EFDA–JET–CP(08)04/20

A. Murari, J. Vega, G. Vagliasindi, J.A. Alonso, D. Alves, R. Coelho, S. DeFiore,
J. Farthing, C. Hidalgo, G.A. Rattá and JET EFDA contributors

Recent Developments in Data Mining and Soft Computing for JET with a View on ITER

“This document is intended for publication in the open literature. It is made available on the understanding that it may not be further circulated and extracts or references may not be published prior to publication of the original when applicable, or without the consent of the Publications Officer, EFDA, Culham Science Centre, Abingdon, Oxon, OX14 3DB, UK.”

“Enquiries about Copyright and reproduction should be addressed to the Publications Officer, EFDA, Culham Science Centre, Abingdon, Oxon, OX14 3DB, UK.”

Recent Developments in Data Mining and Soft Computing for JET with a View on ITER

A. Murari [1], J. Vega[2], G. Vagliasindi[3], J.A. Alonso[2], D. Alves[4], R. Coelho[4], S. DeFiore[3], J. Farthing[5], C. Hidalgo[2], G.A. Rattá[2] 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]Asociación EURATOM-CIEMAT para Fusión, CIEMAT, Madrid, Spain
[3]Dipartimento di Ingegneria Elettrica Elettronica e dei Sistemi-Universitá degli Studi di Catania, 95125 Catania, Italy
[4]CFN, Associacao IST/EURATOM, 1049-001 Lisboa, Portugal
[5]EURATOM-UKAEA Fusion Association, Culham Science Centre, OX14 3DB, Abingdon, OXON, UK
** See annex of M.L. Watkins et al, "Overview of JET Results ",*
(Proc. 21 st IAEA Fusion Energy Conference, Chengdu, China (2006)).

Preprint of Paper to be submitted for publication in Proceedings of the
25th Symposium on Fusion Technology, Rostock, Germany
(15th September 2008 - 19th September 2008)

ABSTRACT

In order to handle the vast amount of information collected by JET diagnostics, which can exceed 10 Gbytes of data per shot, a series of new soft computing methods are being developed. They cover various aspects of the data analysis process, ranging from information retrieval to statistical confidence and machine learning. In this paper some recent developments are described. History effects in the plasma evolution leading to disruptions have been investigated with the use of Artificial Neural Networks. New image processing algorithms, based on optical flow techniques, are being used to derive quantitative information about the movement of objects like filaments at the edge of JET plasmas. Adaptive filters, mainly of the Kalman type, have been successfully implemented for the online filtering of MSE data for real time purposes.

1. INTRODUCTION

Tokamak plasmas are complex and nonlinear systems kept out of equilibrium by powerful external heating systems. In present day devices, to provide the signals required for the interpretation and the control of the experiments, diagnostics have become very complex, to the point of sometimes constituting independent experiments in their own right. They can also produce impressive amounts of data; in the last set of campaigns, JET diagnostics have produced a maximum of more than 10 Gbytes of data per shot and the volume of information is bound to increase in the next generation of devices like ITER (JET whole data base exceeds already 42 Tbytes of data). The explosion of data has become particularly relevant in the last years due to the increased use cameras, both visible and infrared, some of which can produce Gbytes of data per shot. On the other hand, since hot fusion plasmas can seldom be directly probed by diagnostics, the inference of internal parameters must be based on quantities available only outside the plasma, like radiation, escaped particles and external magnetic fields. This leads to complex inversion problems for the interpretation of the data and to the need of new methods for data mining. Therefore data analysis for physical studies requires addressing, among others, at least the following main issues: a) retrieval of the required information and exploration of the database to identify hidden correlations (see section two) b) efficient image processing methods (see section three) c) the development of analysis techniques compatible with feedback control requirements (see section four). The relevance of the presented methods for the operation of ITER is revised briefly in the last section five.

2. INFORMATION RETRIEVAL AND DATA MINING

The first step of any analysis procedure consists of retrieving the information relevant to the physical phenomenon under study. In massive databases like JET's, this cannot be performed efficiently by traditional manual methods. Therefore new approaches are being developed to reduce the amount of data by adaptive sampling [1] and lossless compression [2]. Another significant advance is the development of techniques to store data according to technical and scientific criteria, instead of time intervals and pulse numbers. Since visual inspection is a routine activity in plasma physics, a

"pattern oriented" approach to data analysis is intensively pursued. In particular "structural pattern recognition" allows selecting, with a cursor on a simple interface, the signal or some of its parts and then, in a matter of milliseconds, the algorithms produce the list of shot numbers, time intervals and signals in which the same or similar structures are present [3,4]. An important recent development is the successful extension of this approach to images [5].

At the level of analysis, data mining, the problem of extracting useful hidden correlations from massive databases, is a major time consuming activity for many scientists. Since fusion plasmas, in addition to being very complex, are also often affected by significant uncertainties, it can be very difficult to obtain the required "knowledge" from the available signals, even after the relevant information has been retrieved from the database. The traditional identification techniques, used in other fields to determine dynamical models of the systems under study, are not easily applicable. To help in the direction of deriving physical information from the signals and to cope the high level of uncertainty in the data, several "soft computing" methods are being pursued. Fuzzy Logic [6], Support vector Machines [7], Classification and Regression Trees (CART) [8] and Artificial Neural Networks (ANNs) [9] are among the most systematically pursued approaches. The first three are being introduced to formalise the knowledge of the experts in fields like disruption prediction and regime identification; the fourth has been used for many years to handle problems for which efficient algorithms are not available. Recently ANNs have been used as exploratory data analysis tools to determine whether certain phenomena depend from the historical evolution of the discharge and not only from the plasma state at a single point in time. In particular they have been applied to the analysis of disruptions. The nine signals most relevant for disruption description have been identified by the experts and confirmed with the unbiased and nonlinear CART approach, as described in [10], and they are summarised in table I.

These signals have been then given as inputs to sets of ANNs: the first ANN of a set has been trained with only signals belonging to one time slice, the second ANN has been trained also with the data of the previous time slice and the last with the two previous time slices. The interval between time slices is typically of 20 ms and is mainly dictated by the resolution of the diagnostic signals available. The signals of the various time slices have been multiplied by the weights decreasing with increasing distance to the disruption, to reflect the well known lower predictive power of the signals at earlier times. One example of the results is reported in figure 1 for a series of ANNs trained starting 100ms before the discharge. In this case the chosen weights are 1 for the time slice at 100ms, 0.9 for the time slice at 120ms and 0.8 for the time slice 140 ms before the disruption (these values have been optimised empirically). The results reported in this paper refer to a database of 512 discharges has been analysed (67% used for the training)

The ANNs performance improves when the earlier time slices are provided as additional inputs and, even if the absolute increase is small in percentage terms, the trend is quite consistent and has been confirmed in all the cases analysed. Indeed various training and test sets have been randomly chosen to reduce the chances of ay bias in the choice of the data and the error bars in the figures

account for the resulting uncertainties in the results. These results seem to indicate that some relevant information is present in the history of the signals, since the success rate of the ANNs is improved by including earlier time slices in the list of inputs, which are well known to have per se a lower information content (and therefore a decrease in the performance of the ANNs would be expected if this historical information was not in the signals). This is more evident for some specific type of disruptions, in particular for the ones triggered by a transition from the H to the L mode of confinement. This is illustrated in figure 2. The success rate improves from 93.5 to 95.5 % introducing one additional time slice, increases to 95.7 with one more time slice and then starts decaying again. This is a trend of significant relevance since it is outside the statistical uncertainties. It is also a phenomenology easy to interpret in terms of historical information, since this type of disruption depends directly on the earlier evolution of the plasma.

3. IMAGE PROCESSING

In the last years significant efforts have been devoted to improving JET imaging capabilities, both in the visible and infrared part of the spectrum. Some new cameras have now the potential to produce Gbytes of data per shot and therefore new tools are required to analyse and interpret all this information. One recurrent problem, common also to other devices, consists of trying to quantify the movement in the three dimensional space of objects seen by a single camera and therefore by a single point of view. Of course the information available in bidimensional images is insufficient to derive the displacement of objects in physical space but, if some specific hypotheses are satisfied, quantitative indications can be derived for example with the method of the so called “optical flow” [11]. This approach concentrates on the evolution of the optical emission and in certain conditions this emission can provide indications on the movement of the object generating the emission. A potential application is the propagation velocity of filaments detected at the edge of JET ELMy H mode plasmas as shown in figure 3. Assuming that the filaments move more or less like a rigid body and that the difference in their position between frames is not too high, the time evolution of the intensity can be written as:

$$\frac{dI}{dt} = \partial_t I + v(x)\,\nabla I$$

where I is the intensity of the image pixels and v their velocity. In the hypothesis that the intensity of the emission remains constant the velocity of the filament can be derived by the relation

$$v = -\frac{\partial_t I}{\partial_x I}$$

The extension of this approach in two dimensions consists of minimising a cost function of the form [11]:

$$E(v) = \sum_s \left(\partial_t I + \nabla I \right)^2 + \alpha \left(|\nabla v_t|^2 + |\text{—vy}|^2 \right) \Big)$$

This quadratic functional is a conceptually natural extension of the simple monodimensional case and it is easy to handle numerically since its derivative turns out to be a linear function. On the other hand it is not a robust quantifier since, being quadratic, it tends to weight excessively any error in the data (like sources of light, discontinuities in the movements and so forth). Therefore a different functional, based on Lorentzian functions, has been chosen for the analysis of JET data. Even if additional upgrades have to be implemented in order to improve the method and confirm the robustness of the conclusions, the first results are quite encouraging. An example of preliminary analysis performed with this approach is reported in figure 4, where the movement of a filament against the background of JET poloidal limiters is shown. The estimated velocity of propagation, in this specific case, is of the order of 1 km/s.

4. DATA ANALYSIS FOR CONTROL

The higher energy content of the plasmas, the increase in the sophistication of the configurations and the need to move towards much longer discharges all need the development of more advanced feedback control schemes. In this framework, the requirements in terms of real time signal processing are also becoming more stringent. An example of the difficulties and complexities of this task are well represented by the case of the Motional Stark Effect (MSE), a very important diagnostic to derive the internal profile of the plasma current. The information about the current is derived by measuring the pitch angle of the magnetic field (g), which is linked to the amplitudes *A* of particular spectral components related to the modulation frequencies of the detection system by the equation [12]:

$$\tan(2\gamma(t)) \frac{C_{21}A_{DC}(t) + C_{22}A_{23}(t) + C_{23}A_{46}(t) + C_{24}A_{40}(t)}{C_{11}A_{DC}(t) + C_{12}A_{23}(t) + C_{13}A_{46}(t) + C_{14}A_{40}(t)}$$

The diagnostic has become routine at JET and provides a lot of useful information but the quality of the measurements can be strongly affected by the ELMs. Therefore various approaches have been attempted to mitigate the negative influence of these instabilities, which are believed to generate spurious radiation which is collected by the MSE front end optics.

The best filtering so far has been obtained by an adaptive filter of the Kalman type. These filters minimise the error covariance between the measurements and the linear model (in our case the model is derived by the hardware configuration of the diagnostics and consists of a series of sinusoidal frequencies corresponding to the A components of the previous formula). In our approach, basically, the gain of the Kalman filter is adaptively reduced when the difference between the empirical signal and the model is too high, indicating the presence of the spurious radiation due to the ELMs. The quality of the obtained results can be seen in figure 5, which shows the comparison between the output of the Kalman filter and a single phase lock-in amplifier with an apodization function implementing the hanning window. The superior smoothing achieved by the Kalman filter is quite significant and constitutes a very useful improvement in the quality of the signals provided in real time.

5. THE PERSPECTIVES FOR ITER

Many of the methodologies being developed in JET will become routine in ITER, since the problems presented by the next step devices in terms of data analysis will be more severe. The amount of data collected is expected to be significantly higher since already the IR cameras for surveillance are estimated to produce a couple of Tbytes of data per shot. The energy content of the devices will be also higher and the discharges will have to be sustained for much longer periods. These aspects make more pressing the need for more sophisticated data analysis tools and feedback schemes.

ACKNOWLEDGEMENTS

This work, supported by the European Communities under the contract of Association between EURATOM/[,Ä?.], 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]. G.De Arcas et al *"Self adaptive sampling rate for data acquisition in JET's correlation reflectometer"* submitted to Rev.Sci. Instrum.

[2]. J. Vega et al. Rev. Sci. Instrum., Vol. **67**, No. 12, December 1996

[3]. J. Vega et al Fusion Engineering and Design. 83 (2008) 132-139.

[4]. S. Dormido-Canto, G. Farias, J. Vega, R. Dormido, J. Sánchez, M. Santos et al. Rev. Sci. Ins. **77** (2006)

[5]. J. Vega et al *"Intelligent Technique to Search for Patterns within Images in Massive Databases"* submitted to Rev.Sci. Instrum.

[6]. G.Vagliasindi et al IEEE Transactions on Plasma Science, Vol. 36, Issue 1, Part 2, Feb. 2008

[7]. Cannas B. *et al* 2006 *Support vector machines for disruption prediction and novelty detection at JET Proc. 24th Symp. On Fusion Technology (SOFT 2006) (Warsaw, Poland, 11–15 September 2006)*

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

[9]. D. Rumelhart, G. Hinton, and J. McClelland. Learning internal representations,1986

[10]. A.Murari et al Nucl. Fusion **48** (2008) 035010

[11]. Horn, B.K.P. and Schunck, B.G., *Artificial Intelligence*, vol 17, pp 185-203, 1981.

[12]. R. Coelho, D. Alves, N.Hawkes, M. Brix and JET EFDA contributors "*Real-time data processing and magnetic field pitch angle estimation of the JET motional Stark effect diagnostic based on Kalman filtering*", submitted to†Rev of Scient. Instr.

SIGNAL NAME	UNIT
Plasma current *Ipla*	[A]
Mode Lock Amplitude *Loca*	[T]
Plasma density *Dens*	[m^{-3}]
Total Input Power P_{inp}	[W]
Plasma Internal Inductance L_i	
Stored Diamag. Energy Derivative dW_{dia}/dt	[W]
Safety factor at 95% of minor radius q_{95}	
Poloidal beta β_p	
Net power P_{net}	[W]

Table I: List of the signals used as predictors for the ANNs. These quantities have been identified as the most important or the prediction of disruption using the CART algorithm.

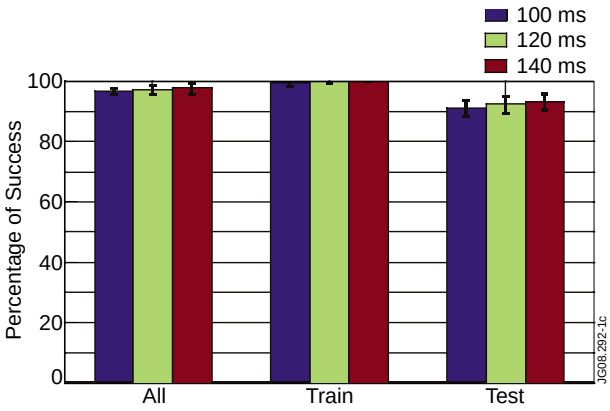

Figure 1: Improved performances of ANNs with historical inputs. The colour code indicates the time before the disruption the various sets of inputs were taken. Train indicates the training set, Test the independent set of discharges used for the test and finally All is the sum of the two sets.

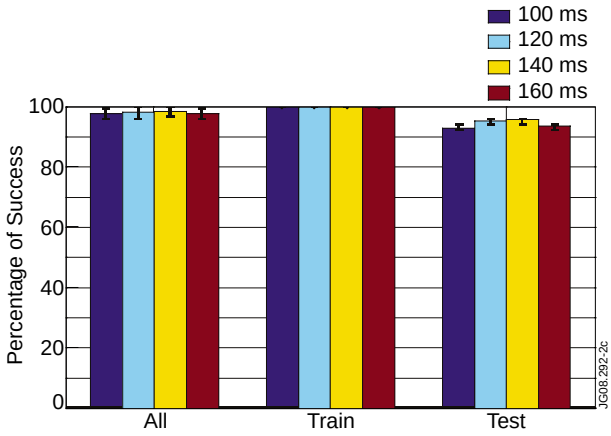

Figure 2: Improved performances of ANNs with historical inputs for the case of disruptions triggered by H-L transitions. The nomenclature in the figure and the method to randomly select the various sets of discharges are the same as in figure 1.

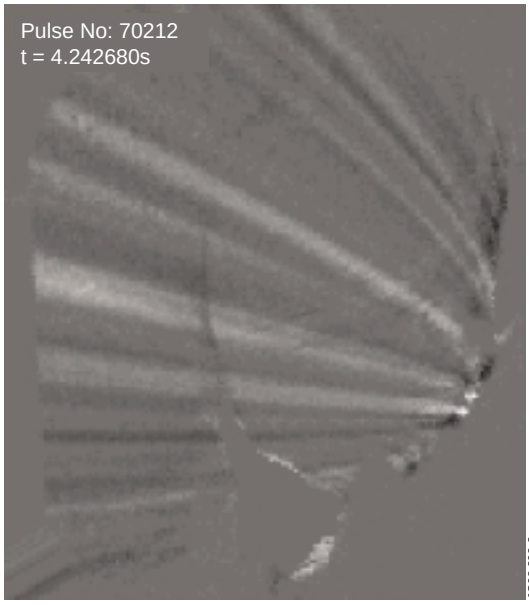

Figure 3: Filaments seen with the fast visible camera during a Type 1 ELMy H mode phase.

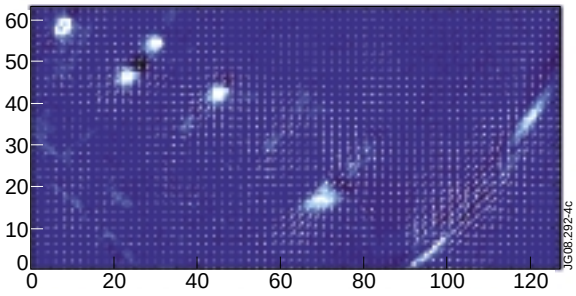

Figure 4: Application of optical flow to filaments moving along JET poloidal limiters (Pulse No: 69903).

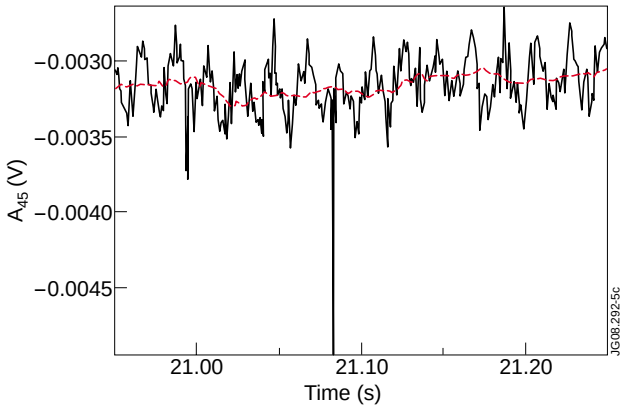

Figure 5: Comparison of the signals obtained with the Hanning apodization window moving average (black line) and Kalman filter(red line) showing the superior quality of the second solution.