Reduction of magnetic confinement fusion data for data mining applications

F. Detering[1] B. D. Blackwell[2] M. Hegland[3]
D. G. Pretty[4] K. Nagasaki[5] S. Yamamoto[6]

(Received 15 August 2008; revised 9 January 2009)

Abstract

We develop a practical, structured analysis of multi-channel time series measurements where the main interest lies in the coherent temporal fluctuations and spatial structures and their time dependence. The current approach to most large scale plasma experiments, tokamak and stellarators alike, is the quest for the experimental data taken under optimal conditions for each study. These data are then analysed in detail and sometimes distributed in a reference database such as the tokamak profile database of the ITER 1D Modelling Working Group. While these results are important for our understanding of future fusion devices, they do not provide easy means to support the evidence based on statistical ensembles. The raw data which are not accessible to simple search queries, are usually kept in large data repositories. At H-1, we routinely log the global experimental parameters in a summary database which is stored in a easily accessible database. In order to facilitate statistical analysis and the search for a wide class of

http://anziamj.austms.org.au/ojs/index.php/ANZIAMJ/article/view/1450 gives this article, © Austral. Mathematical Soc. 2009. Published January 29, 2009. ISSN 1446-8735. (Print two pages per sheet of paper.)

magnetic phenomena, we developed a data processing procedure that reduces the raw signal of an array of Mirnov coils at the H-1 into a series of feature descriptors in time-frequency space which are stored in an SQL-accessible database, which can be used together with the summary database.

Contents

1 Introduction

We develop a practical, structured analysis of multi-channel time series measurements where the main interest lies in the fluctuations and spatial structures. Typical examples from plasma physics research include the spatial array of magnetometers to infer ionospheric activity [12] and so-called Mirnov arrays on magnetic fusion devices that measure magnetic fluctuations inside the fusion plasma [5]. While the spectral analysis of these fluctuations is standard, we propose a tool that additionally captures the evolution and the time dependence in a highly reduced format, which is then studied using statistical tools.

We regard the process that transforms the raw data into a higher level transaction database of 'fluctuations' as a structured analysis. In the developed flow diagram the modularity becomes apparent and the processing steps are analysed and evaluated separately. Section 2 presents this top-down view.

The modules involve the separation of spatial structures with the limited information available. We apply a singular value decomposition (SVD) to a learning set of short time segments in order to identify the dominating spatial modes of the experiment.

The transaction database is primarily for the study of magnetic instabilities which are typically wavelike phenomena with fast growth and which occur over a large range of possible frequencies. The frequency is affected by the environment parameters such as the electron density and the nature of the instability and ranges from a few kHz for ion sound waves to several hundred kHz for Alfven type waves [5].

Therefore, we choose the short time Fourier transform for the time scale representation. The frequency information is immediately amenable to comparison with physical models and the uniform time and frequency resolution accommodates the above mentioned properties of the magnetic instabilities of interest.

The final step of the pre-processing, presented in Section 4, is the transformation of the power spectra into binary information and segmentation into connected regions within the time-frequency domain.

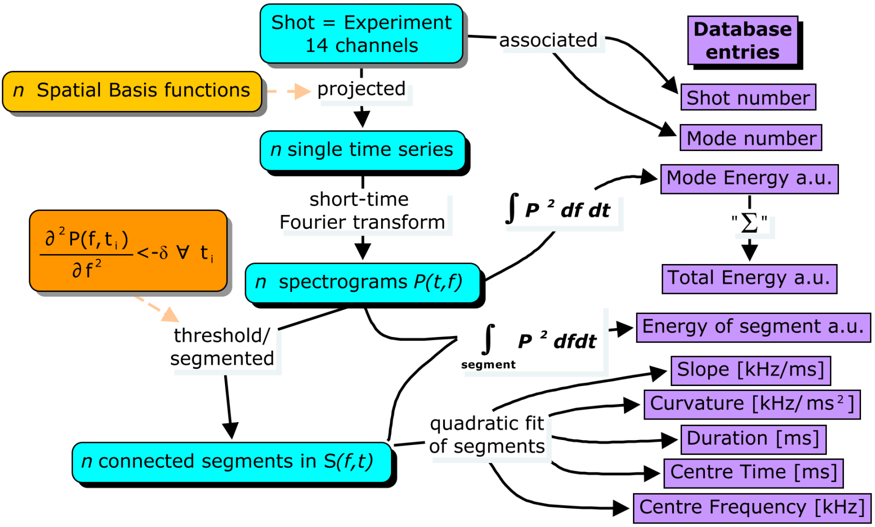

FIGURE 1: The structured analysis from raw data to the transaction database entries.

2 The analysis structure

The raw data reside in a large data repository and includes the measurements for each instantiation of the experiment. The magnetic fluctuation data are recorded from an array of 14 sensors that are distributed poloidally at a fixed toroidal angle around the plasma at approximately equidistant poloidal angles from 0 to π. We consider a fixed time window of about 90 ms which corresponds generally to the active phase of the experiment. The data are sampled at 1 MHz.

The steps of the data processing are summarised in a data flow diagram as shown in Figure 1. The input is the raw data at the top left which is then successively processed (downwards) and transferred into a transaction database (to the right).

3 Mode assignment

In the mode assignment coherent spatial modes are discovered in the data. Here a two step process is suggested which extends the one step SVD approach describe by Nardone [8]. In the first step a SVD is used to extract the main spatial components or modes of short size, strong power time segments. These modes are then clustered in a second step to find representative modes over the whole data set. As the modes of the time signals depend on the geometrical setup which does not change over the data set, we expect that similar modes occur for many different time segments. This makes the second step feasible. It was found that around 90 percent of the energy in the signal can be represented using these representative modes.

3.1 Singular value decomposition

Mode assignment methods decompose the signal into components of the form $s_i(x, t) = u_i(x)v_i(t)$. Early methods assumed sinusoidal temporal dependence of the modes, that is, $v_i(t) = a_i \sin(\omega_i t - \phi_i)$, and used Fourier decomposition methods. The SVD drops the sinusoidal condition and instead requires of the temporal parts $v_i(t)$ only that they are pairwise orthogonal. Note that the sinusoidal functions are also pairwise orthogonal if the ω_i are all multiples of a base frequency and in this case the SVD recovers the original modes. The SVD also recovers travelling waves with real valued amplitudes as a sum of two standing waves that are phase shifted by $\pi/2$. The application of the SVD approach was investigated for plasma experiments at JET [8] and has since been successfully applied to other Tokamak devices [3]. The SVD is widely used in many areas of data analysis and is very popular due to availability of robust and efficient numerical algorithms.

In order to apply the SVD the data are arranged in a $n_t \times n_c$ matrix X where an element X_{ij} of X is the value at a time step t_i and a coil j. The SVD produces a matrix decomposition of the form

$$X = USV^T\,. \tag{1}$$

Here U is an $n_t \times n_t$ orthogonal matrix, S is an $n_t \times n_c$ diagonal matrix and V^T is an $n_c \times n_c$ orthogonal matrix. The SVD, that is, the determination of U, S and V from X is a well studied problem in numerical analysis and various efficient and robust algorithms are available [4]. The SVD does not require any further knowledge and exists and is unique for any data set.

For ergodic data the product $\frac{1}{n_t}X^TX$ is an estimator for the covariance matrix of the spatial observation vectors. It follows that $X^TX = VS^2V^T$ provides a principal component decomposition and the rows of V are the principal vectors. If one selects the k largest principal values one obtains a best rank k approximation to the data. For the data studied here it was found that by only using four to five modes one is able to explain 90 percent of the variation of the data.

3.2 Mode alphabet

The SVD analysis is performed on a 'training' set of time series, which is manually chosen for the existence of strong and coherent signals. Time is divided into 1 ms intervals during which we assume the signals to be quasi-stationary.

The result of this step is an array of mode vectors that are normalised and elements of Euclidian space. Distance based clustering of the modes is therefore straightforward and we employ k-means clustering as a robust method of partitioning the 'mode space'. We use the criterion of finding the 'kink' in the plot of the total within cluster sum of squares versus the cluster number (as described by Hastie et al. [6, p.470 ff]) to find the optimal number of clusters.

The typical number is around eight cluster centers and the analysis of the cross-products shows that usually pairs within correspond to 180 degree

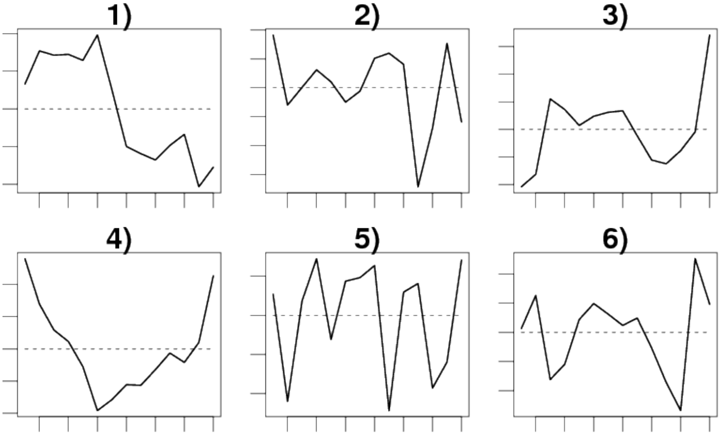

Figure 2: Six basis vector alphabet (normalised), the horizontal axis corresponds to coils approximately equispaced from 0 to π and the zero indicated by dashed lines.

phase shifts. We eliminate these duplicates and the final, automatically obtained, result is a set of six basis vectors in the case of Heliotron J, shown in Figure 2. For example, the third mode strongly resembles a mode with period $m = 3$. We also note large deviations from pure sinusoids which justifies the use of SVD over Fourier analysis.

3.3 Time scale analysis

This database focuses on MHD instabilities that occur at constant or slowly varying frequency and on phenomena which are changing in frequency over many periods of the underlying wave phenomenon. Therefore, the short time Fourier transform is performed to provide data that are easily comparable to theoretical models. Wavelets would be the obvious candidate for a complementary database that concentrates on abrupt changes in the plasma behaviour. The typical power spectrogram $P(f, t)$ is shown in the left panel

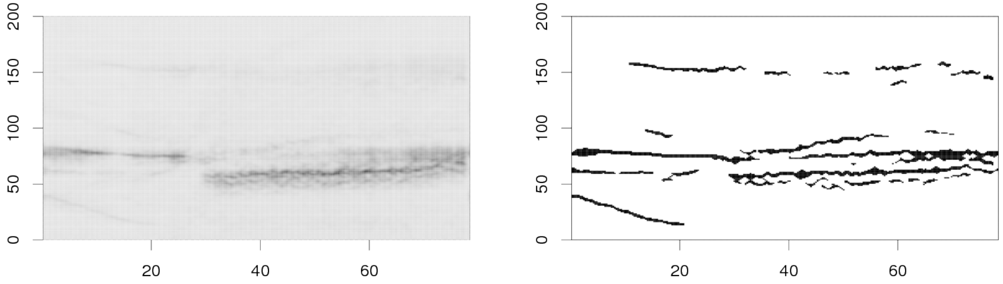

Figure 3: Original power spectrum (left) and segmentation (right) of mode three. The time on the horizontal axis is measured in ms and the frequency on the vertical axis in kHz.

of Figure 3. Some of the fluctuations are present more than one mode , but are usually highly separated in power.

4 Event extraction from spectrograms

The key idea for the reduction of the spectrogram information into a feature database and the later application to data mining techniques such as association rules is to regard the spectrogram matrix of the Fourier amplitudes as an image matrix in which we capture the spatial structure through the mode and the temporal structure, that is, frequency, duration and the time dependence, through image processing techniques applied to the spectrogram.

4.1 Binary images

The instabilities occur in a wide range of amplitudes over different experiments, during the same experiment and at various times. This is influenced by the experimental parameters and physical source of the instability. Consequently, the histogram of the power spectrum is in general not bimodal. Meaningful thresholding has to be performed in an adaptive manner, that is, over local neighbourhoods in the time-frequency domain. The disadvantage is the number of parameters to adjust such as the dimensions of the neighbourhood which determine the size (in time and frequency) of events identified and the offset that identifies a pixel as 'event pixel'.

Therefore we concentrate on frequency slices $P(f, t_i)$ at each, fixed time t_i. The multitude of possible sources and physical processing leading to spectral broadening of the signals do not permit to interpret the spectrum as a line broadened spectrum encountered in spectroscopy. Nevertheless, we borrow some of the techniques known as derivative spectroscopy that enhance spectral features [7]. Following the ansatz by Anderssen et al. [1] for higher order differentiation of non-exact data, we apply a local averaging operator and calculate the second derivative. Any interval, where $\partial^2 P(f, t_i)/\partial f^2 < -\delta$ for a small δ is then identified as 'event pixels'. This method is robust and relies only weakly (through the δ parameter to filter background long wavelength noise) on the amplitude of the events.

4.2 Image segmentation

The watershed transformation [2] is applied to the distance map of the binary image in order to define the segments.

This approach uses a minimum a priori assumptions (for example, well separated events in time, as in a slowly spoken alphabet or specific assumptions about the noise) and is able to disconnect regions that are only connected through a few pixels due to the limited frequency resolution. We use fast routines for the computation of the image distance matrix and the watershed transform from the ImageMagick [14] project that have been previously implemented into the EBImage library for use in biological image processing in R [11].

The distance map [9] is the transform of the binary image that replaces each non-zero pixel with the distance to the nearest background (zero) pixel. The principle of the watershed transform is to consider the negative of the distance map as a topographic surface. This is then flooded from its minima and, if we prevent the merging of the waters coming from different sources, we partition the image into two different sets: the catchment basins and the watershed lines. The former define the segments.

Prior to the watershed transform, the binary image is 'smoothed' with a sequential application of the morphological closing and erosion operation [10] which remove small 'holes' in the segments and smoothen the edges. While all these operations include problem dependent tuning parameters, we found that the power spectrogram images are sufficiently similar (events are small bandwidth in frequency) that the single set of parameters that we determined from a smaller number of shots works for a wide range of experimental conditions and even across the different stellarator experiments. The result of mode, Fourier analysis and segmentation is illustrated for a single shot in Figure 3, where the segmentation is displayed in the right panel.

4.3 Event attributes

The extraction of the event attributes is illustrated in the data flow diagram in Figure 4. The event which is defined by the characteristic function $S(f, t)$ which is one only for segment pixels.

The zeroth moment is defined by the point product with the original power spectrum matrix and results in the energy of that segment (see Figure 1). Since we are concentrating on the time evolution, we divide the temporal extent of the segment into three equal parts and calculate the centres of mass as indicated in Figure 4. The three centres serve as points of a quadratic fit.

The results are stored as floating point numbers in an SQL-table. As such it is accessible as metadata and can be quantised if necessary for data mining

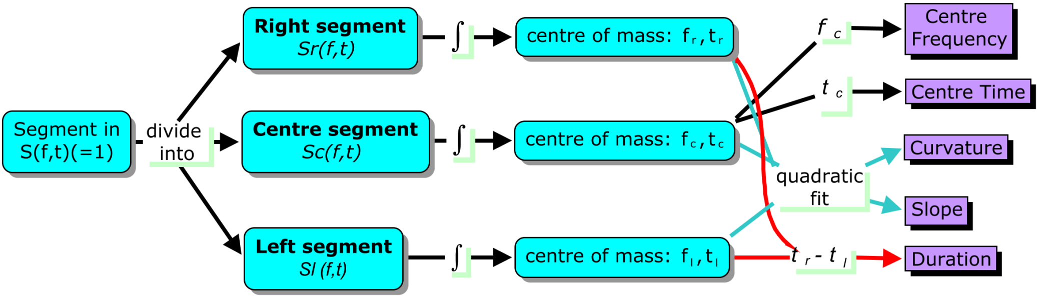

Figure 4: Flow diagram of calculation of the database entries given a segment in the spectrogram.

applications.

The marginal distribution of event frequencies is shown in Figure 5 which shows a bi-modal structure when filtered for high power events. This confirms statistically the two dominating fluctuation types previously observed. The lower frequency events are attributed to ion acoustic of drift-type fluctuations, whereas the higher frequency components which usually also are highly dependent on density and twist of the magnetic field lines have the typical scaling of so-called Alfvén type waves.

5 Summary

We developed a structured analysis of multi-channel time series data from magnetic confinement fusion experiments, that captures spatial and temporal structure and information on time dependence of plasma fluctuations. For example, 'chirps' with rapid changes of the frequency with $\delta f/f > 10\%$ have been observed, which indicate events in which there are sudden changes in plasma parameters. The procedure has already been tested on data from three different stellarator experiments: H-1 in Australia, Heliotron J in Japan and Wendelstein-7 in Germany. The analysis is modular and adaptive and

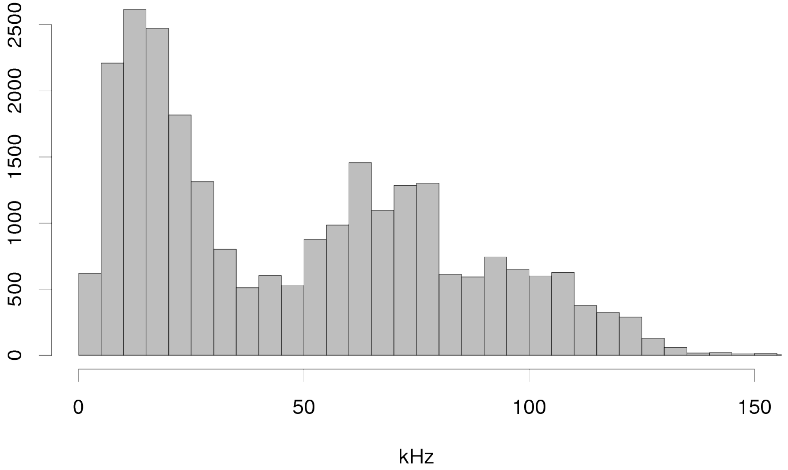

Figure 5: Histogram of centre frequencies (in kHz) in the event database.

as such applicable to different experiments where the underlying physical models are sufficiently similar.

Future work will involve a detailed analysis of the database and will extend the approach to non-equidistant and sparse sampling (in space) beyond the use of SVD techniques.

Acknowledgements This work is funded under the ARC grant DP0451960.

References

[1] R. S. Anderssen, F. De Hoog, and M. Hegland, *A stable finite difference ansatz for higher order differentiation of non-exact data*, Bull. Austral. Math. Soc. **58** (1998), 223–232. C737

[2] S. Beucher, *The watershed transformation applied to image segmentation*, Conference on Signal and Image Processing in Microscopy and Microanalysis, September 1991, pp. 299–314. C737

[3] T. Dudok de Witt, Enhancement of multichannel data in plasma physics by biorthogonal decomposition, *Plasma Physics and Controlled Fusion* **37** (1995), no. 2, 117–135, `http://stacks.iop.org/0741-3335/37/117`. C733

[4] G. H. Golub and C. F. Van Loan, *Matrix computations*, third ed., Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, 1996. MR1417720 (97g:65006) C734

[5] J. H. Harris, M. G. Shats, B. D. Blackwell, W. M. Solomon, D. G. Pretty, S. M. Collis, J. Howard, H. Xia, C. A. Michael, and H. Punzmann, Fluctuations and stability of plasmas in the H-1NF heliac, *Nucl. Fusion* **44** (2004), 279, doi:10.1088/0029-5515/44/2/008. C731

[6] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning*, Springer, August 2001. C734

[7] M. Hegland and R. S. Anderssen, Resolution enhancement of spectra using differentiation, *Inverse Problems* **21** (2005), no. 3, 915–934, `http://stacks.iop.org/0266-5611/21/915`. C737

[8] C. Nardone, Multichannel fluctuation data analysis by the singular value decomposition method. Application to MHD modes in JET, *Plasma Physics and Controlled Fusion* **34** (1992), no. 9, 1447–1465, `http://stacks.iop.org/0741-3335/34/1447`. C733

[9] A. Rosenfeld and J. L. Pfaltz, Sequential operations in digital picture processing, *J. Assoc. Comp. Mach.* **13** (1966), 471–494, doi:10.1145/321356.321357. C738

[10] J. Serra, *Image analysis and mathematical morphology*, Academic Press, Inc., Orlando, FL, USA, 1983. C738

[11] O. Skylar and W. Huber, Image analysis for microscopy screens: Image analysis and processing with EBIimage, *R. News* **6** (2006), no. 5, 1215, http://CRAN.R-project.org/doc/Rnews/Rnews_2006-5.pdf. C737

[12] R. Stening, T. Reztsova, D. Ivers, J. Turner, and D. Winch, Morning quiet-time ionospheric current reversal at mid to high latitudes, *Annales Geophysicae* **23** (2005), 385, http://adsabs.harvard.edu/abs/2005AnGeo..23..385S. C731

[13] "The ITER 1D Modelling Working Group", D. Boucher, J. W. Connor, W. A. Houlberg, M. F. Turner, G. Bracco, A. Chudnovskiy, J. G. Cordey, M. J. Greenwald, G. T. Hoang, G. M. D. Hogeweij, S. M. Kaye, J. E. Kinsey, D. R. Mikkelsen, J. Ongena, D. R. Schissel, H. Shirai, J. Stober, P. M. Stubberfield, R. E. Waltz, and J. Weiland, The international multi-tokamak profile database, *Nuclear Fusion* **40** (2000), no. 12, 1955–1981, http://stacks.iop.org/0029-5515/40/1955.

[14] S. Whitehouse, Magick with images, *Linux J.* (1998), 7, http://www.linuxjournal.com/article/2707. C737

Author addresses

1. **F. Detering**, Research School of Physical Sciences and Engineering, The Australian National University, Canberra, Australia.
mailto:frank.detering@anu.edu.au

2. **B. D. Blackwell**, Research School of Physical Sciences and Engineering, The Australian National University, Canberra, Australia.

3. **M. Hegland**, Mathematical Sciences Institute, The Australian National University.

4. **D. G. Pretty**, Research School of Physical Sciences and Engineering, The Australian National University, Canberra, Australia.

5. **K. Nagasaki**, Institute of Advanced Energy, Kyoto University.

6. **S. Yamamoto**, Institute of Advanced Energy, Kyoto University.