EFDA–JET–PR(09)11

J. Vega, A. Murari, S. González and JET EFDA contributors

An Universal Method for Automatic Event Location in Waveforms and Video-Movies: Applications to Massive Nuclear Fusion Databases

An Universal Method for Automatic Event Location in Waveforms and Video-Movies: Applications to Massive Nuclear Fusion Databases

J. Vega, A. Murari, S. González and JET EFDA contributors*

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

[1]*Asociación EURATOM/CIEMAT para Fusión. Avda. Complutense, 22. 28040 Madrid, Spain*
[2]*Associazione EURATOM-ENEA per la Fusione, Consorzio RFX, 4-35127 Padova, Italy*
** See annex of F. Romanelli et al, "Overview of JET Results",*
(Proc. 22nd IAEA Fusion Energy Conference, Geneva, Switzerland (2008)).

Preprint of Paper to be submitted for publication in
Review of Scientific Instruments

ABSTRACT

Big physics experiments can collect Tbytes (even Pbytes) of data under continuous or long pulse basis. The measurement systems that follow the temporal evolution of physical quantities translate their observations into very large time-series data and video-movies. The off-line analysis of massive amounts of waveforms and video-images is typically affected by two important limitations. Firstly, there are no methods for automatic selection to focus the analysis just on the data of interest for a particular study. Secondly, some physical events take place in a random and very infrequently way and, again, there are not automatic mechanisms to identify and locate these singular events. This article describes a universal and automatic technique to recognize and locate inside waveforms and video-films both signal segments with data of potential interest for specific investigations and singular events. The method is based on regression estimations of the signals using Support Vector Machines. A reduced number of the samples are shown as outliers in the regression process and these samples allow the identification of both special signatures and singular points. Results are given with the database of the JET fusion device.

1. INTRODUCTION

In general, the measurement systems in physics are transducers that convert their observations into electrical signals to be digitized and stored for off-line analysis. The term ‘signal’ is used in this article to represent any kind of output (raw or processed) from a data acquisition system. Therefore, signals can be functions of one variable (like waveforms) and functions of two or more variables (like video-images).

The information content in the signals (i.e. physical knowledge) resides not only in their amplitudes but also in the way in which the amplitude changes between adjacent samples and even in their particular shapes. For example, abrupt/slow variations in Temporal Evolution Signals (TES) are signatures of big/small alterations in the system state. In the case of non temporal evolution signals, changes of amplitude and the presence of specific forms can identify particular phenomena such as characteristic emission lines in spectroscopy, hot spots in infra-red images and specific types of power spectra.

Typically, the measurement systems transform reproducible physical behaviours into similar morphological structures (patterns) in the signals. This implies that the different phenomenology present in the physical systems can be identified through the recognition of patterns. A major problem in data analysis is the identification and location of these patterns inside large data repositories. Traditional data retrieval methods in science use an experiment number and a signal name to obtain the signal samples. The location of events is usually carried out by non-automatic methods (for example visual data analysis to discover the presence of patterns). Obviously, this procedure is far from being an optimal searching method and presents two main drawbacks. First of all, the analysis process is typically very heavy in terms of manpower and therefore often not enough cases are analysed to draw conclusions with a sufficient statistical basis. Secondly, the continuous reliance on human intervention increases the probability of human errors, resulting in a significant lack of reproducibility in the analyses.

The automatic search of patterns in massive databases has recently been considered for nuclear fusion environments to perform data retrieval according to physical criteria [1]. Instead of experiment number and signal name, the input parameters to look for data are the signal patterns that characterise a particular physical behaviour. The outputs are no longer the signal samples but, firstly, the numbers that identify the experiments whose signals contain the required pattern and, secondly, the explicit pattern location (temporal or spatial). Moreover, these techniques provide a similarity factor (a distance in the mathematical sense) to evaluate how similar the signals are to the reference pattern. This factor allows ordering the search results according to the structural resemblance of the signals. Specific applications of intelligent data retrieval methods have been developed [2] for the databases of two fusion devices, the JET Tokamak [3] (the world's greatest fusion device) and the TJ-II stellarator [4]. These particular applications provide a Graphical User Interface (GUI) to display waveforms or images. The pattern to search is selected by the user on a waveform or image.

It should be noted that data retrieval based on pattern recognition techniques requires human intervention to define specific structural forms to search. However, automatic methods without human participation would be of big help not only for data retrieval but also for the analysis of massive databases. On the one hand, they would allow focusing the attention on interesting signal slices and discarding irrelevant signal segments. On the other hand, these methods could also be used for the exact location of singular events in the signals.

This article describes a novel and universal technique (UMEL, Universal Multi-Event Locator) for the automatic location of events in waveforms and video-movies. Here, 'universal' has twofold meaning. First, it indicates that the technique is the same independently from the specific structural patterns of the signals and, therefore, it does not depend on particular physical events. Second, 'universal' stresses the fact that the technique is the same regardless of the type of signals, which means that it can be applied exactly in the same way to waveforms, images or any kind of multivariate signals with arbitrary number of dimensions.

The main achievement of the techniques described in this paper is their ability to 'automatically locate' events. In the case of temporal evolution signals, the existence of events is shown by the presence of specific patterns (spikes, transients or special structures) that are precisely located in time. If non-TES signals are considered, also the existence of particular signatures such as edges, peaks or textures denote the occurrence of well-located patterns inside the signals. It is clear that all the mentioned patterns are characterized by local information either in the time (or space) domain or in the frequency (or spatial frequency/wave number) domain or in both. The goal of UMEL is the automatic location of any kind of special footprint in the signals.

Typically, UMEL provides a double capability. On the one hand, it can be used as a filter to recognize signal segments with relevant behaviours. On the other hand, it can be applied as an exact locator of singular points within signals.

It should be mentioned that the use of UMEL as a filter allows establishing the presence of physical events but perhaps without identifying the particular type of event. This has to be accomplished in a second phase by specific experts.

The capabilities of UMEL as an exact locator of singular points have a very crucial value. UMEL can locate physical events in a high number of experiments and, therefore, big databases with large statistical significance can be created in an automatic way. This is extremely helpful in the case of behaviours that take place in random or non-periodic basis.

UMEL tries to automatically find event signatures inside a signal. The searching process is the core of the technique. UMEL bases its searching capabilities on a specific regression estimation method. Among the several possible regression techniques, the choice has been the use of learning systems, specifically Support Vector Machines (SVM) [5]. The reason of utilizing learning systems rests on the fact that the resulting fits incorporate the maximum information content provided by the data without depending on factors such as a constant sampling period or assumptions about the form of the noise. The choice of SVM is not only related to the robust character of its estimates even when only sparse data is available, but also to the possibility of providing a particular interpretation to the support vectors. This article assumes that, in general, some of the support vectors of a regression represent the most difficult samples to regress and their coordinates in the input domain (note that multivariate signals can be present) determine the exact location of the physical events.

All computations in the article have been performed with Matlab (http://www.mathworks.com). In particular, the SVM implementation for the regression estimation has been ‘The Spider’ software, which is included in public licensed environments for Matlab [6].

This article is divided into 7 sections (including the introduction). Section II briefly reviews the SVM regression method and introduces the required concepts to explain UMEL. Section III is devoted to describing UMEL and particular applications with toy examples are shown. Section IV introduces the strong motivation in nuclear fusion behind the need to use automatic methods to locate singular events (section V) and to detect relevant physical behaviours (section VI). Finally, a short discussion on the technique is given in section VII.

2. SUPPORT VECTOR MACHINE FOR REGRESSION

SVM is a universal constructive learning procedure based on statistical learning theory [5]. SVM was initially developed for the classification problem with separable data. Later it was improved to handle non-separable data and also adapted to solve the regression problem.

Let us consider S training samples $(\mathbf{x}_1, \mathrm{y}_1)....,(\mathbf{x}_\mathrm{s},\mathrm{y}_\mathrm{s})$, $(\mathbf{x}_\mathrm{i} \in \square^n$ and $\mathrm{y}_\mathrm{i} = f(\mathbf{x}_\mathrm{i})$ where $f\colon \square^n \to \square)$. The regression function is given by [7]

$$f^*(\mathbf{x}) = \sum_{k=1}^{S} \gamma^*{}_k H\,(\mathbf{x}_k,\mathbf{x}) \tag{1}$$

The parameters $\gamma^*{}_k$, $k = 1,...,S$, are determined using the solution of the following quadratic optimization problem:

$$\gamma_k^* = \alpha_k^* - \beta_k^*, \quad k = 1,..., S$$

where the parameters α^*_k, β^*_k, $k = 1,...,S$ are determined by maximizing the functional

$$Q\,(\alpha\,,\beta) = -e\sum_{k=1}^{S}(\alpha_k + \beta_k) + \sum_{k=1}^{S} y_k\,(\alpha_k - \beta_k) - \frac{1}{2}\sum_{k,l=1}^{S}(\alpha_k - \beta_k)(\alpha_l - \beta_l)\,H\,(x_k, x_l)$$

subject to the constraints

$$\sum_{k=1}^{S}\alpha_k = \sum_{k=1}^{S}\beta_k,\; 0 \le \alpha_k \le \frac{C}{S},\;\; 0 \le \beta_k \le \frac{C}{S},\;\; k = 1,..., S$$

given the training data $(\mathbf{x}_k, y_k)$, $k = 1,...,S$, an inner product kernel H, an insensitive zone e, and a regularization parameter C.

Different combinations of kernels, e-values and regularization parameter allow determining several degrees of smoothing in the regression estimation. Figures 1a, 1b and 1c show this effect with the same training data. In all cases, C has been set to 1000 and a Radial Basis Function (RBF) kernel has been used. The Kernel parameters are 0.01, 0.2 and 0.2 (from top to bottom) and the insensitive zones are 0.1, 0.1 and 1 respectively.

A. KERNELS FOR ESTIMATING REAL-VALUED FUNCTIONS

To construct different types of support vector machines, different kernels H $(\mathbf{x},\mathbf{x}')$ can be chosen for regression purposes. In particular, kernels which are normally used for classification can also be adopted for regression tasks. Examples of these are kernels generating polynomials

$$H\,(\mathbf{x}, \mathbf{x}') = [(\mathbf{x} \cdot \mathbf{x}') + 1]^d,$$

kernels generating radial basis functions

$$H\,(\mathbf{x}, \mathbf{x}') = H\,(|\mathbf{x}-\mathbf{x}'|)$$

such as

$$\mathrm{H}\,(|\mathbf{x}-\mathbf{x}'|) = \exp\left\{-\,\gamma|\mathbf{x}-\mathbf{x}'|^2\right\}$$

or kernels generating two-layer neural networks

$$H\,(\mathbf{x},\mathbf{x}') = \tanh\left[v\,(\mathbf{x}\cdot\mathbf{x}') + a\right].$$

However, in some cases, in some regression applications the development of special kernels, that reflect special properties of approximating functions, can be necessary [8]: kernels generating expansion on orthogonal polynomials, kernels generating splines and kernels generating Fourier expansions are cases in point.

B. E-INSENSITIVE ZONE

The quality of the approximation produced by a learning machine is measured by the loss $L(y,f(\boldsymbol{x}))$ or discrepancy between the output produced by the system and the learning machine for a given point **x**. By convention, the loss takes on nonnegative values, so that large positive values correspond to a poor approximation. The regression formulation for the SVM uses a special loss function [8]. This loss function is linear with an insensitive zone e (figure 2):

$$\mathrm{L}\left(\mathrm{y}, \mathrm{f}\,(\mathbf{x})\right) = \begin{cases} 0, & \textit{if}\ \left|\mathrm{y} - \mathrm{f}\,(\mathbf{x})\right| \leq e \\ \left|\mathrm{y} - \mathrm{f}\,(\mathbf{x})\right| \mathrm{e}, & \text{otherwise} \end{cases}$$

The loss is equal to 0 if the discrepancy between the predicted and the observed values is less than e.

The insensitive zone *e* provides the required level of accuracy *e* to approximate a function $f\,(x)$ by another function $f^*\,(x)$ such that the function $f\,(x)$ is situated in the e-tube of $f^*\,(x)$. The axis of the tube defines an e-approximation $f^*\,(x)$ of the function $f\,(x)$ (figure 3).

C. REGULARIZATION PARAMETER C

In statistics and machine learning, regularization is used to prevent overfitting. Learning is the process of estimating an unknown (input, output) dependency using a limited number of observations. The challenge in regression is to estimate the values of points different to the training ones. Generalization is the system capability of estimating new points that were not used in the training phase. For a good generalization capability, overfitting should be avoided.

D. SUPPORT VECTORS

Only a subset of the parameters γ^*_k in equation (1) is nonzero. The data points x_k associated with the nonzero γ^*_k are called support vectors. Therefore, the regression function is actually

$$f^*\,(\mathbf{x}) = \sum_{\text{support vector}} \gamma^*_\mathrm{k}\ \mathrm{H}(\mathbf{x}_\mathrm{k}, \mathbf{x}) \tag{2}$$

Figure 4 shows a regression estimation example similar to the one that appears in [7, pp 381]. The training data consists of 40 points obtained according to

$$f\,(\mathrm{x}) = \sin^2\,(2\pi\ \mathrm{x}) + \varepsilon \tag{3}$$

where ε is white Gaussian noise generated with the Matlab awgn() function. The resulting signal-to-noise ratio per sample of $f\,(x)$ is 25dB. The input x has a uniform distribution in [0, 1]. Its values have been generated using MATLAB rand() function. The regressions have been carried out with a RBF kernel ($H\left(x_k,x\right) = \exp\left\{-\frac{\left|x_k - x\right|^2}{2\sigma^2}\right\}$, $\sigma = 0.2$),, a regularization parameter C set to 20000 and e-values of 0.1, 0.2, 0.4 and 0.8 from left to right and from top to bottom respectively. The 'x' symbols in the figure represent the training data, the support vectors are distinguished by square markers, the dot red line is the function $\sin^2(2\pi x)$, the solid black line corresponds to the regression estimation and the dashed lines are the *e*-tube bounds.

By using the same training dataset with identical C and e-value but a polynomial kernel of degree 6, the regression model is

$$f^*\left(\mathbf{x}\right) = \sum_{\text{support vector}} \gamma_k^* \left[(\mathbf{x}_k \cdot \mathbf{x}) + 1\right]^6$$

and the regression estimations are shown in figure 5.

Four empirical facts can be deduced from these simple regressions with two different kernels. Firstly, increasing the e-values produces smoother estimations with less accuracy. Secondly, as *e* increases, less support vectors are obtained but they are subsets of the ones that appear with narrower e-tubes. With the RBF kernel, the number of support vectors diminishes from 13 (32% of the training samples) to 7 (17% of the initial data). In the case of the polynomial kernel, the amount of support vectors goes from 34 (85%) to 12 (30%). Thirdly, the support vectors are samples tending to be the most difficult ones to regress. On the one hand, they correspond to samples on or outside the e-tube boundaries. On the other hand, if they are inside the *e*-tube, the support vectors appear in areas of high curvature or near the limits of the interval. Fourthly, the support vectors are concentrated around the same zones in both regressions and therefore are independent of the kernel function.

These specific conclusions are of general validity in SVM regression problems. Therefore, they suggest that the support vectors are linked to certain morphological characteristics of the signals. In particular, they can be connected with the most difficult samples to regress. This is the central point of the present article because these properties of the support vectors are the basis on which the developed automatic locator of events within signals is founded on.

3. UMEL TECHNIQUE

The aim of UMEL is the automatic location of any type of event inside any class of signal. The formulation of the regression estimations with SVM, equation (1), is valid for any number of dimensions. For this reason, UMEL can be used in feature spaces whose samples can have any number of components.

Equation (1) expresses the fact that the SVM regression estimation of complex datasets can require all samples to obtain a regression model. On the other hand, it should be taken into account that the complexity of a function can be defined in terms of its smoothness, since for smoother functions fewer data points are required for an accurate estimation. This is the meaning of equation

(2). The smoother the function to regress the less support vectors are required.

In a SVM regression estimation, all signal samples that lie on and outward the e-tube are support vectors. This paper refers to them as External Support Vectors (ESV). But also, some samples within the e-tube become support vectors, which are called here Internal Support Vectors (ISV). ISV are necessary samples for the regression estimation, equation (2), but they do not provide the degree of relevance that can be assigned to the ESV (figure 6). Due to the fact that SVM regressions tend to be smooth inside the e-insensitive zone, an original interpretation of the support vectors is proposed: the samples that become ESV are the most difficult samples to regress (they cannot be fitted inside a smooth e-tube) and these samples provide essential additional information in the regression process. The ESV reveal the occurrence of special patterns inside a signal: peaks, high gradients or segments with different morphological structure in relation to the bulk of the signal. Typically, several ESV appear together and they define limited signal segments. On the one hand, singular points show a signature with a strong location in the signal domain and, therefore, they define short segments. On the other hand, large sequences of ESV denote the presence of signal intervals with patterns clearly differentiated from the rest of the signal.

In a SVM regression, three free parameters potentially affect the final model complexity: the regularization parameter C, the kernel parameters and the e-value. The various combinations among them determine the different model selections.

The regularization parameter can be estimated [9, pp 448] as:

$$C = K_C \max \left(\left| \bar{y} + 3\sigma_y \right|, \left| \bar{y} - 3\sigma_y \right| \right) \tag{4}$$

where y is the mean, σ_y is the standard deviation of the training samples and K_C is a constant that can vary between different kind of signals. This estimation can effectively handle outliers in the training data.

In order to establish a criterion in the selection of the e-value, the standard regression formulation should be considered: $y = t(\mathbf{x}) + \xi$. Typically, it is assumed that the standard deviation of additive noise σ_{noise} is known or can be reliably estimated from the data. Then, the *e*-value should reflect the level of additive noise ξ, that is, e−volume $\propto \sigma_{noise}$. In particular and according to [9, pp 449], the selection criterion used in this article is

$$e - \text{value} = K_e \sigma_{noise} \sqrt{\frac{\ln n}{n}} \tag{5}$$

where *n* is the number of training samples, ln *n* is the natural logarithm of the number of training samples and K_e is a proportionality constant. This dependence provides good SVM performance for various dataset sizes, noise levels, and target functions. The value of K_e may vary from case to case just to take into account the aforementioned elements.

Figure 7 shows a toy example in which the function of equation (3) is corrupted in a subinterval with additional Gaussian noise having a signal-to-noise ratio per sample of 20dB. The whole signal is generated with 600 points. The double corrupted segment has a length of 200 samples and its starting point has been determined randomly. In the case of figure 7, the first sample corresponds to an abscissa value of 1.64. The regression model determines the ESV shown in the figure. They just correspond to samples belonging to the signal segment with the worse signal-to-noise ratio. The bottom panel is the histogram of the external support vectors.

Figure 7 is a very clear example of the way that UMEL identifies segments with different morphological structure. In a real case, it would allow the recognition of time intervals in which the signal has properties different from the usual.

As it was pointed out, UMEL is not restricted to identify broad segments of data. It can be used for the exact location of singular points. A new toy example of this potential can be the determination of transition points in pulse trains. A square signal with 3770 samples has been generated with the square() Matlab function and a Gaussian noise per sample of 25dB is added by means of the function awgn(). Figure 8 (top) shows the ESV after the regression estimation process. At the bottom a detailed view of the first transition is shown. It illustrates that samples outside the e-tube are the support vectors which correspond to the closest samples to the transition.

As previously shown, the regression estimation is defined through the selection of the kernel parameter, the e-value and the regularization parameter C. The determination of support vectors around singular points is a robust process in the sense that the choice of the above values is not critical. In figure 8, the number of support vectors around the transitions go from 1 per transition (*e* = 0.8) to 6-8 per transition (*e* = 0.2, that is an average amplitude value of the signal noise). In all cases C = 1000 and RBF kernel $H\left(x_k,x\right) = \exp\left\{-\frac{\left|x_k-x\right|^2}{2\sigma^2}\right\}$ ($\sigma = 0.05$) are used.

If the target function is a sawtooth wave instead of a square signal, the transition times are also determined with a high precision (fig. 9). In this case, the signal-to-noise ratio per sample was set to 20dB.

4. DATABASES IN NUCLEAR FUSION ENVIRONMENTS

The JET Tokamak is the largest nuclear fusion device in the world. Its database stores more than 42 Tbytes of data and, nowadays, more than 11 Gbytes of data per discharge can be collected. This amount of data is provided by diagnostics. Diagnostics are specialised systems devoted to measuring all plasma physical quantities, for instance, temperature, density or energy confinement time. Typically, the diagnostics generate temporal evolution signals (time-series data and video-films) that are stored for off-line analysis.

ITER will be the next generation fusion device and it is expected to acquire about 106 signals (waveforms and video-movies) per discharge, storing Tbytes of data per discharge.

In pulsed fusion devices, like JET or ITER, the plasma is born, lives during a certain period of time and, finally, it is extinguished. The discharge length of ITER will be 1/2 hour and therefore, its signals will record information during 30 minutes. Off-line data analysis should be focused on time intervals where Relevant Physical Behaviours (RPB) take place. A critical point in this respect is to find out the time instants where interesting plasma events occur. Taking into account firstly, the lengths of the ITER discharge and secondly, the large quantity of expected data per discharge, the RPB time instants cannot be found by manual methods. The development of automatic software tools will be essential to achieve two different objectives. The first one is the determination of the exact times when individual physical events appear. Examples can be individual sawteeth or plasma disruptions. The second objective is connected with the creation of specialized software able to suggest the set of temporal intervals in which RPB could be present. The determination of these temporal segments allows discarding data without relevant information and to focus the attention on the time periods most likely to provide important physical knowledge. Of course, the occurrence of false positives is acceptable but they should be minimized.

UMEL is a proper technique to solve the automatic location of both single events and RPB segments. It has been applied successfully to the JET database. The aim of the examples presented in the next two sections is to show the capabilities of the method with real signals. Therefore, the reported results should not be used to draw conclusions about the details of the physics processes described.

Firstly, the location of individual behaviours is accomplished. Secondly, the determination of RPB time intervals is carried out. In all cases, a brief explanation is given about the physical phenomena involved. Then, the results of UMEL are described.

In all the examples of sections V and VI, an RBF kernel ($H\left(\mathbf{x}_k,\mathbf{x}\right) = \exp\left\{-\frac{\left|\mathbf{x}_k-\mathbf{x}\right|^2}{2\sigma^2}\right\}$) has been used for the regression estimations. This kernel has a single parameter s, which defines the width of the region around x for which H is large. A small value of s typically tends to produce rough regression estimations while a large value yields smoother estimations. For the UMEL applications shown here, σ is chosen proportional to the value given by the Normal Reference ruLe (NRL): $\sigma = Kh_{REF}$. The NRL is a criterion to choose the window width to estimate non-parametric probability density functions by means of the Parzen Window estimator [10]. For n training samples, Gaussian functions (like the present kernel) and univariate distributions the rule can be written as $h_{REF} = 1.06sn^{-1/5}$, where a suitable estimate for s is the standard deviation of the distribution [11]. The advantage of using the normal reference rule in UMEL resides in the general validity of the NRL for any kind of signals whose samples can be completely scattered (without a uniform sampling period) through the domain. This contributes to increase the *universal* character of UMEL.

5. LOCATION OF SINGULAR EVENTS

The UMEL technique has been applied in JET to the recognition and location of two different events related to plasma physics: sawteeth and major disruptions.

It should be emphasized that in both cases, the way of computing the regression estimation is the same, as it corresponds to the universal character of the technique. Appendix I contains the Matlab code for this computations with the Spider toolbox.

5.1. SAWTEETH

Sawtoothing is a commonly found instability in nuclear fusion plasmas [12]. The name comes from the sawtooth shape of temporal evolution signals corresponding to central channels of the soft x-ray emission.

The heat pulse due to the sawteeth crash is assumed to spread diffusively to the edge of the plasma. Figure 10 shows the signatures of sawtooth activity in soft x-ray signals. From top to bottom, the waveforms correspond to chords of soft x-ray intensities at different plasma radii, from the plasma core to the plasma edge. Figures 10a and 10b are central chords. The footprint of sawtooth activity in these channels is a periodic increasing emission of soft x-rays followed by an abrupt drop. Figure 10c shows the inversion (q = 1) radius. The sawtooth crashes are revealed here by small spikes. The other signals show how the heat pulse propagates outward. The periodic signature in the signals is an increasing emission after the crash up to a maximum (heat pulse) followed by a slow decreasing. Figure 11a and 11b show an expanded view of figures 10a and 10d respectively.

The application of UMEL to central chords allows an exact estimation of the time instant in which the sawtooth crash happens. Figure 11a shows the ESV of a central channel after the SVM regression. The dot lines are the bounds of the e-tube and its axis (red, bold line) provides the regression estimation. The external support vectors have been represented by circles over the soft x-ray signal. It should be noted that the regression model is smooth but good enough to delimit the time intervals where the sawtooth crashes take place. In these cases, the specific time instant in each sawtooth corresponds to the external support vector with the highest amplitude. Figure 11b shows the regression estimation of a soft x-ray channel beyond the inversion radius but close to it.

The computation time per signal (15000 training samples per waveform) executing Matlab on a computer with one Intel Core 2 Quad processor (2.5GHz, 2GB of RAM) and Windows XP is about 7s. In all signals, the constant K_C of the regularization parameter, equation (4), has been set to 1. The RBF kernel parameter to determine all the maxima in all of the soft x-ray signals has been s = 1000hREF. Regarding the e-tube, several *e*-values have been used for each soft x-ray channel. This is a consequence of the different amplitude levels of the sawtooth signals compared with the background noise (figure 10). The constant K_e in equation (5) to determine the specific e-value for each x-ray signal has varied between 2 and 13.

A further UMEL application with sawteeth can be the automatic estimation of the plasma diffusivity. According to the time-to-peak method [13], it is possible to find an analytic solution for the plasma incremental diffusivity:

$$\chi_{e,Inc} = \frac{r^2 - r^2_{reconnection}}{8t_p}$$

where $r_{reconnection}$ is related to the radius at which the flattening of the electron temperature profile takes place during the sawtooth crash (reconnection radius), r represents the plasma radii of the soft x-ray measurements and, finally, the t_p parameter measures the propagation time of the heat pulse from the inversion radius to the edge (figure 12). The automatic determination of tp allows the automatic determination of ce,Inc because the radial coordinates of the soft x-ray chords are known beforehand. UMEL is used to determine the time instants where the maximum soft x-ray emissions in the vicinity of the crash are reached.

5.2. MAJOR DISRUPTIONS

Disruptions are plasma instabilities in tokamaks that produce sudden losses of confinement [14]. A disruption is a rapidly developing process involving a redistribution of current over the plasma column cross section with a decrease in poloidal field energy and an expulsion of part of the poloidal flux beyond the boundary of the plasma column. The most dangerous ones (major disruptions) provoke a violent and unforeseen termination of the discharge with potential damage to the integrity of the fusion device.

The disruptive instability has traditionally been one of the main concerns in tokamaks. Big efforts have been devoted to understanding the causes of this instability. However, due to the complex non-linear interactions that can originate a disruptive behaviour, methods for avoiding its occurrence or, at least, mitigating its effects continue to be the subject of intensive research.

UMEL can be used in off-line analysis for both the automatic recognition of disruptive behaviours and the location of the disruption time. To this end, two time-series data from the JET database have been chosen: the plasma current (signal PPF/MAGN/IPLA of the JET database) and the loop voltage (PPF/MG2/VSU). The pattern of a major disruption leading to the termination of the plasma is an abrupt change of the current amplitude from its operational value (several MA) to zero in tens of ms. Simultaneously, the loop voltage also shows a characteristic spike. These signatures can be identified by the ESV of UMEL.

The identification of a disruptive behaviour cannot be determined with only one of the previous signals. In a controlled end of discharge, the VSU spike always appears and, therefore, it does not discriminate between disruptive and non-disruptive plasma terminations. With regard to the plasma current, it is usually programmed to evolve in a smooth way. It can be seen as a ‘piecewise linear’ function with a long flat top (figure 13). The time instants in which the plasma current is pre-programmed to start changing level, at 13.88s in the figure 13 and at 16.12s in the same figure, require specific external support vectors to be fitted. Therefore, in some cases, UMEL wrongly identifies the samples around these inflections of the plasma current as due to a disruption. Therefore, disruptions are recognized with a higher success rate by exploiting the simultaneous presence of support vectors in both signals in the same time interval.

Figure 13 shows the plasma current and the loop voltage of a non-disruptive JET discharge together with their regression estimations (axis of the e-tube bounds). UMEL finds ESV in the voltage signal (circle markers) but none appears in the plasma current. Therefore, it is concluded that the discharge is non-disruptive. Other example of non-disruptive behaviour can be seen in figure 14, in which the current evolution is not so flat.

Figure 15 shows the proper identification of a disruptive behaviour that takes place even when the current is decreasing in a controlled way. Figure 16 details the last part of the discharge shown in figure 15. Only one external support vector is identified in the plasma current (time 25.07s) but several ones appear in the VSU signal. The support vector in VSU at time 65.07 is marked in bold. The regression factors for both signals have been $K_C = 1$, $\sigma = h_{REF}$ and $K_e = 30$. Computing times have been 0.33s per signal, with 1024 and 1000 samples for the plasma current and voltage signals respectively. It should be noted the difference with soft x-ray signals. In the present case, the regression model complexity is very low.

In the recognition of the disruptive behaviour in shot 70440 only one external support vector appears. In general, several ESV can be present. To automate the disruption time computation with an absolute error equal to the sampling period of the plasma current, a general criterion consists of choosing the external support vector which shows a maximum difference with the next sample. This support vector is typically the nearest to the current quench.

The validation of the recognition method has been carried out with 4400 JET discharges (343 disruptive and 4057 non-disruptive) between shots 65115 and 70722. The global success rate has been 99.02% (4357 discharges recognized properly). By considering only disruptive shots, the success rate is 93.59% (321 shots well classified). Taking into account just the non-disruptive cases, the success rate in the recognition process is 99.48% (4036 discharges well classified).

6. LOCATION OF RELEVANT PHYSICAL BEHAVIOUR

The capability to recognise of temporal segments with interesting events is illustrated in this section with video-movies of the JET database. An example is provided by the JET KL7 InfraRed (IR) camera. This camera is mounted in a fixed position on an endoscope providing a wide angle view of the JET main chamber. Design requirements established viewing the divertor, the inner wall, the outer poloidal limiters, the ITER-like ICRH antenna and the top limiter, which implies a field view of 70 degrees. The diagnostic was designed to measure from JET operating temperature of 200^{o}C up to a maximum temperature of 2000^{o}C.

Infrared sources are coming from the Plasma Facing Components (PFC) due to plasma power deposition on the first wall. IR images reveal the presence of hot spots, *i.e.* regions of the first wall where the PFC reach temperature high values. These regions usually are well located in the images (divertor, limiters or antennas) and each one is known as a ‘Region Of Interest’ (ROI). The aim of the measurements with IR cameras is twofold. On the one hand, they can be used for real-time control of the discharge when the surface temperature of the PFC approaches dangerous levels. On the other hand, they can help to perform off-line analysis of physical phenomena, for example heat load measurements and plasma-wall interactions.

Off-line analysis can be used to locate the time instants along a discharge where infrared emissions take place. However, infrared cameras can collect hundreds of frames per second. Therefore, just to avoid a manual search by visual inspection of the movies, automatic means to detect high IR emission are needed. UMEL can be used for this purpose.

Each individual frame of a movie can be seen as a two-dimensional function $f\,(x,y)$where x and y are spatial coordinates and the amplitude of f at any point of coordinates (x, y) is the intensity or gray level of the image at that point (or pixel). If no IR emission is present, all pixels are black. A hot spot can be seen as a peaked value of over specific pixels, just the ones that define the ROIs. The peak is revealed by the existence of gradients in and the gradients can be recognized by UMEL through the presence of external support vectors.

It should be noted that depending of the plasma evolution, several hot spots can be present simultaneously and, therefore, more than one peak per frame can be expected. UMEL does not need to recognize the number of individual peaks, their maximum amplitudes or even the size of their respective ROIs. Just a single regression estimation of each frame is enough to show the presence of hot spots (regardless of the number of them). The evidence comes from the number of ESV that are determined in the regression process. By following the number of ESV that appear frame by frame in a movie, it is possible to identify the time instants that show a greater activity in IR emissions.

Figure 17 (top) shows the temporal evolution of the number of external support vectors in the discharge 70231 of JET after applying the UMEL technique to a series of frames (abscissa axis). The plot was generated from a movie with 300 frames. Each frame (515x505 pixels) is regressed with UMEL and the computing time (with the same computer mentioned in the section devoted to sawteeth) has been 120 s/frame. The regression parameters have been K_C = 1, s = 5h$_{\mathrm{REF}}$ and K_e = 100. At this point, part of the knowledge enclosed in the movie has been condensed in a single waveform. A further step is needed to know the exact location of the peaks. To this end, UMEL is again applied to the last waveform. The corresponding external support vectors allow the temporal location of the maximum amplitudes (figure 17 bottom). In this way, it has beenpossible to automatically identify temporal segments with higher infrared activity along the discharge. Once identified the time instants, specialized analyses can be focused around these times.

Finally, it should be remarked that the regression estimations of images have been performed exactly with the same software that was used for time series data (see Appendix I).

DISCUSSION

The physical properties of systems that evolve under stationary conditions scarcely vary. Therefore, the signal amplitudes collected by their measurement systems practically remain constant or, at most, can show smooth trends. The existence of events that alter the quiet system evolution (for example internal instabilities or external perturbations) is revealed by the presence of specific patterns that are fully located in time. For non-TES measurements, also the existence of particular signatures denotes the existence of well-located patterns inside the signals.

UMEL allows the automatic recognition and location of signal segments with relevant data and singular events. This is accomplished through a specific interpretation about the meaning of the support vectors in regression estimations based on SVM.

UMEL is a robust technique in the sense that the selection of the regularization parameter, the *e*-value and the kernel parameter are not critical. Several regression models with different complexity levels provide the same results (external support vectors appear around the same points).

UMEL uses exactly the same software independently of the signal type and the signal dimensionality. This is a key feature of the technique because it completely avoids the need of adapting different software to the characteristics of the different signals: noise levels, amplitude ranges, structural forms and dimensions.

However, one important point to emphasize is the long computation times to process complex signals. This time is especially high to analyze complete video-movies. However, it should be noted that the method is easily parallelizable and, therefore, the use of high performance computing can help in reducing computational time.

ACKNOWLEDGEMENTS

This work was partially funded by the Spanish Ministry of Science and Innovation under the Project No. ENE2008-02894/FTN. This work, supported by the European Communities under the contract of Association between EURATOM/CIEMAT, 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]. J. Vega. “Intelligent methods for data retrieval in fusion databases”. Fusion Engineering and Design. **83** (2008) 382-386.

[2]. J. Vega, A. Murari, A. Pereira, A. Portas, G. A. Rattá, R. Castro and JET-EFDA Contributors. “Overview of intelligent data retrieval methods for waveforms and images in massive fusion databases”. Fusion Engineering and Design (in press).

[3]. J. Paméla et al.”“The JET programme in support of ITER”. Fusion Engineering and Design. **82** (2007) 590-602.

[4]. C. Alejaldre et al. “First plasmas in the TJ-II flexible Heliac “. Plasma Phys. Controlled Fusion **41**, 1 (1999), A539-A548.

[5]. V. N. Vapnik. “The Nature of Statistical Learning Theory”. Second edition. Springer. (1999).

[6]. “The Spider - A machine learning in Matlab”. http://www.kyb.tuebingen.mpg.de/bs/people/spider/main.html. Max-Planck Institute for biological Cybernetics, Tuebingen, Germany.

[7]. V. Cherkassky, F. Mulier. “Learning from data“. John Wiley & Sons, Inc. (1998), **380**.

[8]. V.N. Vapnik. “Statistical Learning Theory”. John Wiley & Sons, Inc. (1998), **460-471**.

[9]. V. Cherkassky, F. Mulier.”“Learning from data“. Second edition. John Wiley & Sons, Inc. (2007), 449.

[10]. R.O. Duda, P. E. Hart, D.G. Stork.“”Pattern classification”. Second edition. Wiley-Interscience. (2001).

[11]. W. L. Martinez, A. R. Martinez. “Computational Statistics Handbook with MATLAB”. Chapman & Hall/CRC. (2002).

[12]. G. Bateman. “MHD Instabilities”. The MIT Press. (1978).

[13]. M. Soler, J. D. Callen. Nuclear Fusion **19** (1979) 703.

[14]. J. Wesson. “Tokamaks”. Oxford University Press. (1987).

APPENDIX I

```
% Spider svr object
a = svr;

% Input: kernel type (linear, poly, RBF, …)
a.child = kernel(‘kernelType’);

% Input: kernel parameter (depends on kernelType)
a.kerparam = kernelParam;

% Input: e-tube value
a.epsilon = eValue;

% Input: regularization parameter
a.C = CParameter;

% The training data are in variables x and y. The former is a matrix of size rows x cols, where rows is the number of training samples and cols is the number of sample dimensions. The y variable carries the function values at points x
ddat = data(x, y); % Conversion to Spider object

% THE ONLY SENTENCE TO PERFORM REGRESSION ESTIMATION IS
[estimy algo] = train(a, ddat); % valid for any dimensions

% Output: support vectors
xsv = get_x(algo.Xsv);

% Output: fit of the training data, i.e. the axis of the e-tube
sfit = get_x(estimy);

% Output: top bound of the e-tube
sfitPlusE = sfit + a.epsilon;

% Output: bottom bound of the e-tube
sfitMinusE = sfit - a.epsilon;
```

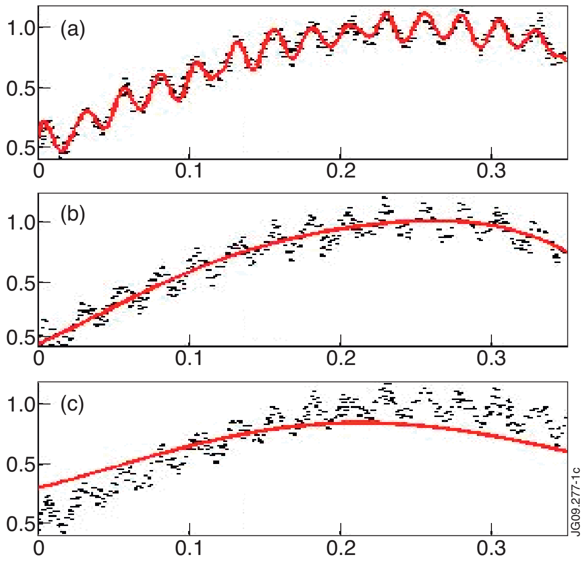

Figure 1: Different combinations of C, e-value, and kernel parameter provide several degrees of smoothing. The training data are represented by dots and the red continuous lines are the regression estimations.

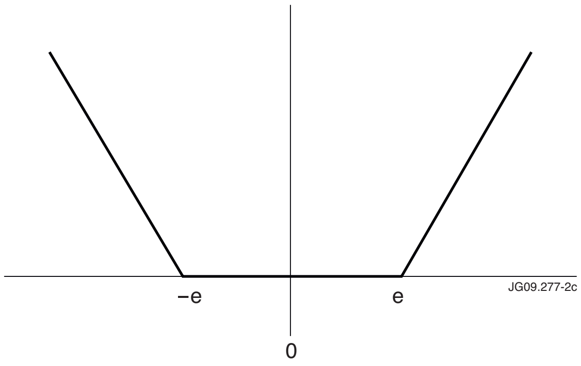

Figure 2: e-insensitive linear loss function for SVM regression estimations.

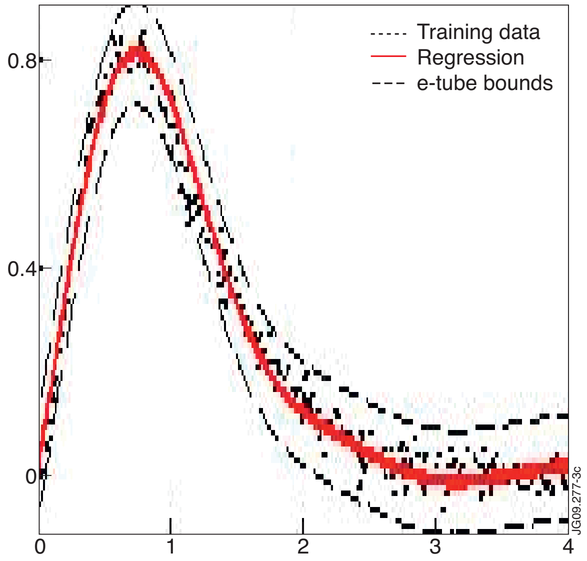

Figure 3: The solid line is the axis of the tube and defines the regression estimation.

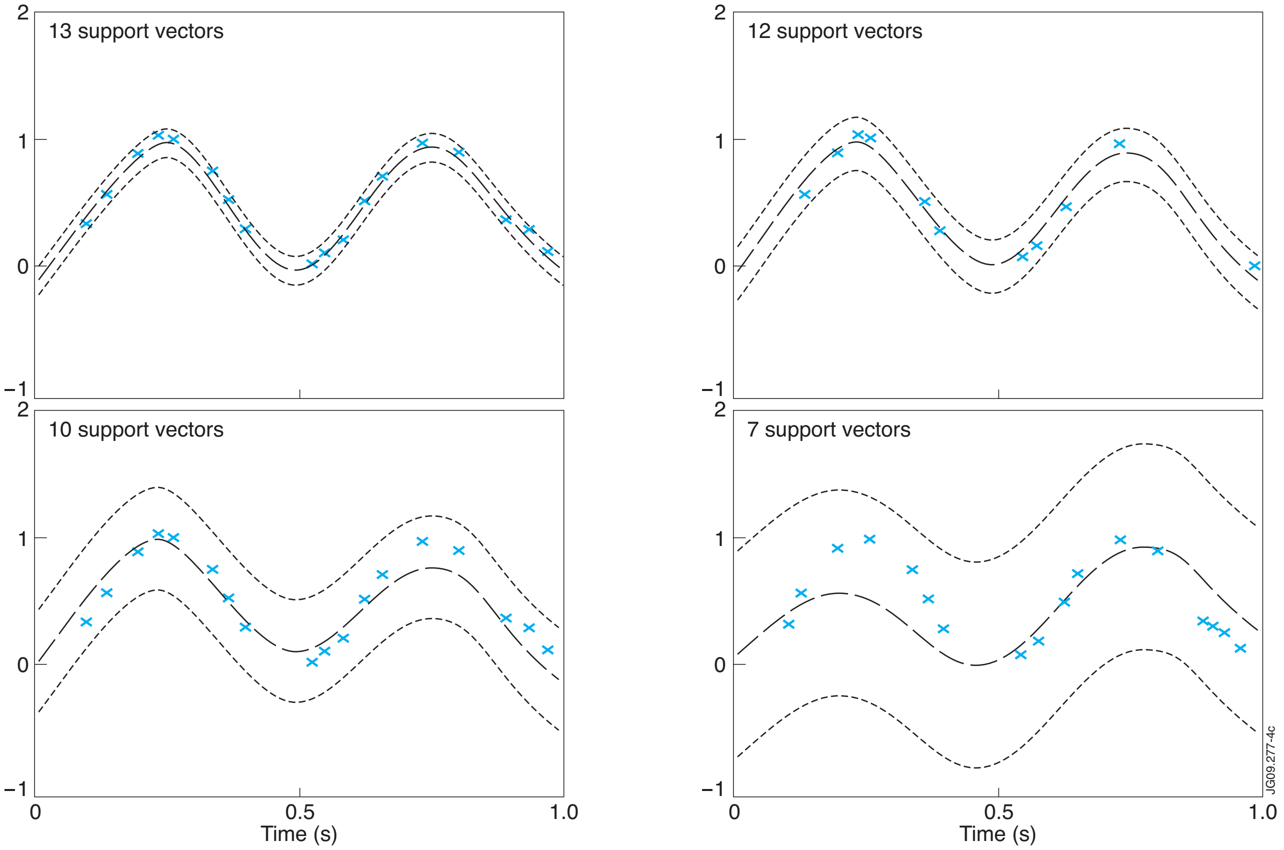

Figure 4: Regression estimations of $\sin^2(2\pi x) + \varepsilon$ with a RBF kernel and the number of support vectors for different e-values.

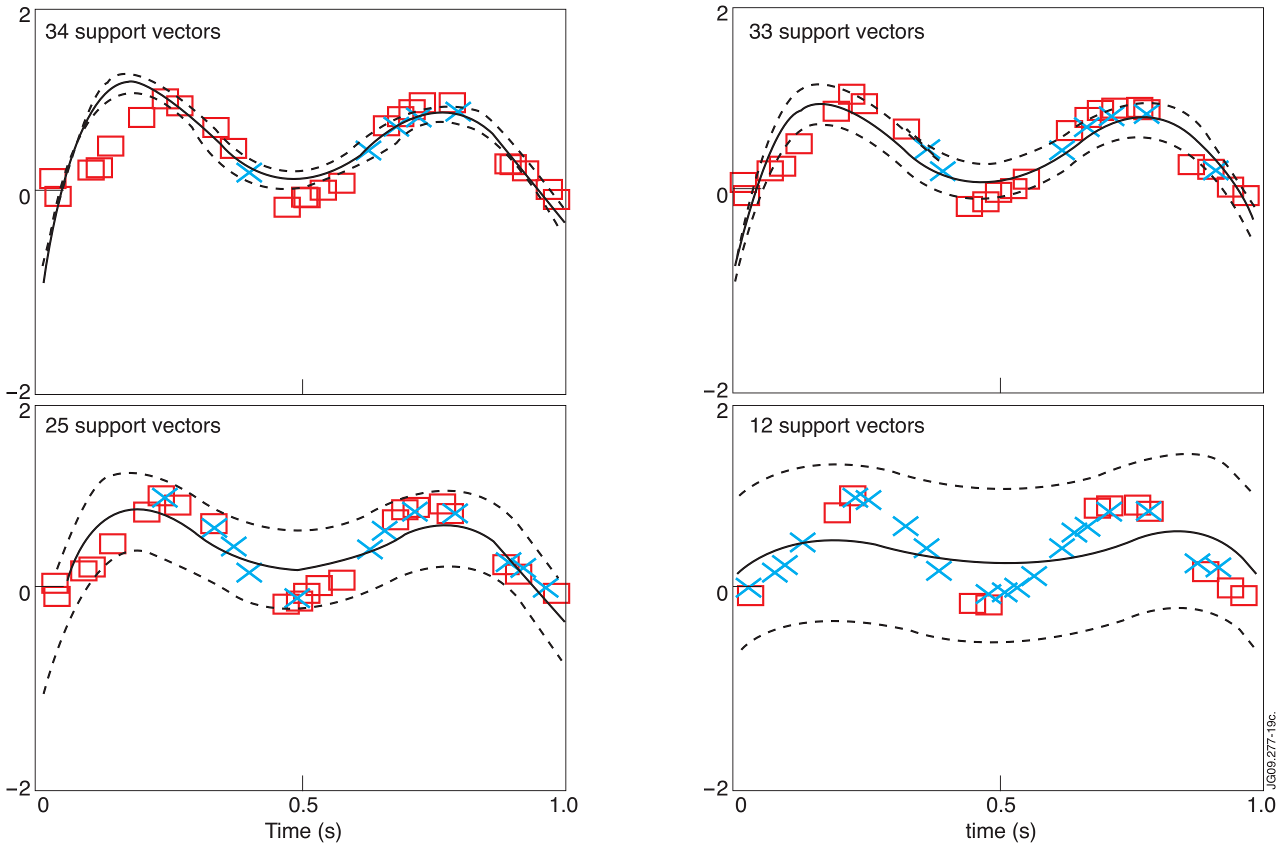

Figure 5: Regression estimation of $\sin^2(2\pi x) + \varepsilon$ with a polynomial kernel of degree 6.

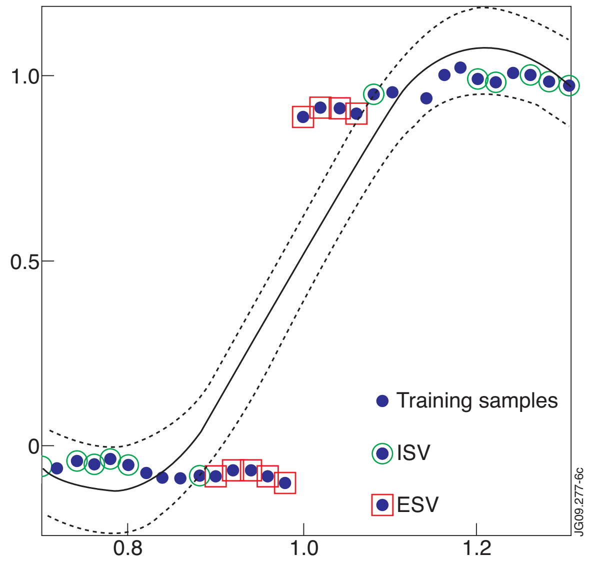

Figure 6: Heaviside function corrupted with pink noise and regressed with an RBF kernel. The dashed lines are the e-tube bounds and the plain line is the regression estimation: $f^*(x) = \sum_{\text{support vectors}} \gamma_k^* H(x_k, x)$

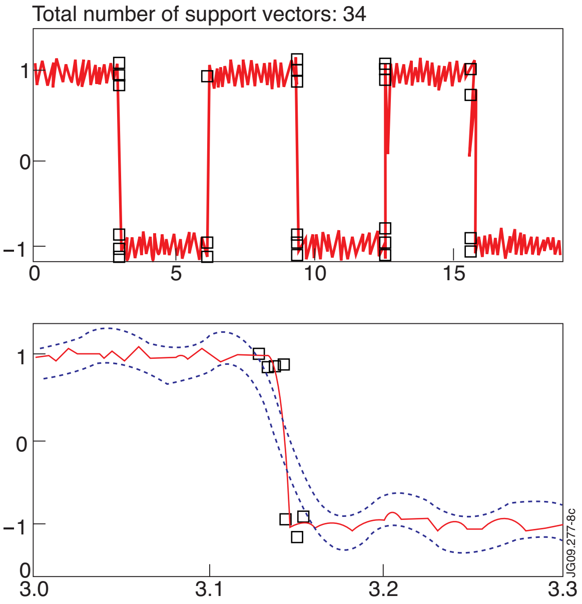

Figure 8: Location of singular points in a square signal. The samples outside the e-tube (black, square markers) become the external support vectors and define the transition very accurately.

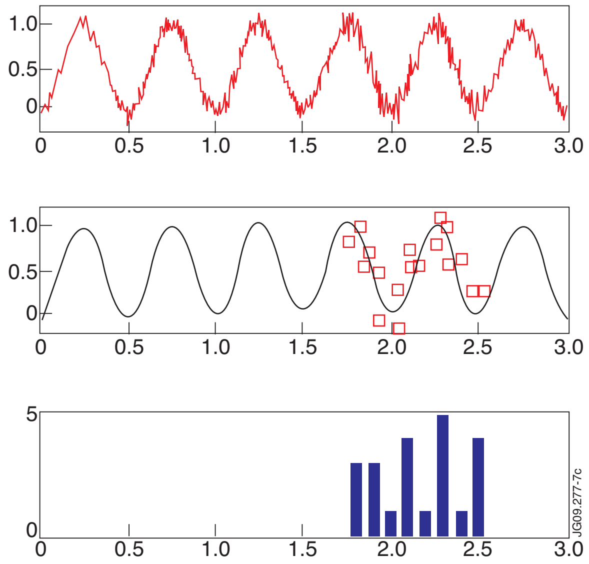

Figure 7: Top plot shows 600 samples generated with equation 3. Two hundred points (starting at an abscissa value of 1.64) are corrupted with a worse signal-to-noise ratio. In the middle, the regression estimation (plain line) and the external support vectors (square symbols) are represented. At the bottom, the histogram of the external support vectors appears.

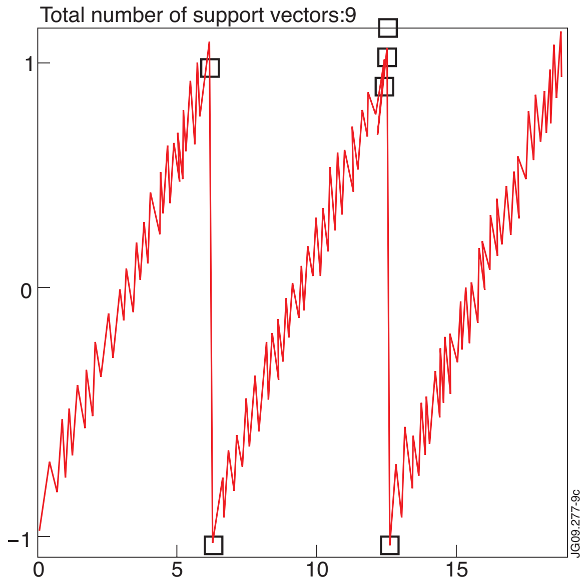

Figure 9: The ESV are the square markers. The total number of samples is 3770.

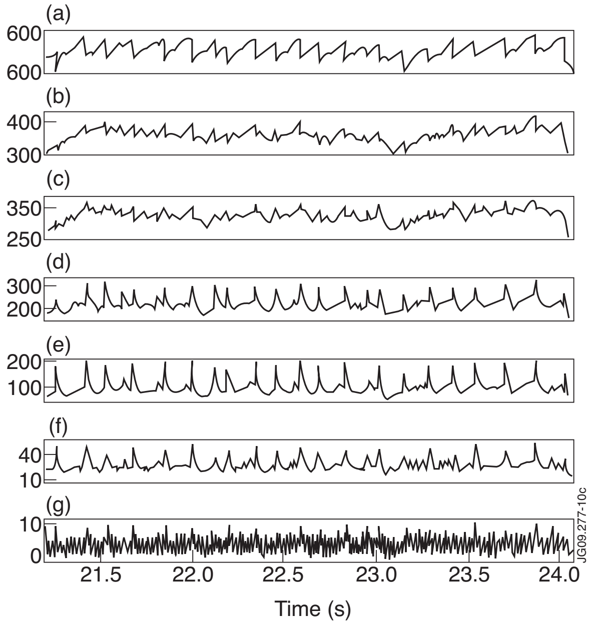

Figure 10: Sawtooth activity in JET Pulse No: 60908.

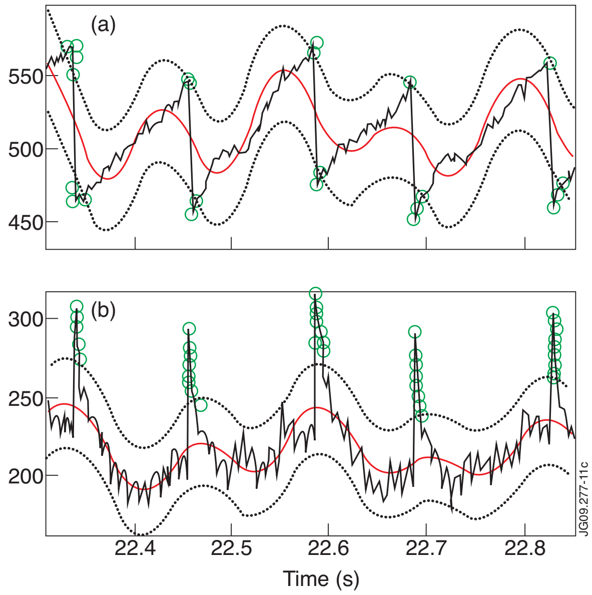

Figure 11: Determination with UMEL of time instants showing the maximum of the soft x-ray emissions. Data correspond to the same discharge of figure 10.

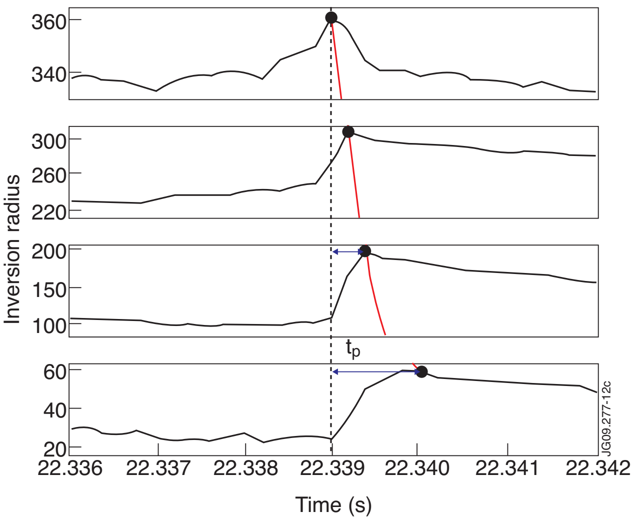

Figure 12: Automatic determination of the incremental plasma diffusivity according to the time-to-peak method with UMEL. From top to bottom the signals correspond respectively to the soft x-ray signals (c), (d), (e) and (f) of figure 10.

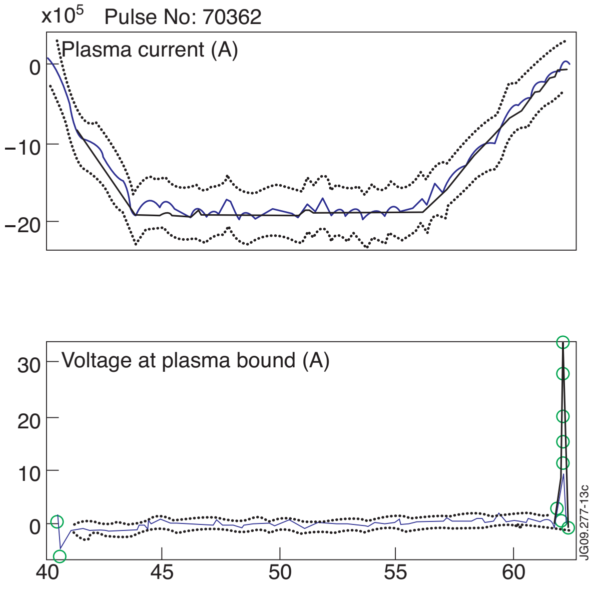

Figure 13: Non-disruptive discharge. UMEL identifies ESV only in the SVU signal.

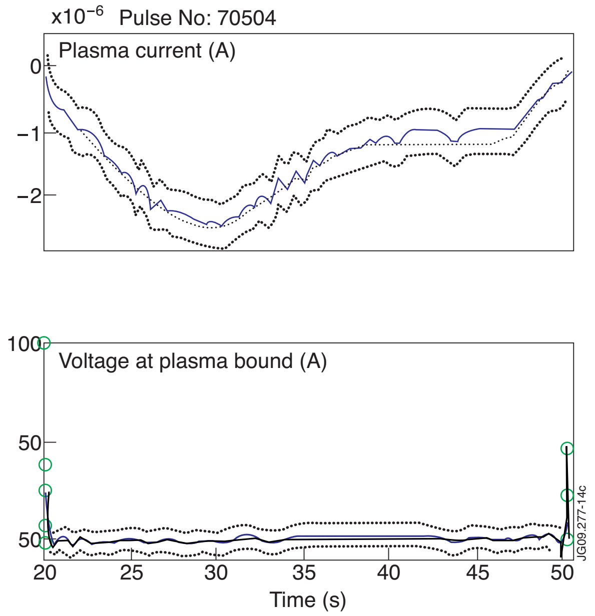

Figure 14: Non-disruptive discharge. The recognition of a disruptive behaviour requires the simultaneous presence of external support vectors in both signals around the same time interval.

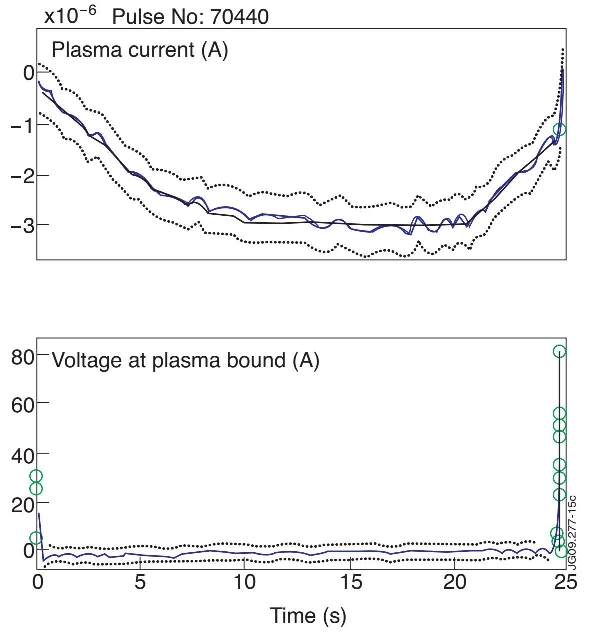

Figure 15: Disruptive behaviour. Simultaneous ESV appear in both signals around a common time instant.

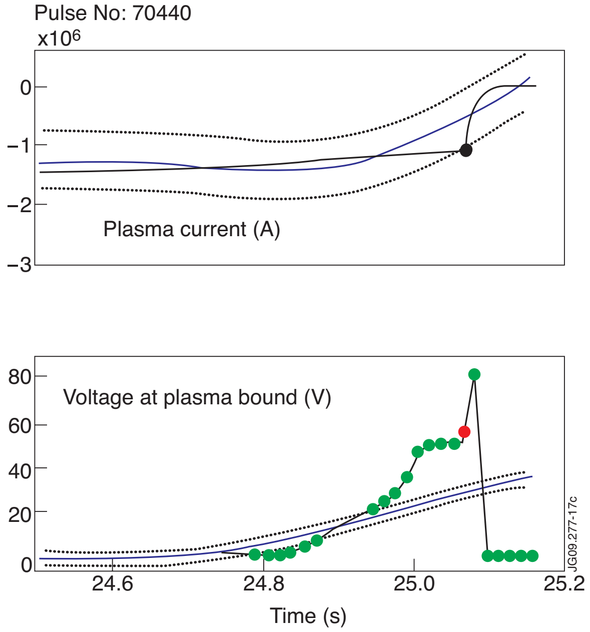

Figure 16: The e-tubes and the regression estimations are shown for both signals. External support vectors are represented with circles and the one identifying the disruption instant appears in bold.

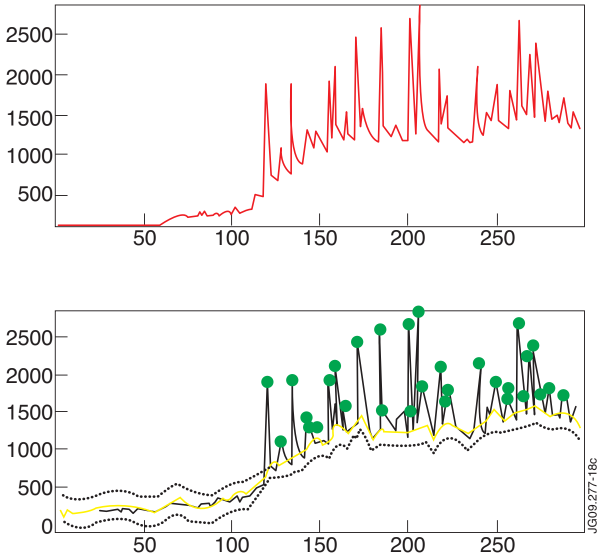

Figure 17: Top: ESV temporal evolution for Jet Pulse No: 70231. Bottom: automatic identification of frames with higher IR activity. The regression parameters have been $K_C = 1$, RBF kernel with $\sigma = 0.5h_{REF}$ and $K_e = 20$. The computing time was 3.38s.