EFDA–JET–CP(07)04/04

A. Murari, J. Vega, J.A. Alonso, E. De La Luna, J. Farthing, C. Hidalgo,
G.A. Rattá, J. Svensson, G. Vagliasindi and JET EFDA contributors

How to Extract Information and Knowledge from Fusion Massive Databases

How to Extract Information and Knowledge from Fusion Massive Databases

A. Murari [1], J. Vega[2], J.A. Alonso[2], E. De La Luna[2], J. Farthing[3], C. Hidalgo[2], G.A. Rattá[2], J. Svensson[4], G. Vagliasindi[5] 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]EURATOM/UKAEA Fusion Association, Culham Science Centre, Abingdon, UK
[4]Max-Planck-Institut für Plasmaphysik, Teilinstitut Greifswald, EURATOM Association, Wendelsteinstr.1, 17491 Greifswald, Germany
[5]Dipartimento di Ingegneria Elettrica Elettronica e dei Sistemi-Università degli Studi di Catania, 95125 Catania, Italy
** See annex of M.L. Watkins et al, "Overview of JET Results ", (Proc. 21st IAEA Fusion Energy Conference, Chengdu, China (2006)).*

Preprint of Paper to be submitted for publication in Proceedings of the
Inernational Workshop on Burning Plasma Diagnostics, Villa Monastero,Varenna, Italy.
(24th September 2007 - 28th September 2007)

ABSTRACT.
The need to understand and control the dynamics of reactor grade fusion plasmas requires the analysis of increasing amounts of data, which at JET can reach easily the level of several GBytes per shot. Therefore a series of new approaches are being pursued to store the data and to retrieve the required information. They range from loss less data compression techniques, to wavelets and Structural Pattern Recognition methods. Since the information available is very often affected by high level of uncertainties and the phenomena to be studied are complex and nonlinear, the inference problems in this field of plasma physics are particularly delicate. Even in this perspective innovative solutions are under development. In particular a range of Soft Computing approaches have already been implemented at JET. The most successful are Bayesian statistics for the integration of diagnostic measurements, Data Mining techniques to study the nonlinear correlation of various variables and Fuzzy Logic to include the knowledge of the experts even if formulated in linguistics terms.

1. INTRODUCTION

Fusion plasmas are nonlinear, complex systems, which have to be controlled to a high degree of reliability in order to achieve the final goal of thermonuclear fusion for energy generation. They are also quite delicate physical entities, very difficult to access and therefore, in many cases, the required information must be derived indirectly on the basis of the natural emission of the plasma, like electromagnetic radiation or particles. The objectives of data processing are also quite different compared to other traditional fields of physical research, like particle physics. In high energy physics, very often the main task of the analysis consists of isolating the products of a specific reaction, among a myriad of particles generated by competing and unwanted phenomena. This is not the case in Magnetic Confinement Nuclear Fusion (MCNF), where, at least in theory, one should be able to make sense of all the data collected in each single discharge. In this respect, the similarities are higher with the health sciences, in which a holistic view of the patient is required to optimize the result of therapies or operations. Such a comprehensive view of fusion plasmas will be also desirable in the final reactor, to find the best compromise between the often contrasting objectives of maximizing the efficiency of the machine and minimizing the risk of incidents. On the other hand, this global understanding of all the available data is not an easy task in modern machines, since the amount of information, derived from each plasma shot, keeps increasing. At JET, for example, already more that 10 GBytes of data have already been acquired in some specific discharges and in normal experiments several Gbytes of data are the typical output. Since the first discharge at JET, the increase in the amount of collected data has followed the Moore law (doubling every two years) and this trend is expected to continue in the future.

The management of such an amount of information presents of course a series of issues. The most evident are related to the data transmission and storage. In particular various forms of first data processing and data compression are under study to optimize the data management over JET network, with particular attention to the implications for the real time control of complex phenomena in context of scenario development for ITER (see section 2). In the last years, the problem of information retrieval has also become a very sensitive issue. In addition to producing more data, JET has also become a much more international experiments, with hundred of scientists visiting from all the countries of the EU and also from outside Europe. New tools, based mainly on Structural Pattern Recognition techniques, have therefore been developed to help the scientist in the first data screening (see section 3).

Even if progress in magnetic confinement fusion has been significant in the last decades, several issues remain obscure and comprehension of many aspects far from satisfactory. The correlation between many variables, together with various sources of uncertainties, increases the difficulty of many problems. Since many experiments have already been carried out, more data driven theory could be a solution to improve the rate of progress in the next years. The first steps in this direction have already been moved with the adoption of various data mining techniques to explore in particular the non linear correlation between the many variables which can cause disruptions (see section 4).

In addition to the issue of information retrieval, fusion plasmas of reactor relevance pose also a formidable inference problem. Since they are very delicate physical objects, they cannot be perturbed too much in order to perform the necessary measurements. Therefore information about their internal status must be derived indirectly, sometimes with a significant level of uncertainties. These difficulties, combined with the fact that very often non linear complex phenomena dominate their behaviour, have motivated the community to develop more robust statistical inference methods to both validate and combine the data of various diagnostics and to falsify theories. The most successful approach tested so far is based on Bayesian theory, which has been applied, among other things, to derive the magnetic topology of JET discharges (see section 5). Other Soft Computing methods, other than the statistical ones already mentioned, are also receiving increased attention. From Fuzzy Logic to Neural computing, the successful applications increase in number and ambition (see section 6).

2. DATA STORAGE AND FIRST SIGNAL PROCESSING

As already mentioned, JET has demonstrated that its diagnostic can produce more than 10 Gbytes of data per shot and this amount of information is expected to reach more than 50 Gbytes with the installation of the new measuring systems. Since during a typical two shifts operational day JET can have between 30 and 40 good shots, the quantity of information to be handled is clearly already enormous and it is going to increase. In ITER the data to be acquired will certainly be 1-10 Tbytes per shot. On the other hand, as in any typical experiment, it is difficult to tell which signals will be the most relevant for the understanding of the physics and the progress of the scientific programme. Any loss of information is therefore to be avoided because tit could affect the interpretation of the experiments. Therefore in order to manage the vast amount of data produced only loss less compression techniques are an option in a fusion device. In JET and other facilities, like TJ-II, the so called delta methods have been successfully adopted. They are based on the observation that magnetic confinement fusion plasmas are relatively stable physical entities, whose parameters tend to change quite slowly in time. Therefore, the difference between two subsequent time slices of the same signal has a probability distribution which peaks at zero and decreases rapidly as the value of the difference increases. The strategy therefore consists of storing only the difference between subsequent time slices, assigning the shortest symbol to the value of zero and progressively the longer symbols to higher incremental values, which have a much lower probability of occurring. The detailed description of the solutions tested on TJ-II and JET are reported in [1]. With this technique, the average compression rate on TJ-II waveforms (almost 3 million signals) is 69.87%, where the compression factor is defined as

$$f = \left(1 - \frac{\textit{compressed storage}}{\textit{no compressed storage}}\right) *100$$

i.e. a 70% of compression means that only 30 bytes out of 100 bytes are written. Application to JET waveforms (62567 time-series data) gives f = 49.59%. It is worth pointing out that this approach can be applied very successfully also to bidimensional signals and its effectiveness has already been verified. Table I shows a comparison of this technique with well-known compaction procedures. All of them were applied to typical video frames of JET infrared and visible cameras. In addition to the higher compression rate, it should be mentioned that the above delta method can be used under real-time requirements whereas the other procedures are delayed techniques (compression only can be carried out having available all the data). The real-time software compression on frames of 256×256 pixels takes 5.5ms (< 200 frames/s) in a Pentium IV computer with Windows XP. The use of dedicated hardware, for instance Field Programmable Gate Arrays (FPGA), will allow a significant reduction of this time.

3. INFORMATION RETRIEVAL

Even once the data has been efficiently and safely stored, the problem of retrieving the desired information remains a huge task. In JET, this has become particularly evident under the new diffuse organization, with scientists from all over Europe participating in the experimental campaigns and the following analysis of the measurements. Only the preliminary task of identifying the most relevant discharges, to concentrate on them the subsequent analysis efforts, has become an incredibly time consuming and sometimes frustrating work. In order to alleviate this problem and render the first data screening a much more efficient process, completely new way to organise the data bases is under intensive investigation. In the new paradigm adopted, signals are not organized according to the traditional scheme of signal name, shot number and time slice. On the contrary, the objective consists of giving the scientists the opportunity to select, on a simple visual interface, the part of a signal they are interested in and then it is the task of the software system to extract the shot numbers and time intervals in which the same signal shape is present. This new approach, which is meant to complement not to substitute the old database structure, is illustrated in fig.1.

The methods to obtain these objectives are based on pattern recognition, which is a mathematical formulation of the more general problem of object identification. The general task of recognising objects is a characteristic goal of all living beings and pervades their daily life, since it is essential in basic activities like search for food or navigation, identification of friends and enemies etc. Pattern recognition is the discipline which studies the mathematical aspects of these activities, so that they can be carried out automatically by computers. In fusion the most promising techniques for the first data screening are the ones of structural pattern recognition. They are based on the general structural shape of the signals more than on their detailed mathematical properties (unlike correlation or coherence estimators). This choice is based on the fact that in plasma physics many phenomena can indeed be identified by the form of the signals they generate in the diagnostic measurements. And these shapes in the signals are indeed what the physicists normally use to identify the various phenomena of interest during the visual inspection of the measurements, a preliminary activity to all type of subsequent data analysis. The main idea therefore consists of developing methodologies that can perform a first screening of the data extracting the signal shapes which are similar to a prototype defined by the user. The first step in achieving this goal is the extraction of the relevant features from the signals. On the basis of these features, criteria of similarity have to be defined, together with a suitable distance, which would allow to assess automatically the degree of similarity between different signals.

To exemplify these concepts the approach based on primitives is particularly instructive, in addition to being very effective in finding patterns within time traces. The version, already implemented in JET, is based on the division of time traces in a series of short intervals, inside which the gradient of the signal is calculated. The gradient levels are codified in a suitable number of groups, depending on their value, and a letter is attributed to each group. The signals can therefore be represented as a sequence of letters, codifying the amplitude of the signals gradient in each time interval (see fig.2). Finding a certain signal pattern is therefore reduced to a string comparison problem, which can be carried out very efficiently. The method provides also a lot of flexibility because differences from the original letter sequence can be allowed, relaxing the constraints on the search degree of similarity. Also filters can be applied to impose the search for signal with the same time base and amplitude or more generally simply signals parts with the same shape can be looked for (independently from amplitude and time duration). Alternative approaches have also been tested. For example, to retrieve entire time traces and not only parts of them, an efficient method is base on the wavelet transforms, which play the role of sophisticated filters [2]. Then the search is performed in the space of the transforms on the appropriate parameters. Some pilot software, implementing these solutions for monodimensional signals, is already operational at JET. Figure 3 illustrates the typical performance for the case of a generic time trace. This methodology has also been recently extended to images. In this case, again a wavelet transform is applied to the image and then the original shape is retrieved.

4. DATA MINING

As described in the previous sections, techniques for storing big amounts of data and for retrieving information efficiently from massive databases have been introduced. These methods have already been implemented at JET and other devices and they have proven to be very useful. On the other hand, plasmas of fusion relevance are very complex and nonlinear systems and therefore very important information can remain hidden in the signals, without the scientists being aware of it. Another important issue could be the tendency to oversimplify physical issues, which is always a danger when very complex phenomena have to be studied and solutions are dispersed in large amounts of information. The methods used to address these problems fall in the category of “data mining”, which is defined as the collection of techniques used to extract hidden information from large amounts of data. Data mining techniques have been originally developed by private companies for various purposes, ranging from product development to market targeting. These approaches allow, for example, automated prediction of trends and behaviour and the automated discovery of unknown patterns (whereas the pattern recognition methods described in the previous section of course assume that the signal shape to be retrieved has already been identified).

In the context of MCNF, a particularly reach application field of data mining techniques is expected to be the quest for nonlinear correlations among signals, which is also a scientific issue of wider application than simply plasma physics. The example described in the following uses classification and regression trees to determine the relevance of various quantities for the prediction of disruptions. The instabilities, which can evolve to the point of causing disruptions, have very complex dynamics and they can derive their energy from many different sources. It has therefore resulted impossible so far to reach a satisfactory theoretical understanding of the disruption phenomenology. As a consequence, no satisfactory model is available to predict the occurrence of disruptions and in the last years significant efforts have been devoted to devising non algorithmic predictors. In order to better understand the interplay between the various causes and to formulate better predictors, an analysis of the importance of various signals for the prediction of disruptions has been undertaken using Classification and Regression Trees (CART) methods. CART is a non-parametric and fully nonlinear technique to solve classification ad regression problems. In our case it is used to describe on output variable, the fact that a discharge is disruptive or not, as a function of different interpretative variables. Given the value of the output variable and the values of the signals which are believed to be of relevance for the interpretation of the phenomenon, the method builds automatically a classification tree. This is achieved by traversing the whole database of the input variables and trying to find the signal which better divide the discharges into classes, one with only disruptive and one with only non-disruptive discharges. This approach is said to try to maximize the purity of the child nodes, in the sense to split the classes in the most coherent way. In general, of course, no variable is enough to obtain two pure child nodes so the process is repeated or the child nodes using other variables until a pure classification is obtained (or the list of variables is exhausted). The hierarchy of the nodes in the final tree allows defining the relative importance of the various variables for the classification of the phenomenon under study, the disruptions in our case.

Table II summarises the results of an application of the CART method to a JET database of disruptions. The various columns refer to different time intervals before the occurrence of the disruption. It appears very clearly how the relative importance of the various signals changes significantly depending on the considered time. This information has been included in the design of a Fuzzy Logic predictor to improve its performance as described in section 6. This technique and other classification methods, based on automatic clustering and unsupervised learning, are being investigated further and this field of research is expected to gain significant momentum in the next years.

5. BAYESIAN STATISTICS

For the reasons described in the introduction, large nuclear fusion experiments present a remarkable interpretation problem and require complex processes of inference, which sometimes have to rely on sophisticated physics models and complex inversion algorithms, to provide a good estimate of the physical quantities of interest from the available external measurements. Due to these difficulties, the extraction of physical information from measurements and its validation is undergoing a fundamental methodological revision and more systematic, statistical techniques are being implemented. The general approach of 'Bayesian Probability Theory' is proving very suited to the difficulties inherent in associating error bars to the diagnostic signals and in deriving better estimates of the physical parameters needed to analyse the experiments. Based on Bayes theorem, it constitutes now a solid body of work that can be applied to a series of different problems, ranging from measurement interpretation to theory falsification.

In JET Bayesian statistic is being considered now to treat in a coherent way the uncertainties in the measurements, both random and systematic. This is expected to provide a better evaluation of errors bars and the determination of various parameters, for which independent measurements exist, with higher accuracy. More ambitiously, it constitutes the backbone of the diagnostic integration programme, which aims at deriving the plasma state from a unique Bayesian estimator, capable of including a vast majority and possibly all the diagnostic information available. This methodology is based on the observation that all the element of the inference process, from the model parameters to the measurements and prior information, should be represented as probability distribution functions. Bayesian statistic is an already developed tool to combine all these probability distribution functions, the inputs, to obtain the probability distribution functions of the required physical quantities, the output of the model [3].

On the route of this ambitious programme, the first step is the derivation of the magnetic topology of the plasma. In addition to providing the basic element of confinement in a Tokamak, the magnetic fields represent also an essential ingredient to interpret the measurements of practically all the diagnostics. The topology of the magnetic fields indeed are the most natural choice for the reference frame into which combine all the available measurements. The main problem is that in the past the magnetic filed geometry was determined with the help of complex equilibrium codes, which rely on various hypotheses and do not allow an easy determination of the error bars to be associated with the final results. It was therefore decided in JET to develop a current tomography i.e. a Bayesian derivation of the magnetic fields of the plasma on the basis of the pure measurements and without any hypothesis on the equilibrium. To this end, JET plasmas and the surrounding conducting structures have been modelled with a series of current beams. The values of the currents in these beams are optimised on the basis of the prior information in order to reproduce best the available measurements of the pick-up coils in a statistical sense. The results are very positive and the error bars to be associated to essential parameters of the magnetic topology, like the magnetic axis or X-point position, can be easily derived as shown in fig.4 [4]. The next step in this research programme will be the systematic inclusion in the inference process, leading to the magnetic topology, of the internal magnetic measurements, like the Motional Stark Effect and Polarimetry.

6. FUZZY LOGIC

In addition to the complexity of the plasmas and the uncertainties in the systems and the measurements, and sometimes because of these facts, some specialist knowledge of difficult phenomena is often patrimony of experts, who do no have the background or the opportunity to formulate it in mathematical terms. Sometimes therefore there is a relevant amount of information, which is difficult to express in a way useful for the wide community because it remains at linguistic or in any case at an informal level. A case in point, to use an example already mentioned, is the one of disruptions. These sudden events, which produce a complete loss of confinement and can be very dangerous for the integrity of the machines, are induced by many different causes, which can interact nonlinearly in complicated way. Moreover, for a series of reasons, the boundary between the safety and dangerous regions of the operational space are very often blurred and difficult to pinpoint exactly. All these aspects have suggested the investigation of “Fuzzy Logic” as a possible tool for both understanding and prediction of disruptions. This recent form of logic is indeed based not on crisp sets, like Boolean logic, but on fuzzy sets, whose membership is defined in completely different way from the traditional one, which hinges on the so-called “Law of the Excluded Middle”, which states that an element X must either be in set A or in set not-A. On the contrary, in the scheme of Fuzzy Logic the membership of an element to a set can take any value in the whole interval [0,1] and not only the Boolean discrete values 0 and 1. In more detail, a fuzzy set F defined on U (called the universe of discourse) is given by $F = \{(u, \mu_F (u))| \mu \in U\}$where $\mu_F (u)$ is the membership function, a curve that specifies how each element of U is mapped to the real interval [0, 1], that is $\mu_F (u)$: U → [0, 1]. In other terms, for each $u \lfloor U$, the membership function defines the degree of membership of u to the fuzzy set F, which can vary continuously from zero (no membership) to one (full membership) according to the particular properties of the fuzzy set. It can be demonstrated that fuzzy logic is in reality a superset of traditional Boolean logic. Therefore, also the traditional logic operations, like AND, OR etc, can be generalised to operate on fuzzy sets. On this basis, entire fuzzy systems can be defined, which consist of a series of fuzzy rules and an inference mechanism. In practice they allow to operate on fuzzy sets and perform inferences to derive conclusions again expressed in terms of fuzzy sets. If numerical values are required as the result of the computation, specific rules exist to perform weighted averages over the fuzzy sets to obtain such numerical outputs.

Pioneering work on the application of these fuzzy methods to disruption prediction has been carried out at JET [5]. Various predictors, with about ten signals as inputs and almost forty fuzzy rules, have been devised to provide the probability of disruption for any phase of the discharge. Their results are very competitive with other alternatives but their main advantage has proven to be their transparency. Indeed, using the information obtained from the correlation analysis performed with the CART approach and described in section 4, it has been possible to optimise the predictors for the various phases of the discharge. This work culminated in the definition of a series of predictors, each one giving the probability of disruption for a certain time in the future [6]. The outputs of these predictors, being probabilities of disruptions about future times, can be integrated in proper controller, which can evaluate the risk and decide what safety actions to undertake accordingly. Indeed, on the basis of the probability given by the predictor and the potential danger of the discharge, a coherent evaluation of the risk can be performed. Moreover, since these fuzzy predictors provide the probability for various times in the future a disruption is to be expected, the intervention strategy can be optimised taking into account the tools available to mitigate the disruption and their time response.

REFERENCES

[1]. J. Vega et al. “Encoding technique for high data compaction in databases of fusion devices”. Rev. Sci. Ins. **67**, 12 (1996)pp. 4154-4160.

[2]. J. Vega et al. “Data mining technique for fast retrieval of similar waveforms in Fusion Massive databases”. Accepted in Fus. Eng. Des.

[3]. J. Svensson et al. “Integrating Diagnostic data analysis for W7-AS using Bayesian graphical models”. Rev. Sci. Ins. **75**, 10 (2004).

[4]. J. Svensson, A. Wemer. “Current Tomography for Axisymmetric Plasmas”. Submitted to Nuclear Fusion.

[5]. G. Vagliasindi, A. Murari, L. Fortuna, P. Arena. “Fuzzy logic approach to disruption prediction at JET”. Submitted to IEEE Transactions on Plasma Science

[6]. A. Murari, G. Vagliasindi, P. Arena, L. Fortuna, O. Barana, M. Johnson. “Prototype of an adaptive disruption predictor for JET based on fuzzy logic and regression trees”. Submitted to Nuclear Fusion.

	(%)
Delta approach	93.05
WinZip	87.30
Compress() UNIX command	85.52
pack() UNIX command	36.78

TABLE 1. Compression rates for images.

[td – 440, td – 100]		[td – 200, td – 100]		[td – 320, td – 220]		[td – 440, td – 340]	
Variable	Importance	Variable	Importance	Variable	Importance	Variable	Importance
dW_{dia}/dt	100.0	dW_{dia}/dt	100.0	dl_i/dt	100.0	Ipla	100.0
dl_i/dt	82.88	dl_i/dt	50.46	dW_{dia}/dt	99.20	dl_i/dt	69.42
Ipla	70.88	$d\beta_p/dt$	37.01	Ipla	79.40	P_{net}	68.23
P_{net}	67.96	P_{net}	35.03	l_i	70.47	l_i	54.78
dq_{95}/dt	54.90	q_{95}	27.54	q_{95}	62.06	dq_{95}/dt	53.92
Dens	53.24	β_p	24.80	P_{net}	59.00	P_{inp}	49.58
$d\beta_p/dt$	52.96	l_i	24.37	β_p	57.05	dW_{dia}/dt	38.78
l_i	52.36	Loca	22.74	dq_{95}/dt	56.19	q_{95}	37.97
Loca	46.75	P_{inp}	20.49	$d\beta_p/dt$	38.68	Loca	37.39
P_{inp}	45.26	Ipla	16.36	P_{inp}	38.41	$d\beta_p/dt$	36.57
β_p	43.47	Dens	13.34	Dens	37.91	Dens	35.55
q_{95}	33.55	dq_{95}/dt	9.87	Loca	30.13	β_p	28.77

Table 2: The importance of various variables for the prediction of disruptions as derived by application of the CART software. The various columns refer to different time intervals before the disruption.

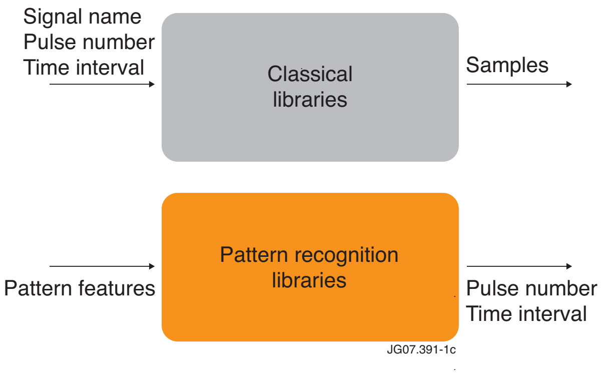

Figure 1: Comparison between the traditional structure of fusion data (top) and the new one.

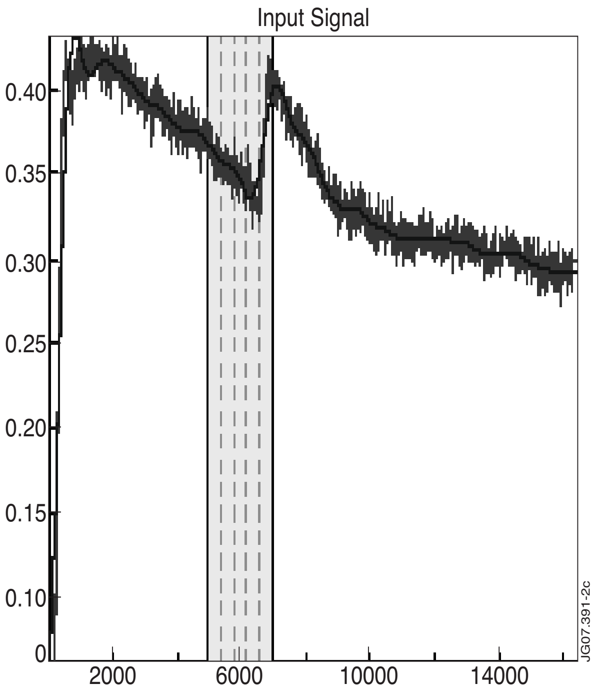

Figure 2: Pattern recognition of patterns inside signals using primitives based on the gradient.

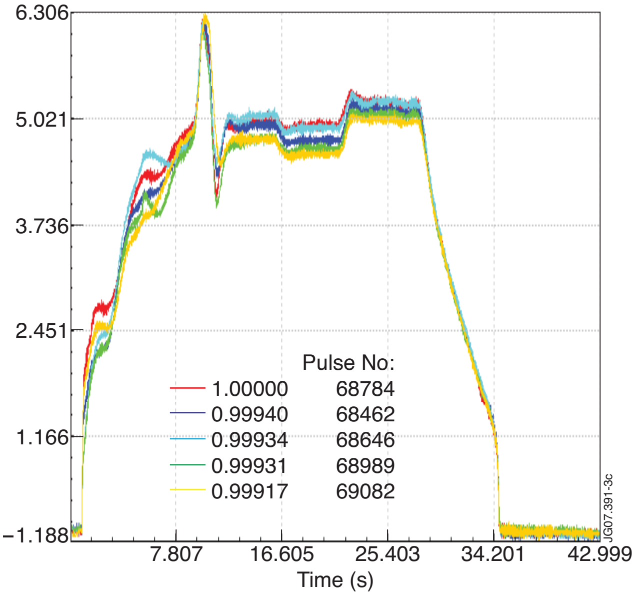

Figure 3: Pattern recognition of entire waveforms. The essential features are extracted by applying a wavelet transform. The signals are then grouped in similarity classes using a tree structure to expedite the search by avoiding the need to traverse the whole database.

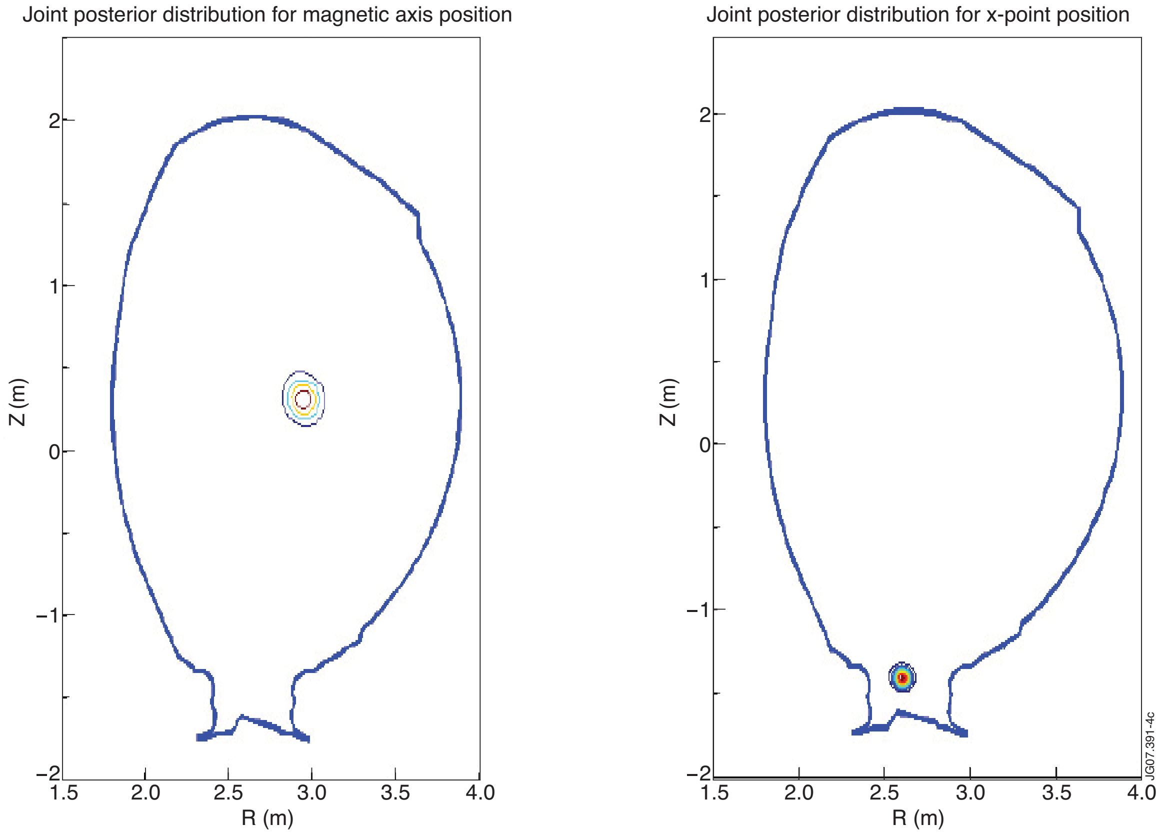

Figure 4: Error bars for the magnetic axis position (left: 6-8cm) and X-point position (right: 2-3cm).