EFDA–JET–PR(08)33

A. Murari, D. Mazon, N. Martin, G. Vagliasindi, Y. Andrew, A. Meakins
and JET EFDA contributors

Exploratory Data Analysis Techniques to determine the Dimensionality of Complex Non Linear Phenomena: The L to H transition at JET as a Case Study

Exploratory Data Analysis Techniques to determine the Dimensionality of Complex Non Linear Phenomena: The L to H transition at JET as a Case Study

A. Murari[1], D. Mazon[2], N. Martin[2], G. Vagliasindi[3], Y. Andrew[4], A. Meakins[4]
and JET EFDA contributors*

JET-EFDA, Culham Science Centre, OX14 3DB, Abingdon, UK

[1]Consorzio RFX-Associazione EURATOM ENEA per la Fusione, I-35127 Padova, Italy.
[2]Association EURATOM-CEA, CEA Cadarache, 13108 Saint-Paul-lez-Durance, France
[3]Dipartimento di Ingegneria Elettrica Elettronica e dei Sistemi-Università degli Studi di Catania, 95125 Catania, Italy.
[4]EURATOM-UKAEA Fusion Association, Culham Science Centre, OX14 3DB, Abingdon, OXON, UK
** See annex of F. Romanelli et al, "Overview of JET Results ", (Proc. 22 [nd] IAEA Fusion Energy Conference, Geneva, Switzerland (2008))*

Preprint of Paper to be submitted for publication in
Nuclear Fusion

ABSTRACT

In this paper a strategy to identify and select the most relevant variables to study problems in the exact sciences, when large databases of data have to be explored, is formulated. It consists of a first exploratory stage, performed mainly with the Classification and Regression Tree method, to determine the list of most relevant signals to be used in the analysis of the phenomenon of interest. A linear followed by a non-linear correlation technique (Principal Component Analysis and Auto-associative Neural Networks respectively) are then applied to reduce the number of signals to the ones containing non redundant information. The potential of the approach is illustrated by an application to the problem of identifying the confinement regime in the Joint European Torus. The minimum set of signals has been used to train a neural network and its performance is compared with various theoretical models. The success rate of the neural network is very high and it significantly outperforms the theoretical models in terms of classification accuracy.

1. INTRODUCTION

The plasmas produced in Magnetic Confinement Fusion are very complex, open systems. They are kept out of equilibrium by injection of material and energy in order to sustain configurations capable of maximising the fusion rate. To study and control these plasmas, an increased number of diagnostics take many measurements of all the most important physical parameters. In the largest devices these diagnostics can produce very large amounts of information. In JET, for example, more than 10Gbytes of data, which is the equivalent of a 2 hour digital movie in terms of information content, can be generated per shot. JET's entire database now exceeds 40 Tbytes and projections indicate that the volume of data will be orders of magnitude larger in the next generation of devices. It is therefore increasingly difficult to analyse all this information manually.

All the aforementioned issues motivate the development of new exploratory and data analysis methods to be able to process automatically large amounts of information and derive the required knowledge from the information stored in the databases. On the other hand, the methods developed to help in this respect have been originally conceived for the private sector, to understand market and the consumer behaviour. Therefore these techniques are more suited to provide qualitative confirmation or in any case they are not explicitly designed to provide answers in the form more suited to quantitative investigation. As a consequence significant developments are often necessary for the effective application of these methods to the exact sciences. One particularly relevant issue is the determination of the minimum number of variables that are required to study a physical phenomenon. This variable selection is of the highest importance or an efficient theory formulation and model selection process.

In this paper various exploratory techniques have been revisited, combined and applied to the problem of identifying the confinement regime, whether the plasma is in the L (Low confinement) or H (High confinement) mode. The threshold to access the H mode remains one of the most important physical aspects of the Tokamak configuration, which has not been fully understood yet

[1]. Even after more than twenty five years after its discovery and even if the H mode can be achieved routinely in many devices, the exact physical mechanisms leading to the transition remain unidentified and this can have significant implications, particularly for the scaling of the power threshold to larger devices like ITER [2]. An example of a typical JET discharge with an L to H and an H to L transition is shown in figure one.

The main objective of the present analysis consists of determining the most relevant and independent signals among the thousands routinely acquired at JET to study this subject. To this end, after a pre-selection performed by the experts, a general database is traversed using Classification And Regression Tree [3] (CART) algorithms (see section two). The CART approach identifies the most relevant signals and their relative importance but it does not provide clear information about the level of correlation among these variables. Therefore a Principal Component Analysis (PCA) [4] is applied first to cluster the signals, determine the level of linear correlation between them and eliminate the redundant ones (see section three). In order to extract also the nonlinear component of the correlation, the approach of Auto-associative Neural Networks (ANNs) [5] has been adopted, which allows determining the dimensionality of the minimum space required to properly model the phenomenon. To also identify the actual variables, a couple of independent methods have been devised which provide consistent results (see section three). In order to assess the validity of the approach, a specific Neural Netwrok (NN) has been trained using only the inputs identified by the previously described steps. The performance of the network has been compared with the predictive capability of some theoretical models, confirming the choice of the signals (section four). Further developments of the approach are briefly described in the last section.

2. DATA BASE CORRELATION AND REGRESSION TREES

As is typical of all the applications of automatic learning methods, a crucial preliminary step consists of making sure that a valid database is available for the techniques to be properly trained. To this end, a set of about 50 discharges has been carefully analysed by the experts, who have provided the best possible estimates of the transition times from the L to the H mode of confinement. The available discharges cover the shot range between 55211 (21/03/2002) and 62723 (28/01/2004) and they therefore refer mainly to JET divertor configuration with the Septum. This general database, whose main characteristics are described in detail in [6], has been divided in two sets, one for the training and one for the final test of the methods. The composition of these two sets has been changed randomly to achieved a sounder statistical basis with the number of shots available. The time slices analysed in the rest of the paper belong to the interval between plus and minus 300 ms around the L-H transition.

For the discharges in this specific database, thousands of signals have been acquired. In order to reduce the dimensionality of the problem, the first step has been the selection of a subset of signals (about 40), which the experts consider as possibly linked to the confinement regime transition, given the basic understanding of the physics involved. In this indeed well known (see [7]), that

including an excessive number of variables in the exploratory phase can result in the algorithms detecting a high number of spurious correlations, which can be due to casual fluctuations in the data and are not relevant to the phenomenon under study. To assess the relative importance of the signals in this reduced dataset, consisting of the measurements and the information for every time slice whether the plasma is the L or H mode, the CART method has been applied first. The CART algorithm is a supervised approach, which traverses the entire database and tries to find which variable and which value allows best division of the time slices in two groups, one comprising the H mode phases and one the L mode phases of confinement. Since for the problem under study no signal can completely separate the two classes, after the variable with the highest exploratory value has been determined, the process is repeated for the resulting subclasses until a full discrimination has been obtained. The output of the technique is a tree in which, by construction, the most relevant signals are located toward the root. The importance of the method resides in its non linear and unbiased character. Indeed the algorithm explores the entire database in its pristine form and not even a renormalization of the data is required.

Even if the CART method is very powerful, it is not completely satisfactory for the present application. Indeed, it is well known that the nonlinear, sequential approach to the building of the tree typically produces results which are sensitive to the details of the input database. Therefore running the algorithm on different versions of the database can produce significantly different hierarchies of important signals. Therefore the most conservative use of the technique consists of applying the method to different versions of the input database and then selecting all the signals that in any of the runs the CART identifies as important. This is the approach followed and the resulting more important variables are β_N, Wmhd, Te, LiD4, Bt80, Q80, RXPL, ZXPL as summarised in table I.

Short name	**Description**
Bndiam	Beta normalised with respect to the diamagnetic energy
LID4	Outer interferometry channel
Wmhd	Magnetohydrodynamic energy
Te	Electron temperature at the intersection between interferometry vertical view and magnetic surface ($\psi = 0.8$)
RXPL	X point horizontal coordinate
ZXPL	X point vertical coordinate
Q80	Safety factor on magnetic surface ($\psi = 0.8$)
Bt80	Axial toroidal Magnetic Field at a magnetic surface ($\psi = 0.8$)

Table I : The most relevant variables for the study of the L-H transition as provided by the CART method.

3. LINEAR AND NONLINEAR CORRELATIONS

Another important aspect of the previous analysis, which must be taken into consideration, is the

fact that the CART output just identifies the most relevant signals for the problem at hand but it does not provide any information about the interdependence of these quantities. This redundancy of the variables identified by CART has been therefore analysed with the PCA approach. To this end, the signals identified with CART have been divided into two main groups: the first one includes the quantities which change significantly during the same shot and are therefore called dynamic (β_N, Wmhd, Te, LiD4), the second comprise the parameters which are almost constant during a discharge and are therefore identified as static (Bt80, Q80, RXPL, ZXPL). This distinction, which is probably of quite general value but is certainly valid for our database, allows a significant optimisation of the PCA analysis in terms of computational efforts. The degree of linear correlation within these two groups of variables is reported in figure 2 and 3. Bin charts represent the principal components and their cumulative energy is defined as the percentage of total variance in the data which is accounted for. It is calculated by summing percentage of total variance describe by each component. Among both the static and dynamic signals, we have selected the three first components which represent more than 90% of variance in the original data, as shown in figure three. Then each variable has been plotted in the subspace created by the principal components. For each principal component then only the most relevant variable has been retained. On the basis of the results of the PCA analysis we are able to cluster signals into five groups and so to have the possibility to select a subset (one in each group) of independent signals without degrading the quality of the prediction. The following signals have thus been identified: Wmhd, Te, Bt80, Q80, ZXPL.

The correlation determined by the PCA method allows reducing significantly the number of relevant signals. On the other hand the PCA is a linear technique and therefore it extracts only the linear correlation between the variables. To quantify the remaining degree of non linear correlation, the architecture of ANNs [5] has been adopted. These networks have the structure described in figure four and they are used to match inputs on themselves. By reducing progressively the number of neurons in the central layer and monitoring the identity mapping of the network, it is possible to determine the dimensionality of the problem under study. Indeed, when removing an additional neuron in the central layer causes a significant increase in the errors, the degradation in performance indicates that all the neurons are essential to properly model the phenomenon. Applying this approach to the five signals Wmhd, Te, Bt80, Q80, ZXPL indicates very clearly that the minimum number of neurons in the intermediate layer is three.

At this point, identified the dimensionality of the problem, a strategy is required to finally select the subset of really indispensable variables. A logical approach consists of calculating the total of the weights connecting each input to the three neurons in the bottleneck layer. Such technique allows the determination of both the correlations between variables and their importance to the L-H transition. Indeed, as reported in table II, the five variables Wmhd, Te, Bt80, Q80, ZXPL can be clearly grouped into three different clusters, (Wmhd/Te), (Q80/Bt80) and ZXPL, depending of the strength of the weights connecting them to the bottleneck layer.

	Coefficients NLPCA				
Neuron	**Wmhd**	**Te**	**Q80**	**ZXPL**	**Bt80**
A	1	0.85	0.4	0.15	0.45
B	1	0.85	0.35	0.15	0.45
C	1	1	0.4	0.2	0.55

Table II: Relative importance of each variable as derived from the sum of the weights connecting each input to the bottleneck layer. The column Neuron indicates the three neurons of the bottleneck layer and so the intrinsic dimension.

To test the validity of this original approach of summing the weights from the inputs to the neurons in the bottleneck, a brute force method has been followed. First, four neural networks have been trained on the database: one using all the signals and three eliminating in turn one of the clusters. This set of networks allows checking the degradation in the prediction accuracy for each of the clusters. From the results reported in table (III-a), it appears very clearly that the relative importance of the three clusters identified with the sum of the weights is fully confirmed. We can also note that the best result in terms of performance for regime identification is obtained with the full set of the five variables, confirming at the same time the validity of the technique to identify the most relevant signals.

	Variables					
Number of variables	**Wmhd**	**Te**	**Q80**	**Bt80**	**ZXPL**	**Error**
5	1	1	1	1	1	2.2%
3	X	X	1	1	1	**35.4%**
3	1	1	X	X	1	**8.2%**
4	1	1	1	1	X	**4.6%**

*Table III(a): Brute force technique to confirm the relative importance of the three clusters of variables identified with the sum of the weights. The **X** indicates the inputs omitted.*

Then, six neural networks have been trained on the database: one with all the five signals (Wmhd, Te, Bt80, Q80, ZXPL) and four without one of the inputs each. This set of networks allows checking the importance of variables in the same cluster and the results are reported in table III-b.

Wmhd	**Te**	**Q80**	**Bt80**	**ZXPL**	**Error**
1	1	1	1	1	**2.2%**
X	1	1	1	1	**4.8%**
1	X	1	1	1	**11.2%**
1	1	X	1	1	**5.3%**
1	1	1	X	1	**6.7%**

Table III(b): Brute force technique using neural network prediction to check the relative importance of each variable in the same cluster

It is worth mentioning that the NN used to provide the results reported in table III-a and III-b are all traditional multilayer perceptrons, trained with the back propagation method.

The selection of the five most important signals being confirmed, the next step consists of determining the minimum number of inputs, which are still sufficient to characterise the L to H transition, without an excessive degradation in performance. The motivation resides in the need to reduce the number of variables to three, in order to compare the classification capability of neural networks, using these three quantities, with the most widely accepted theoretical models, which the typically utilise four quantities (see section four).

At this point, two possible alternatives can be considered for the choice of the most relevant variables. One would consist of eliminating the two signals with the lowest sum of the weights connecting them to the three neurons in the bottleneck layer. On the other hand, given the strong correlation between the quantities in the first two clusters, it is not evident that the alternative choice of selecting one signal per cluster should be discarded a priori. To identify the most appropriate alternative, two different NNs have been trained. One NN uses one signal in each cluster and the second uses only the three variables with the highest weights from the first two clusters. Table III-c summarizes choice of the two different subsets of more relevant variables and the corresponding error rates.

Wmhd	**Te**	**Q80**	**Bt80**	**ZXPL**	**Error**
X	1	X	1	1	**5.7%**
1	1	X	1	X	**5.2%**

Table III(c): The performance of the ANNs using the two alternatives of selecting the there most relevant variables for the analysis of the L-H transition

This exhaustive test has confirms the results of the previous method. The variables surviving this selection are finally Te, Wmhd and Bt80. For clarity sake, the described process of variable reduction is represented graphically in figure five.

4. CLASSIFICATION CAPABILITIES OF THE IDENTIFIED VARIABLES AND COMPARISON WITH THEORETICAL MODELS

To summarise the results of the previous sections, with the three most relevant variables identified, a traditional multiplayer perceptron can discriminate the L from the H mode of confinement in our database with a success rate of bout five percent (last row of table III-c). This performance is very high an in a certain sense testify on its own the quality of the selection process developed. In any case, to further confirm the high information content of the variables identified with the described approach, the performance of the network has been compared with the classification capability of various theoretical models. Only models which do not depend on specific scale lengths have been considered since they use more reliable quantities less dependent from complex calculations and less affected by noise. The model developed in [8,9] assumes that the drift wave turbulence is the

main transport mechanism in the L mode and the resistive skin effects are considered the main stabilising factor. The analytical function provided by the model depends on the electron temperature Te, the total axial toroidal magnetic field B, the safety factor q, the atomic mass A of the plasma, the major radius R of the machine and the ratio τ between Te and the ion temperature Ti. In [10] a model based on drift wave instabilities has been particularised for two different assumptions about the transport (convective or conductive) in the Scrap Off Layer (SOL). Two different equations are therefore arrived at for the temperature threshold, one for convective and one for conductive transport in the SOL and they involve the same variables as the previous model excluding τ and including the electron density n. In the model described in [11], also the effects of the neutral particles in the core are taken into account and the variables appearing in the final expression are the same as model [10] plus the plasma radius a of the plasma. A more detailed discussion of these models can be found in [6] and their failure rates for the database used in this paper are shown in table IV.

The analytic expressions for the scaling of the electron temperature, which these models provide, depend on similar variables but not exactly the same as the ones selected by the method proposed in this paper. The significantly higher success rate of the NNs, trained with the variables identified with our approach, supports the validity of the proposed method. This seems also to indicate that these theoretical models do not use the most significant quantities. In this respect and in order to dissipate possible doubts that the difference in performance between the theoretical models and the NNs is really due to the choice of the input signals, a network has been trained with the same inputs as the theoretical model with the best success rate [10]. The failure rate of this network is reported in table V and is similar to the one of the theoretical model. This indicates that the improvement in performance of the networks can really be attributed to the choice of the input signals and not on the inherent classification power of the networks themselves.

A further test of the proposed method capability, to discriminate the most inportant variables, has been performed using a mixture of real and synthetic signals. In addition to the three quantities Wmhd, Te, Bt80 the autoassocitive neural network has also been given as input random signals completely uncorrelated with the L to H transition. The resulting weigths of the ANNare clearly much lower for these synthtetic signals and the quantities with a real physical meaning are again easily identified.

	NN using variables from theoretical model [10]	NN using the three most relevant variables
Failure rate	18.8%	5.2%

Table V: Failure rate of the NN using the variables of model [10] non collisional and the most relevant variables obtained with the proposed selection procedure.

CONCLUSIONS AND FURTHER DEVELOPMENTS

The results presented in the previous section indicate quite clearly that the proposed procedure to

select the most relevant variables is quite effective. On the basis of these variables the network provides performance, which are significantly superior not only to the theoretical models but also to neural networks trained to use the same signals as the theoretical models. Furthermore, their high success is obtained with only three signals, compare to the four used in the most performing theoretical model. Therefore the described method to select the most relevant variables can be considered a validated paradigm, which can be probably extended to other physical problems. Particularly innovative is the interpretation of the ANN weights, which has been proved to allow a fast identification of the most relevant inputs. The fact that the vast majority of the theoretical models available in the literature do not seem to use the most informative quantities for the classification emphasizes also the importance of exploratory methods, like the one described in this paper, as a preliminary or complementary step to the process of model development.

An obvious extension of the present work could include the transition back to the L mode from the H mode. Preliminary evaluation indicates that the best choice of variables is different from the one for the L to H transition. A more involved exploratory data analysis approach could contribute to clarify this point and determine also the similarities and differences between these two transient phenomena.

From a methodological point of view, it would be interesting to apply the same procedure to even more nonlinear phenomena. The step involving the PCA and the ANN could require some additional refinements.

ACKNOWLEDGEMENTS

This work, supported by the European Communities under the contract of Association between EURATOM/CEA and ENEA Consorzio RFX, was carried out within the framework of the European Fusion Development Agreement. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

REFERENCES

[1]. Wagner, F. et al., Physical Review Letters **49**, 1408 (1982).

[2]. Y Andrew et al "H-mode access on JET and implications for ITER" *Plasma Phys. Control. Fusion* **50** No 12 (December 2008) 124053 (8pp)

[3]. Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. 1984 *Classification and Regression Trees* (Belmont, CA:Wadsworth Inc.) (1993, New York: Chapman and Hall)

[4]. J.Lattin, J.D.Carroll, P.E.Green "*Analyzing Multivariate data*" Duxbury Applied Series, Thomson, 2003

[5]. Nonlinear Principal Component Analysis Using Autoassociative Neural Networks / aut. Kramer Mark A.// AIChE.Massachusetts Institute of Technology, Cambridge: Laboratory for Intelligent Systems in Process Engineering, Dept. of Chemical Engineering, 1991.2: Vol. **37**.-p233-243.

[6]. Meakins, A. J. (2008) "*A Study of the L-H Transition in Tokamak Fusion Experiments*". PhD Thesis. Imperial College London.

[7]. K.P.Burnham and D.R.Anderson "*Model selection and multimodel inference*" Second Edition, Springer Science and Business Media, 2002
[8]. Chankin A V 1997 Plas. Phys. Cont. Fus. **39** 1059
[9]. Chankin A V, Saibene G 1999 Plas. Phys. Cont. Fus. **41** 913
[10]. Kerner W et al. 1998 Contrib. Plas. Phys. **38** 118
[11]. Rogister A L 1994 Plas. Phys. Cont. Fus. **36** A219
[12] . Y Andrew et al "H-mode access on JET and implications for ITER" *Plasma Phys. Control. Fusion* **50** No 12 (December 2008) 124053 (8pp)

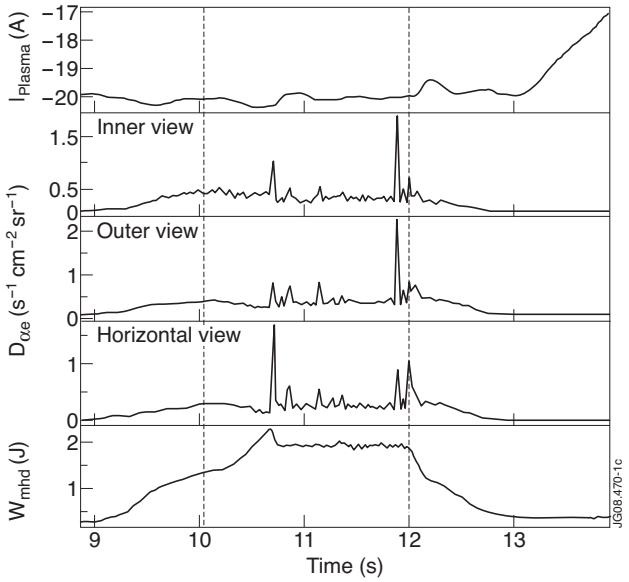

Figure 1. Time evolution of various signals for a JET discharge with an L and H mode phase. The L to H transition is at 18s, the H to L transition is at 25s as indicated by the last variable "Mode" , which assumes value one when the plasma is in the H mode.

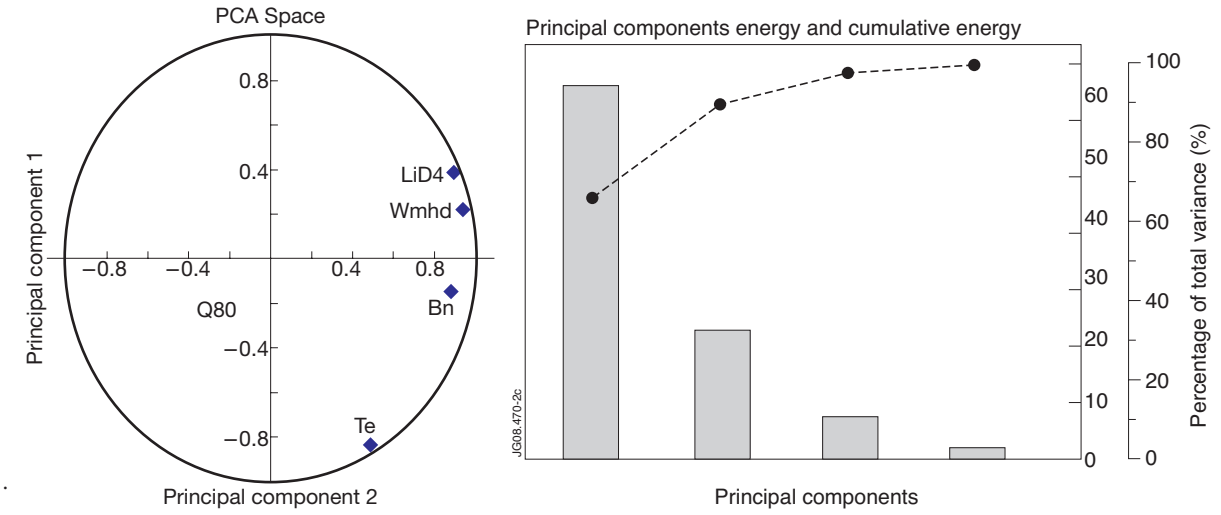

Figure 2: Correlation between the dynamic signals ≤n, Wmhd, Te and LiD4. The bin chart gives the energy of the principal components and the red line show the cumulative energy in term of percentage of total variance in data. The plot below illustrates the degree of correlation of the variables with respect to the principal components (one and two in the figure)

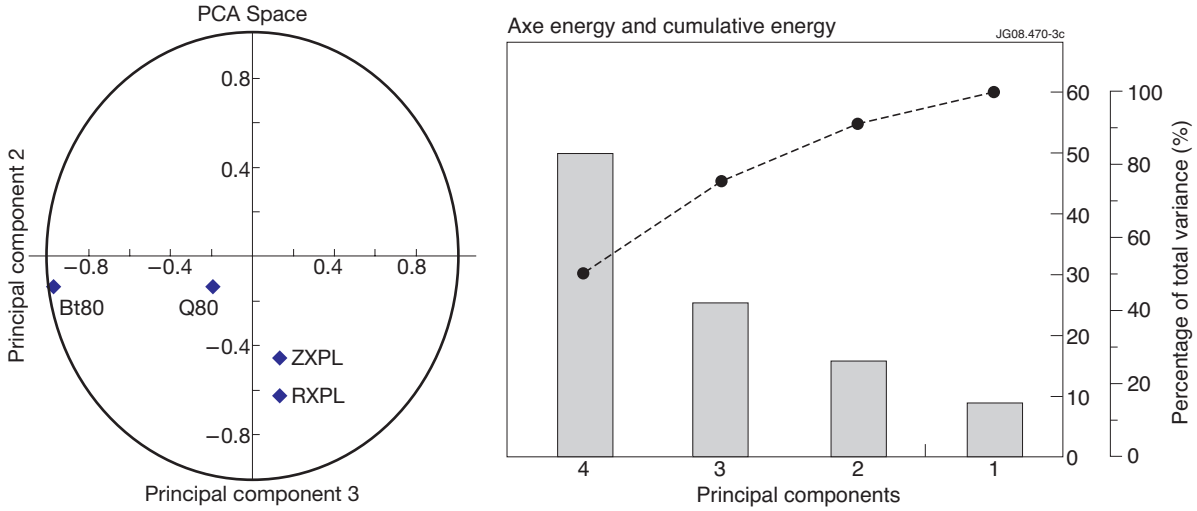

Figure 3: Correlation between the static signals Bt80, Q80, RXPL, and ZXPL. The top bin chart gives the energy of the principal components and the red line show the cumulative energy in term of percentage of total variance in data. The plot below illustrates the degree of correlation of the variables with respect to the principal components (two and three in the figure)

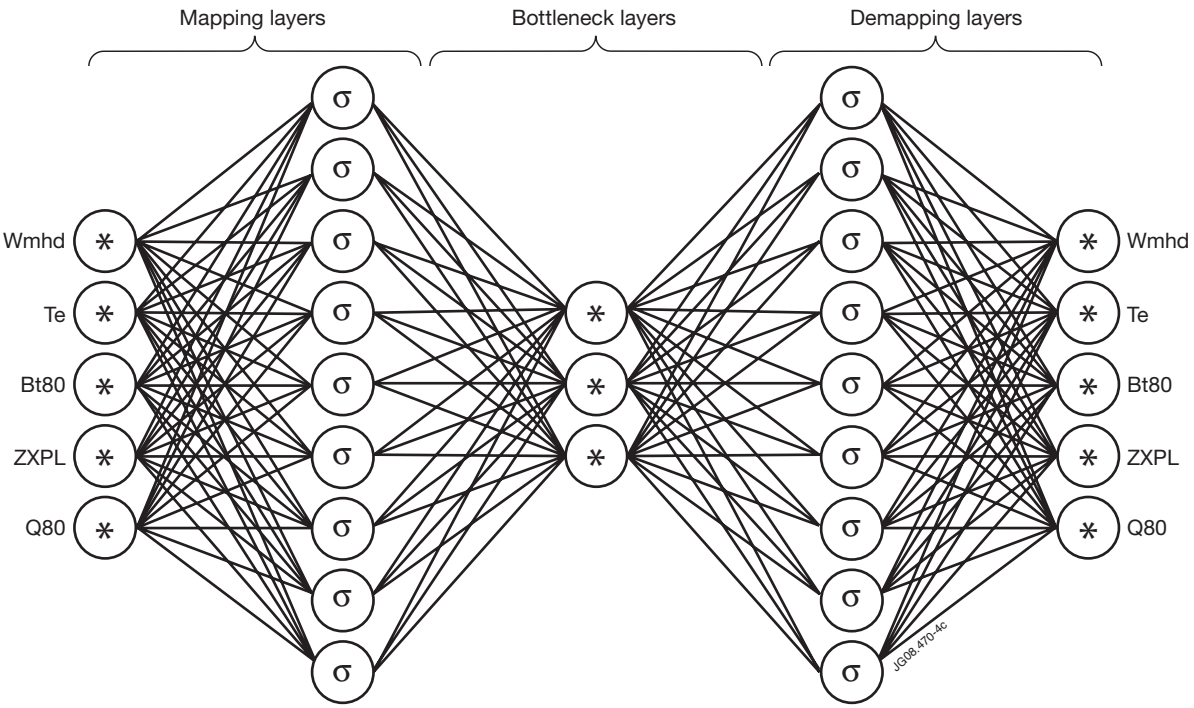

Figure 4: The topology of Autoassociative Neural Networks

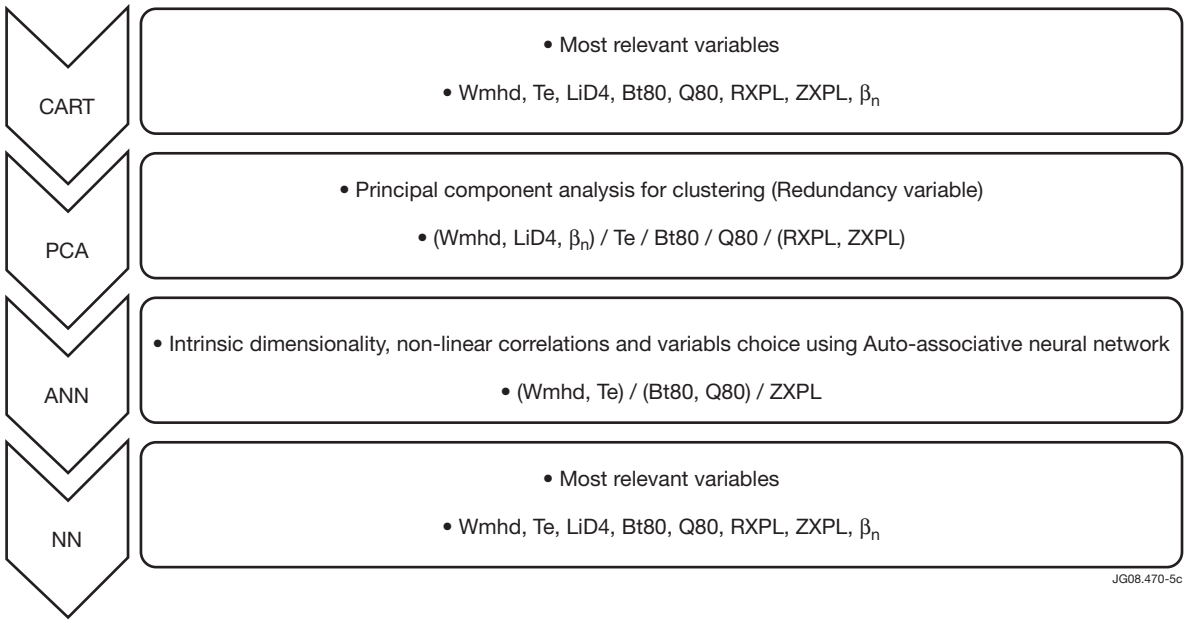

Figure 5: The main steps of the proposed exploratory methodology to extract the most relevant signals from the database.