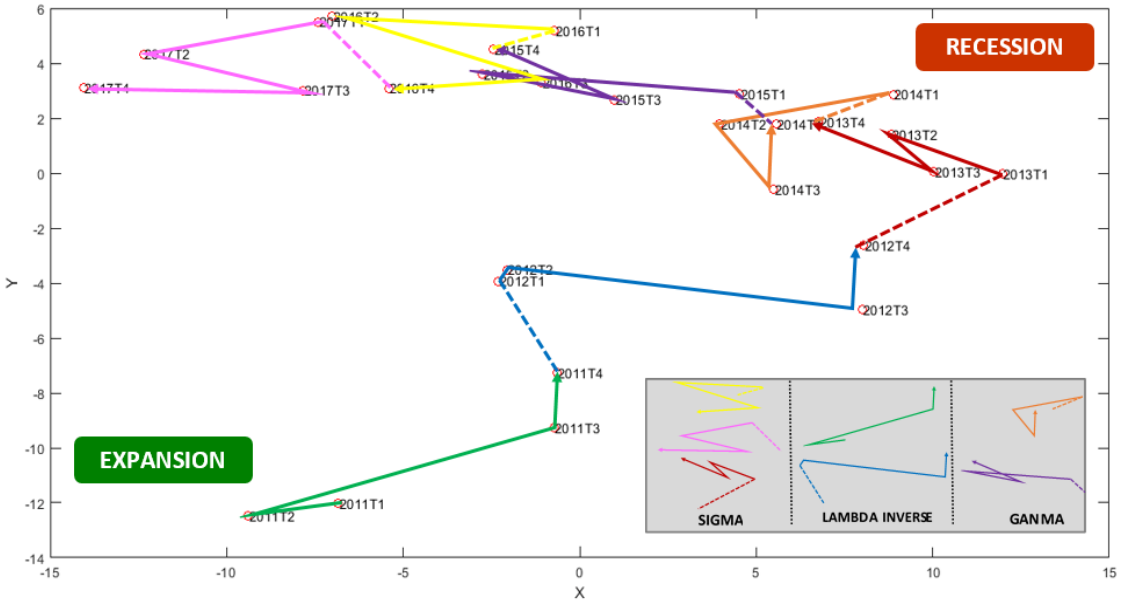


*Figure 2: Enriched PCA projection of transport data (Figure 1) that incorporates temporal (yearly) trajectories. In the bottom-right corner, the classification of such trajectories by Greek letters (Sigma, Lambda (inverse), and Gamma) is shown.*

The analysis of Fig. 2 reveals quite a coherent picture for interpretation: while the inverted Lambda forms of 2011 and 2012 point towards the most extreme area of recession, somewhere around 2013, a Sigma change occurred pointing to an area of expansion. It was prolonged in a way that favoured growth by two weaker Gamma type formations in the years 2014 and 2015, returning to create two somewhat firmer Sigma-type formations in 2016 and 2017.

In a complementary way, the dataset was also processed with CMLHL, given the interesting results that were obtained in earlier investigations [19]. The CMLHL projection is shown below in Fig. 3.

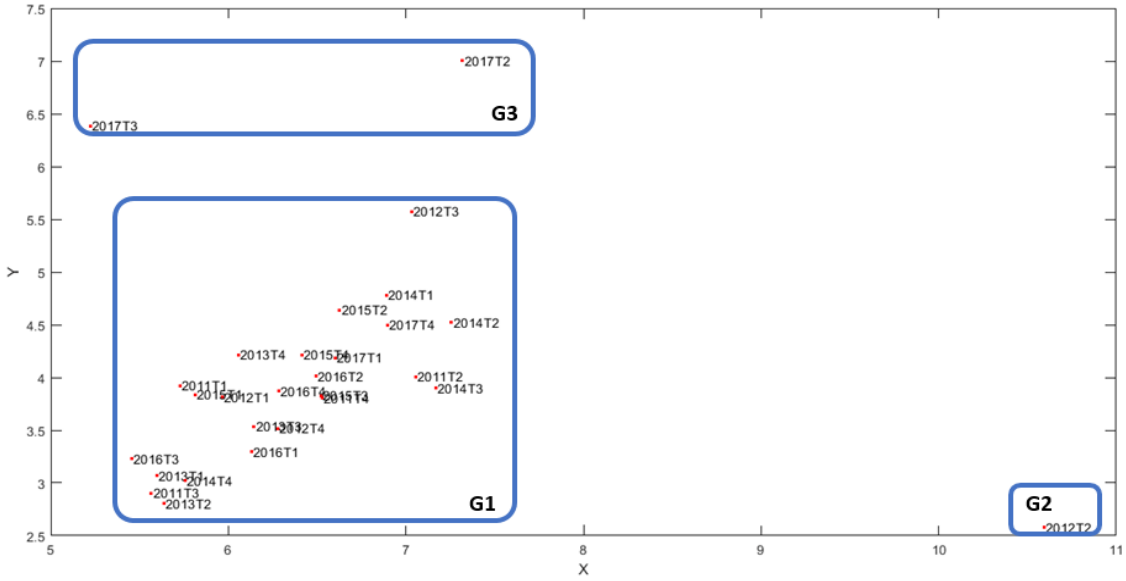


*Figure 3: 2D projection obtained by CMLHL when applied to transport data extracted from the ERTF survey as described in section 3.1. Data instances are labelled according to the year and quarter and identified groups are labelled for subsequent analysis.*

As may be seen for the transport data, the data structure was not projected by CMLHL as clearly as it was by PCA. The analysis of the seasonal patterns and the topographical zones of economic profitability and adversity were replicated, without such clear results as in the case of PCA. After this analysis, the

distribution of the majority of quarters is worth highlighting, which can be seen in a common area (G1 in Fig. 3) with different economic nuances. The technique only differentiates two cases from among the majority; in the first place, the second quarter of 2012 (G2 in Fig. 3) that underlines the unmistakable entry into recession, with the peculiarity that it is the only quarter where total debt decreases. In contrast, it also appears to have differentiated the second and the third quarters of 2017 (G3 in Fig. 3) with good economic growth and high growth of employment.

### 3.2.2 Results of Clustering Methods

Once again, as done with the EPP techniques, the clustering methods were exclusively applied to the transport data. As shown in present section, interesting results have been obtained by some of those methods when analysing the intermeshed patterns of road transportation and macroeconomic indicators. Experiments were conducted with both the different distance criteria and the linkage criteria used for the different methods in use (*k*-means and agglomerative). In addition, tests were conducted with different *k*-parameter values (number of groups to obtain): 2, 3, 4, 6, and 8. Among the results obtained, it may be indicated that *k*-means was not capable of finding a coherent structure that will give the best quarterly transport results. However, the agglomerative method is capable of doing so, as the better results demonstrate, in table 2. These better results were obtained for values of *k* 1 and 2 and the following linkage methods: average (L1), centroid (L2), complete (L3), median (L4), single (L5), Ward's (L6), and weighted (L7).

| **Quarter** | **k = 2** | | | | | | | **k = 3** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | L1 | L2 | L3 | L4 | L5 | L6 | L7 | L1 | L2 | L3 | L4 | L5 | L6 | L7 |
| **2011T1** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 |
| **2011T2** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 |
| **2011T3** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 1 |
| **2011T4** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 1 |
| **2012T1** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 1 |
| **2012T2** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 1 |
| **2012T3** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| **2012T4** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| **2013T1** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| **2013T2** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| **2013T3** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| **2013T4** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| **2014T1** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| **2014T2** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| **2014T3** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| **2014T4** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| **2015T1** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| **2015T2** | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 3 | 2 | 2 |
| **2015T3** | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 3 | 2 | 2 |
| **2015T4** | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 3 | 2 | 2 |
| **2016T1** | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 3 | 2 | 2 |
| **2016T2** | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 3 | 2 | 2 |
| **2016T3** | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 3 | 2 | 2 |
| **2016T4** | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 3 | 2 | 2 |
| **2017T1** | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 3 | 2 | 2 |
| **2017T2** | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 3 | 2 | 2 |
| **2017T3** | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 3 | 2 | 2 |
| **2017T4** | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 3 | 2 | 2 |

*Table 2: Cluster assignation for the best clustering results on transport data.*

From the results presented in table 2, it may be seen that the clustering methods with two groups were perfectly confined to the two large situations that were produced in the economic field; to be in outright depression (quarters from 2012T3 to 2015T1) or otherwise (quarters from 2011T1 to 2012T2 and from 2015T2 to 2017T4). All the linkage methods produced the same result, except for Single (L5), which reveals the concerning exception that the situation in 2015T2 was closer to economic depression and was still assigned to it. The clustering into 3 groups was clearly defined by the three perceived economic situations with a finer approach; decline (quarters 2011T1 to 2012T2), depression (2012T3 to 2015T1) and recovery (2015T2 to 2015T4). Once again, the Single (L5) exception appears to identify two different

circumstances within the economic decline, and once again the observation concerning a prolonged depression.

### 3.2.3 Results of Neural Models for TSP

It was considered feasible, on the basis of the promising results, both with the EPP techniques and the clustering methods, to try to predict some of the transport indicators using the transport variables as well as the macroeconomic variables. In this section, the interesting results are shown that were obtained with this proposal by the different models (NARX, NAR and NIO). Exhaustive experimentation was conducted for those three models, testing them with the following values for the parameters (where applicable):

- Number of input time-delays: 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10.
- Number of output time-delays: 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10.
- Number of neurons in the hidden layer: 1, 5, 10, 15 and 20.
- Training algorithm: 1 - Levenberg-Marquardt, 2 - Batch Gradient Descent, 3 - Gradient Descent with Momentum, 4 - Adaptive Learning Rate Backpropagation, 5 - Gradient Descent with Momentum and Adaptive Learning Rate, 6 - Scaled Conjugate Gradient, and 7 - Broyden–Fletcher–Goldfarb–Shanno Backpropagation (Quasi-Newton).

Among the available data, a standard selection strategy was employed, using 75% of the data for the training dataset and the remaining 25% for the validation dataset. The Mean Squared Error (MSE) is shown for the experiments that were conducted. Taking that figure into account, very high errors were obtained when trying to predict the majority of variables associated with transport. However, some interesting results were obtained for a few of those variables; the best of them are shown in table 3.

| **Variables** | | **NARX** | | **NAR** | | **NIO** | |
|---|---|---|---|---|---|---|---|
| **Name** | **Max value** | **Min MSE** | **%** | **Min MSE** | **%** | **Min MSE** | **%** |
| **Price of Transport** | 139.60 | 0.0930 | 0.0666 | 0.2866 | 0.2053 | 0.2211 | 0.1584 |
| **Average age of the vehicle fleet** | 10.18 | 0.0006 | 0.0057 | 0.0036 | 0.0353 | 0.0118 | 0.1167 |
| **Average age of the vehicle fleet in empty running** | 10.49 | 0.0011 | 0.0104 | 0.0034 | 0.0327 | 0.0026 | 0.0254 |
| **Transport costs** | 164.00 | 0.3443 | 0.2099 | 5.1588 | 3.1456 | 0.5287 | 0.3224 |

*Table 3: Best results when predicting some of the transport features.*

With regard to table 3, the MSE was normalized in accordance with the maximum value for each of the variables (Max value in table 3). The lowest MSE obtained in all the experiments is therefore shown for each of the models and the % that it represents with regard to the maximum value of that variable.

From those results, it is seen that the best predictions used both sources of data with the NARX model, in some cases undergoing important reductions in error. The use of solely macroeconomic data (with the NIO model) yielded better results than the use of data from the transport sector (with the NAR model) in 3 out of the 4 cases. Especially significative was the case of the transport costs, in which the error obtained with the NAR model was significantly higher than the error obtained by the other two models.

The values of the different parameters that were used to gain those best results are shown in table 4.

| **Model** | **Input delays** | **Output delays** | **Number of neurons** | **Algorithm** |
|---|---|---|---|---|
| **Transport price** | | | | |
| NARX | 8 | 2 | 20 | 7 |
| NAR | - | 10 | 10 | 1 |
| NIO | 5 | - | 10 | 1 |
| **Average age of the vehicle fleet** | | | | |
| NARX | 2 | 10 | 10 | 1 |
| NAR | - | 9 | 15 | 5 |
| NIO | 9 | - | 15 | 5 |
| **Average age of the vehicle fleet in empty running** | | | | |
| NARX | 3 | 7 | 20 | 1 |
| NAR | - | 8 | 10 | 1 |
| NIO | 2 | - | 20 | 7 |
| **Transport costs** | | | | |
| NARX | 3 | 9 | 5 | 5 |
| NAR | - | 10 | 15 | 7 |

| | | | | |
|---|---|---|---|---|
| NIO | 10 | - | 5 | 7 |

*Table 4: Parameter Values for the most predictive results (Table 3).*

Analysing the values shown in table 4, high values (both for the input and for the output) delays were seen to predominate, in order to obtain the best results. In other words, a broad time window must be considered for the models, in order to minimize the error. In the case of both data sources (model NARX), an imbalance arises in the time window between both data sources; in most cases (except for the Price of Transport), the model with the best results employed a high number of delays at the output (8.33), and a low number of output delays (2.66). With regard to the number of neurons in the hidden layer, the lowest value (1) never yielded the best results, but all the other values did in at least 3 cases. The number of 10 neurons stood out in the highest number of best results (4). Finally, there appeared to be greater consensus over the learning algorithm, as only 3 of the 7 options that were used obtained some of the best results: Levenberg-Marquardt (1), Gradient Descent with Momentum and Adaptive Learning Rate (5), and Broyden–Fletcher–Goldfarb–Shanno Backpropagation (7). The first obtained the best results in 5 cases and the other two in 3 and 4 cases, respectively.

## 4 Conclusions

Given the results presented in the preceding section, it can be concluded that the proposed HAIS is a valuable tool for the analysis of the data on macroeconomic indicators and goods transport. The techniques have proved useful for the study that has been designed; the EPP methods and the clustering techniques have provided an initial approach to understand the structure of the dataset under study. More specifically, PCA and the method of agglomerative grouping yielded the best results for non-supervised exploration of the data, leading to improved predictive accuracy for non-supervised exploration of the data. Besides, the neuronal models can also be used for the prediction of some of the objective time series, with a reduced error margin. The use of the macroeconomic data in those models yielded improved prediction. With respect to the other parameters (number of delays, number of hidden neurons, and maintenance algorithm), there was no successful combination that always yielded the best results. It was therefore necessary to adapt them to each case, conducting experiments to test them and to choose the one that presented the lowest error.

In view of the field of application, we should reflect on two aspects. 1) On the one hand, the point of view that transport activity is, when correctly analysed, a reliable indicator of economic activity. According to that assumption, a proper study of the results obtained with PCA and the agglomerative clustering of the new data generated over the timeframe of the study might alert us to and even confirm changes in economic trends. 2) On the other hand, the point of view that economic activity provokes changes in the levels of transport activity. In that sense, the neuronal models were shown to be especially useful; the fact that they served as predictors of prices and costs of that activity convert them into useful tools for the sector and/or for those requesting its services. Not only in the framework of price negotiations between the parties, but in predictable scenarios for investment and assignation of costs to complex project logistics.

Future work will be initially based on the comparison to some other competitive methods. Additionally, techniques specifically oriented towards the management of times series with low data volumes will be applied, in order to adapt in a better way to the reality of the available data. Likewise, scaling a study at a European level will be explored, taking data of a transnational type and thereby valuing the intermeshed relations that exist at that level. The predictability of the age of the fleet of transport vehicles has a strong linkage, with a view to taking action against environmental damage and road safety that is linked to an aging vehicle fleet. It may also serve as a possible basis for the verification of projects offering governmental support for the renovation of that equipment. Accordingly, future work will also address this issue.

## References


1. Bennathan, E., Fraser, J. y Thompson, L.S., *What determines demand for freight transport?* 1992
2. Crainic, T.G. y Laporte, G., *Planning models for freight transportation.* European Journal of Operational Research, 1997. **97**(3): p. 409-438. DOI: 10.1016/S0377-2217(96)00298-6
3. Costa, P., *Using Input-Output to Forecast Freight Transport Demand*. 1988, Springer, Berlin, Heidelberg. p. 79-120. 10.1007/978-3-662-02551-2_3

4. McKinnon, A. y Woodburn, A., *Logistical restructuring and road freight traffic growth.* Transportation, 1996. **23**(2): p. 141-161. DOI: 10.1007/BF00170033

5. Eurostat_Press_Office, *Energy, transport and environment indicators 2018 edition*. 2018. DOI: 10.2785/326009

6. Ballingall, J., Steel, D. y Briggs, P. *Decoupling Economic Activity and Transport Growth: The State of Play in New Zealand* 2003. URL: https://atrf.info/papers/2003/2003_Ballingall_Steel_Briggs.pdf.

7. Alises, A. y Vassallo, J.M., *Comparison of road freight transport trends in Europe. Coupling and decoupling factors from an Input–Output structural decomposition analysis.* Transportation Research Part A: Policy and Practice, 2015. **82**: p. 141-157. DOI: https://doi.org/10.1016/j.tra.2015.09.013

8. Profillidis, V., Botzoris, G. y Galanis, A., *Decoupling of economic activity from transport-related energy consumption: an analysis for European Union member countries.* International Journal of Innovation and Sustainable Development, 2018. **12**(3): p. 271-271. DOI: 10.1504/IJISD.2018.091518

9. Murko, D. y Štok, Z.M., *The impact of the transport policies on marketing charges in the rail transport comparisons between the European Union member states.* International Journal of Logistics Systems and Management, 2017. **27**(2): p. 187-207. DOI: 10.1504/IJLSM.2017.083816

10. Alises, A. y Vassallo, J.M. *The Impact of the Structure of the Economy on the Evolution of Road Freight Transport: A Macro Analysis from an Input-output Approach. Transportation Research Procedia*. 2016. Elsevier B.V. DOI: 10.1016/j.trpro.2016.05.404

11. Yu, H., *A review of input–output models on multisectoral modelling of transportation–economic linkages.* Transport Reviews, 2018. **38**(5): p. 654-677. DOI: 10.1080/01441647.2017.1406557

12. Peng, Z.-R. y Yu, H. *Economic Analysis Framework for Freight Transportation Based on Florida Statewide Multi-Modal Freight Model. Final Report*. Florida Department of Transportation Research Center. University of Florida. 2018. URL: https://rosap.ntl.bts.gov/view/dot/35548.

13. Russo, F. y Musolino, G., *A unifying modelling framework to simulate the Spatial Economic Transport Interaction process at urban and national scales.* Journal of Transport Geography, 2012. **24**: p. 189-197. DOI: https://doi.org/10.1016/j.jtrangeo.2012.02.003

14. Liimatainen, H. y Pöllänen, M., *The impact of sectoral economic development on the energy efficiency and CO2 emissions of road freight transport.* Transport Policy, 2013. **27**: p. 150-157. DOI: 10.1016/j.tranpol.2013.01.005

15. Courtonne, J.Y., Longaretti, P.Y. y Dupré, D., *Uncertainties of Domestic Road Freight Statistics: Insights for Regional Material Flow Studies.* Journal of Industrial Ecology, 2018. **22**(5): p. 1189-1201. DOI: 10.1111/jiec.12651

16. Herrero, Á. y Jiménez, A., *Improving the Management of Industrial and Environmental Enterprises by means of Soft Computing.* Cybernetics and Systems, 2019. **50**(1): p. 1-2. DOI: 10.1080/01969722.2019.1560961

17. Herrero, Á., et al., *A hybrid proposal for cross-sectoral analysis of knowledge management.* Soft Computing, 2016. **20**(11): p. 4271-4285. DOI: 10.1007/s00500-016-2293-9

18. Sáiz-Bárcena, L., et al., *Easing knowledge management in the power sector by means of a neuro-genetic system.* International Journal of Bio-Inspired Computation, 2015. **7**(3): p. 170-175. DOI: 10.1504/ijbic.2015.069556

19. Herrero, Á., Jiménez, A. y Bayraktar, S., *Hybrid Unsupervised Exploratory Plots: a Case Study of Analysing Foreign Direct Investment*. Complexity, 2019. **2019**: p. 6271017

20. Herrero, Á. y Corchado, E., *Mobile Hybrid Intrusion Detection: The MOVICAB-IDS System*. Studies in Computational Intelligence. Vol. 334. 2011: Springer

21. Arroyo, Á., et al., *A Hybrid Intelligent System for the Analysis of Atmospheric Pollution: a Case Study in Two European Regions.* Logic Journal of the IGPL, 2017. **25**(6): p. 915-937. DOI: https://doi.org/10.1093/jigpal/jzx050

22. Mrowczynska, B., et al., *A COMPARISON OF FORECASTING THE RESULTS OF ROAD TRANSPORTATION NEEDS.* Transport, 2012. **27**(1): p. 73-78. DOI: 10.3846/16484142.2012.666763

23. Kuptcova, A., et al., *Data mining workspace as an optimization prediction technique for solving transport problems.* Transport Problems, 2016. **11**

24. Abdirassilov, Z. y Sładkowski, A., *Application of artificial neural networks for shortterm prediction of container train flows in direction of China–Europe via Kazakhstan.* Transport Problems, 2018. **13**

25. Simić, D., et al., *A hybrid clustering and ranking method for best positioned logistics distribution centre in Balkan Peninsula.* Logic Journal of the IGPL, 2017. **25**(6): p. 991-1005. DOI: 10.1093/jigpal/jzx047

26. Woźniak, M., Graña, M. y Corchado, E., *A survey of multiple classifier systems as hybrid systems.* Information Fusion, 2014. **16**: p. 3-17. DOI: https://doi.org/10.1016/j.inffus.2013.04.006

27. Asencio-Cortés, G., et al., *Using principal component analysis to improve earthquake magnitude prediction in Japan.* Logic Journal of the IGPL, 2017. **25**(6): p. 949-966. DOI: 10.1093/jigpal/jzx049

28. Bishop, C.M., *Neural Networks for Pattern Recognition*. 1996: Oxford University Press. 482

29. Friedman, J.H. y Tukey, J.W., *A Projection Pursuit Algorithm for Exploratory Data Analysis.* IEEE Transactions on Computers, 1974. **23**(9): p. 881-890. DOI: https://doi.org/10.1109/T-C.1974.224051

30. Corchado, E. y Fyfe, C., *Connectionist Techniques for the Identification and Suppression of Interfering Underlying Factors.* International Journal of Pattern Recognition and Artificial Intelligence, 2003. **17**(8): p. 1447-1466. DOI: https://doi.org/10.1142/S0218001403002915

31. Corchado, E., MacDonald, D. y Fyfe, C., *Maximum and Minimum Likelihood Hebbian Learning for Exploratory Projection Pursuit.* Data Mining and Knowledge Discovery, 2004. **8**(3): p. 203-225. DOI: http://doi.org/10.1023/B:DAMI.0000023673.23078.a3

32. Seung, H.S., Socci, N.D. y Lee, D., *The Rectified Gaussian Distribution.* Advances in Neural Information Processing Systems, 1998. **10**: p. 350-356

33. A.K. Jain, M.N.M., P.J. Flynn, *Data Clustering: A Review.* ACM Computing Surveys, 1999. **31**(3)

34. Macqueen, J. *Some methods for classification and analysis of multivariate observations*. *Fifth Berkeley Symposium on Mathematical Statistics and Probability*. 1967.

35. Andreopoulos, B., et al., *A roadmap of clustering algorithms: finding a match for a biomedical application.* Briefings in Bioinformatics, 2009. **10**(3): p. 297-314. DOI: http://doi.org/10.1093/bib/bbn058

36. Zhuang, W., et al., *Ensemble Clustering for Internet Security Applications.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2012. **42**(6): p. 1784-1796. DOI: https://doi.org/10.1109/TSMCC.2012.2222025

37. Leontaritis, I.J. y Billings, S.A., *Input-output parametric models for non-linear systems Part I: deterministic non-linear systems.* International Journal of Control, 1985. **41**(2): p. 303-328. DOI: 10.1080/0020718508961129

38. Haykin, S., *Neural Networks: a Comprehensive Foundation*. 1994: Macmillan

39. Basso, M., et al., *NARX models of an industrial power plant gas turbine.* IEEE Transactions on Control Systems Technology, 2005. **13**(4): p. 599-604. DOI: 10.1109/TCST.2004.843129

40. Consejo_Unión_Europea., *REGLAMENTO(CE) 70/2012 sobre la relación estadística de los transportes de mercancías por carretera (refundición)*, in *Diario Oficial de la Unión Europea, 145*. 2012, Oficina de publicaciones de la Unión Europea