A data mining approach to the analysis of Mirnov coil data from a flexible heliac

D G Pretty, B D Blackwell, J H Harris[†], D Oliver, J Howard and Santhosh T A Kumar

Plasma Research Laboratory, Research School of Physical Sciences and Engineering, Australian National University, Canberra, ACT 0200, Australia
[†] Fusion Energy Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Abstract.

We present a data mining technique for the analysis of geometrically ordered multichannel data and show an application using poloidal Mirnov arrays installed in the H-1 heliac. The procedure is mostly automated, and scales well to large datasets.

Timeseries data are split into short time segments to provide time resolution, and each segment is represented by a singular value decomposition (SVD). By comparing power spectra of the temporal singular vectors, singular values are grouped into subsets which define *fluctuation structures*. Thresholds for the normalised energy of the fluctuation structure and the normalised entropy of the SVD are used to filter the dataset.

We assume that a distinct class of fluctuations is localised in the space of phase differences $\Delta\psi(n, n+1)$ between each pair of nearest neighbour channels. An expectation maximisation clustering algorithm is used to locate the distinct classes of fluctuations, and a *cluster tree* mapping is used to discover the well defined clusters.

Using this method, we present the various classes of magnetic fluctuations seen in RF powered hydrogen/helium plasmas throughout a wide range of configurations of the H-1 magnetic geometry.

1. Introduction

The H-1 flexible heliac [1, 2] is a three field-period helical axis stellarator with major radius $R = 1\,\mathrm{m}$ and minor radius $\langle r \rangle = 0.2\,\mathrm{m}$. Optimisation of the H-1 power supplies for low current ripple allows precise control of the ratio of secondary (helical, vertical) coil to primary (poloidal, toroidal) coil currents, resulting in a finely tunable magnetic geometry. Slight variation in the current ratio between shots (plasma discharges) in a sequence corresponds to a high resolution parameter scan through magnetic configurations (ie: rotational transform profile, magnetic well). The programmable control system allows for repetition rates of around 30 shots per hour, limited by data acquisition time and magnet cooling time.

Automated parameter scans allow for the generation of arbitrarily large datasets, requiring alternative methods of analysis. For a method of analysis to be scalable with the size of the dataset, manual interaction should be limited to operations on the dataset as a whole. The application of data mining techniques is widespread among diverse fields of scientific research such as astronomy, bioinformatics and the geosciences [3] where large datasets are commonplace. Generally, in fusion research a small number of shots are analysed in a campaign due to the complexity of the experiment and the high cost of operations. In the present work, we show a data mining process for the identification of classes of magnetic fluctuations using data from Mirnov coils in H-1 for a helical current ratio (κ_h) parameter scan and offer an interpretation of the results.

Perhaps the most challenging aspect in the application of data mining procedures is the pre-processing stage [4], that is the sufficient preparation of the dataset for the main algorithm to be effective. Here, we use the singular value decomposition (SVD) of the magnetic fluctuation data for short time segments, and apply energy- and/or entropy-based filters on the dataset. The filtered data are then processed by a clustering algorithm to distinguish classes of fluctuations by their phase structure.

In section 2 we explain the algorithm used for the data mining process, in section 3 we implement our method to analyse results from a configuration scan on the H-1 heliac, and a discussion of the results follows in section 4.

2. Method

Shown below is our algorithm for data mining classes of fluctuations from raw multichannel timeseries data:

Algorithm 1 Discovering fluctuation structures

```
for each shot in campaign do
  for each time segment in shot do
    take SVD of the set of RMS normalised channel timeseries
    calculate entropy $H$ of singular values
    group singular values with similar power spectra into fluctuation structures
    for each fluctuation structure in time segment do
      calculate normalised signal energy of the fluctuation structure $p$
      calculate nearest neighbour phase differences $\Delta\psi(n, n+1)$ at peak frequency
      using the inverse SVD
    end for
  end for
end for
apply filters to dataset: $H < H', p > p'$
perform clustering algorithm on the set of fluctuation structures in the space of $\Delta\psi$
select well defined clusters using cluster tree mapping
```

Each step up to, and including, the filtering process is part of the the preprocessing stage, where the dataset is conditioned for the clustering process. For each shot i, the $N_c \times N_s$ data matrix S_i has N_c rows of scalar timeseries channels; N_s is the number of samples. To provide time resolution the data are split into short time segments $S_{i,j} \equiv S_i(j\Delta t \leq t < (j+1)\Delta t)$ with $N'_s = \Delta t/f_s$ samples where f_s is the sample frequency. For each time segment we take the SVD of the ordered set of RMS normalised data, represented by the factorisation $\langle S_{i,j}\rangle_{RMS} = UAV^*$. Here, the columns of U and V contain the spatial (topo) and temporal (chrono) singular vectors respectively and the diagonal matrix A contains the $N_a = \min(N_c, N'_s)$ singular values. The convention is for the singular values to be sorted in decreasing monotonic order.

We calculate the normalised entropy H of the singular values a_k in A [5]:

$$H = \frac{-\sum_{k=1}^{N_a} p_k \log p_k}{\log N_a}, \tag{1}$$

where p_k is the dimensionless energy:

$$p_k = \frac{a_k^2}{E}, \quad E = \sum_{k=1}^{N_a} a_k^2. \tag{2}$$

The extreme case of $H = 0$ occurs when there is only one non-zero singular value, meaning $S_{i,j}$ has separable spatial and temporal singular vectors (i.e. a standing wave). In the other extreme, $H = 1$ and all singular values are equal, which occurs when $S_{i,j}$ consists only of noise. The scalar quantity H can be used to filter physically interesting signals from noise without any investigation of the structure of $S_{i,j}$.

We define a *fluctuation structure* α as a subset of singular values which have chronos with similar frequency power spectra. We measure the similarity between two chronos with a normalised cross-correlation $\gamma_{\alpha,\beta} \equiv \langle |u_\alpha \star u_\beta|^2\rangle / \langle |u_\alpha \star u_\alpha| \cdot |u_\beta \star u_\beta|\rangle$, where $u_\alpha \star u_\beta$ denotes the cross-correlation between chronos u_α and u_β. The structures are built as follows: starting with the largest unallocated singular value a_ζ, the set of unallocated singular values a_ξ for which $\gamma_{\zeta,\xi} > \gamma_{min}$ define a structure. This is repeated until all singular values have been allocated to a fluctuation structure. The normalised energy p of a structure is defined as the sum of the normalised energies of its constituent singular values. In general, fluctuation structures will consist of several singular values. For example, a rotating mode has two singular values which have their topos and chronos phase-shifted by $\pi/2$.

To distinguish fluctuation types, we use the set of phase differences between nearest neighbour channels. For each fluctuation structure α_l we take the inverse SVD to get $S'_{i,j,l}$, where singular values not in the structure are set to zero. The phase difference $\Delta\psi(n, n+1)$ between nearest neighbour channels $n, n+1$ is evaluated at the dominant frequency of the fluctuation.

We then apply filters to the dataset in order to remove noise. In the general case, we require the signal entropy to be below some threshold H', $0 < H' \leq 1$, and the normalised energy of the fluctuation structures to be greater than some value p', $0 \leq p' < 1$. A randomly selected subset of the data can be used to reduce computation.

We assume that a class of fluctuations is localised in the N_c-dimensional space of $\Delta\psi(n, n+1)$. We use a clustering algorithm to locate the classes of fluctuations, in this case an expectation maximisation (EM) clustering algorithm as implemented in the WEKA suite of data mining tools [6]. The EM algorithm is a method for estimating the most likely values of latent variables in a probabilistic model. Here we assume that each type of fluctuation can be described by a N_c-dimensional Gaussian distribution in $\Delta\psi-$space where the latent variables are the mean μ_i and standard deviation σ_i for each cluster i.

Given the initial conditions, in the form of random initial μ_i and σ_i values for a prescribed number of clusters, the EM algorithm consists of two steps which repeat until a convergence criterion is met. Firstly, the expectation step assigns to each datapoint a probability, or expectation value, of belonging to each cluster which is calculated with the Gaussian distribution function. Secondly, μ_i and σ_i are recalculated using the new expectation values as weight factors.

The WEKA algorithm cannot calculate metrics in cylindrical coordinates, so we map the $\Delta\psi-$space from the N_c-dimensional torus to a $2N_c$-dimensional cube

$[-1,1]^{2N_c}$ by taking the $\sin(\Delta\psi)$ and $\cos(\Delta\psi)$ components. The 10-fold cross-validated log-likelihood ratio is used as a measure of how well the cluster assignments fit the data. The cross-validation process involves partitioning the dataset into random subsamples and comparing results from each subset to avoid oversensitivity to outliers in the data. The likelihood is the conditional probability of obtaining the cluster means and standard deviations given the observed data.

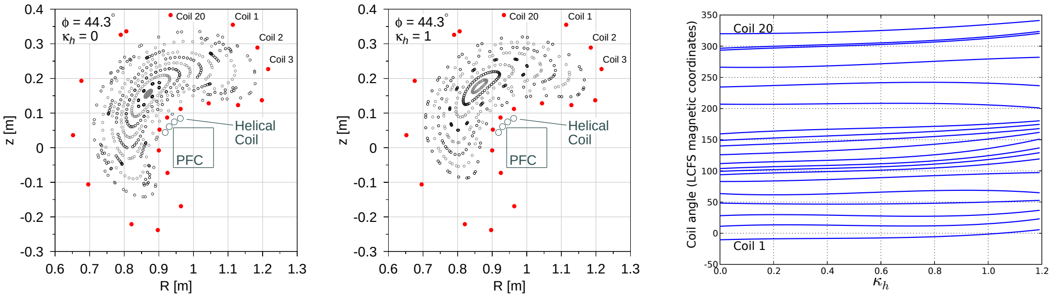
The identification of the correct number of clusters N_{Cl}, or of which ones are important, is a task that is by no means trivial to automate. We have found a *cluster tree* mapping to be a practical tool for identifying the important clusters. The cluster tree displays all clusters for each N_{Cl} below some value, with the clusters for a given N_{Cl} forming a single column. Each cluster is mapped the cluster in $N_{Cl} - 1$ with the largest fraction of common datapoints. Cluster branches which do not fork over a significant range of N_{Cl} are deemed to be well defined. The point where well defined clusters start to break up again suggests that N_{Cl} is too high.

3. Experiment

We now describe the implementation of algorithm 1 for configuration scan data from the H-1 heliac. The present work involves analysis of magnetic fluctuation data from sets of Mirnov coils at three toroidal locations. Two of these sets (at toroidal angle $\phi = 44°, 284°$) are identical 20-coil poloidal arrays, as shown in Figure 1. The coils are 100 turns with diameter of 3.2 mm and inductance of 20 μH, electrostatically shielded and isolated from the vacuum region by a bean-shaped stainless steel tube of diameter 12.7 mm and thickness of 1.2 mm. The skin-depth frequency f_δ of the tubing is 130 kHz, and the digitisation Nyquist frequency f_N is 500 kHz. The third set of Mirnov coils is a 5-coil linear array above the plasma at $\phi = 35°$, where the coils have same geometry but lie in a thinner stainless steel tubing with $f_\delta \sim 220\,kHz$. Due to calibration issues, this array is not used for toroidal mode number determination, however it can still be used for cluster definitions. Figures 1(a) and 1(b) show the locations of the Mirnov coils in the poloidal arrays and computed flux surfaces for the $\kappa_h = 0$ and $\kappa_h = 1$ configurations respectively. The positions of the coils have been chosen so that they encompass the last closed flux surface (LCFS) for all accessible magnetic geometries. As a result, for most configurations there are a few coils in sub-optimal locations. For the present analysis, we normalise the signal amplitudes and classify fluctuations by their phase structure, taking into account the shifting magnetic coordinates of the coils though the configuration scan as shown in 1(c). The set of available Mirnov coils for this campaign was $M_{44} =$ [1, 2, 3, 4, 7, 8, 9, 10, 15,16,17,18], $M_{284} =$ [1, 2, 3, 5, 7, 8, 9, 10, 15,17,18,19,20], and $M_{35} =$ [2, 4, 5].

A scan through magnetic geometry was performed by varying the ratio of helical coil to main coil current (κ_h). The range of $0 < \kappa_h < 1$ with $\Delta\kappa_h = 0.01$ corresponds to a range of rotational transform at the axis (edge) of 1.122 (1.234) $< \iota\!\!\!-_{0(a)} <$ 1.436 (1.445) with $\Delta\iota\!\!\!-_{0(a)} = 0.0031$ (0.0021). The transform profile changes from monotonic positive shear ($\iota\!\!\!-' > 0$) at $\kappa_h = 0$ to central reversed-shear at $\kappa_h = 1$. In the vaccum field model used here, the magnetic well increases from 0.4 % to $\sim$ 5.0 % over this range, with a local magnetic hill at the outer edge for low κ_h, as shown in figure 2; however recent results from error field mapping suggest a reduced magnetic well with negligible change to the transform profile [7].

Plasma discharges of 60 ms duration were produced using $50 - 60$ kW of 7 MHz ICRF in a H:He = 3:2 mixture. The ICRF antennas are conformal picture-frame coils



(a) The $\kappa_h = 0$ configuration.

(b) The $\kappa_h = 1$ configuration.

(c) Poloidal magnetic angles of Mirnov coils, evaluated at LCFS, for the parameter scan. Angles are measured clockwise, with $\phi(\text{coil1}) = 0$ for $\kappa_h = 1$.

Figure 1. The location of Mirnov coils and flux surfaces in H-1

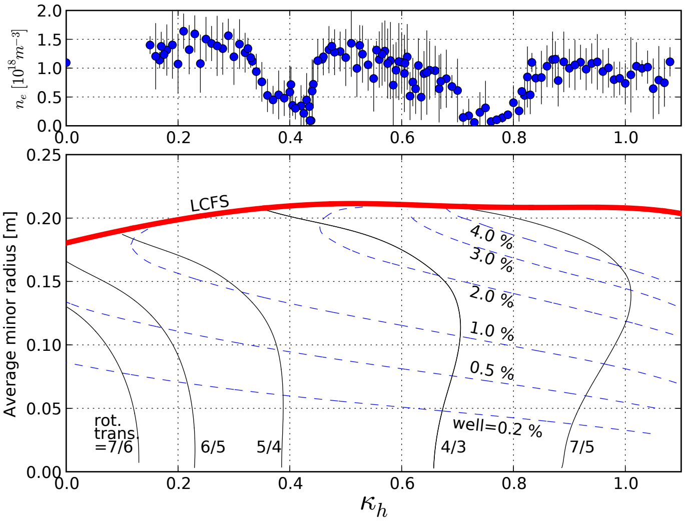


Figure 2. The top panel shows the line-averaged electron density for all H,He shots with standard RF power in the H-1 database, excluding values of κ_h which have only a single record this totals 4072 shots. The dots represent the mean value for a given value of κ_h, and the error bars are one standard deviation. The lower panel shows the radial profiles of rational ι surfaces (—) and magnetic well (- -) for the range of magnetic geometries $0 \leq \kappa_h \leq 1.1$. The bold line shows the last closed flux surface (LCFS).

located $3-4\,\mathrm{cm}$ outside the last closed flux surface [2,8]. The primary coils produce a base magnetic field of $B_0 = 0.46\,\mathrm{T}$, which is the same for all shots in the scan. Plasma parameters for these discharges are typically $n_e = 1-2\times10^{18}\,\mathrm{m}^-3$, $T_e \sim 10-20\,\mathrm{eV}$ and $T_i < 100\,\mathrm{eV}$. The normalised plasma pressure $\beta = 2\mu_0 p/B^2$ is quite low, $\ll 1\,\%$, so that the magnetic flux surfaces are those of the vacuum configuration.

A complication to the data mining procedure arises in the case of a magnetic configuration scan as the magnetic coordinates of the Mirnov coils depend on the configuration. If we were to use the phase difference between coils for analysis, then the changing magnetic angles of the coils would introduce a hidden variable throughout the configuration scan; hence we map the phase differences to κ_h-averaged (virtual coil) angles instead, where the angles are evaluated at the outermost flux surface. During the preprocessing stage, we use $\gamma_{min} = 0.8$ to define fluctuation structures, apply an energy threshold $p' = 0.4$. From this dataset, 2000 randomly selected datapoints are used in the clustering step with the remainder mapped back to the clusters once the clusters are defined.

The cluster tree representation is shown in figure 3 in the frequency and magnetic configuration (κ_h) coordinates. The Gaussians distributions of nearest neighbour phase differences which define the clusters are shown in figure 4. As N_{Cl} is increased the clusters become more well defined by thinner Gaussian distributions. At $N_{Cl} = 3$ we have the well defined clusters 5 and 6 centred on the $\iota(\iota' = 0) = 4/3$ and $5/4$ configurations respectively. The branch stemming from cluster 9 at $N_{Cl} = 4$ contains modes with $(n, m) = (0, 0)$. At $N_{Cl} = 10$ the $\iota(\iota' = 0) = 7/6$ and $6/5$ configurations are represented by clusters 54 and 50 respectively, while cluster 51 is associated with $\iota \simeq 7/5$.

4. Discussion

We note that the frequency spectra seen in the Mirnov signals have corresponding n_e fluctuations which are observed with an electronically scanned interferometer [9], as shown in figure 5; however, it is not possible to do an accurate correlation between $\dot{B}$ and $\dot{n}_e$ with these diagnostics. The fluctuations cannot be attributed simply to magnetic island activity as at least some fluctuations remain active for clusters with $\iota(\iota' = 0) \gtrsim n/m$ where the (n, m) rational surface is not present. The dependence of fluctuation frequency on rotational transform is suggestive of torsional Alfvénic activity. The global Alfvén eigenmode (GAE) frequency lies below the Alfvén continuum which, in a cylindrical model, has a lower boundary of $f_{GAE} = (2\pi)^{-1}(m/R)|\iota - n/m|v_A$, where $v_A = B/\sqrt{4\pi\rho}$ is the Alfvén velocity, ρ is the mass density, and R is the major radius [10, 11].

Shown in figure 6 are the observed Mirnov frequencies compared to GAE frequencies scaled by a factor of $\lambda = 1/3$ for two configurations in cluster 5. The existence of the $(n, m) = (4, 3)$ rational surface within the plasma volume for configurations below $\kappa_h = 0.74$ introduces a root in the radial f_{GAE} profile; in this case we take the frequency below the local maximum (at $\langle r\rangle \sim 0.12\,\mathrm{m}$ in figure 6(a)). The scale factor λ may be explained by an increased effective mass density caused by neutral collisions or the presence of impurities, although the scale factor is smaller than expected for these phenomena. Also, comparisons between Alfvén eigenmodes computed for cylindrical and stellarator geometries have shown the latter can exist at lower frequency [12]. A very small offset $\delta_\iota \simeq 7\times10^{-3}$ from the vacuum rotational transform profile from is required for an optimal fit to the GAE scaling. It is likely

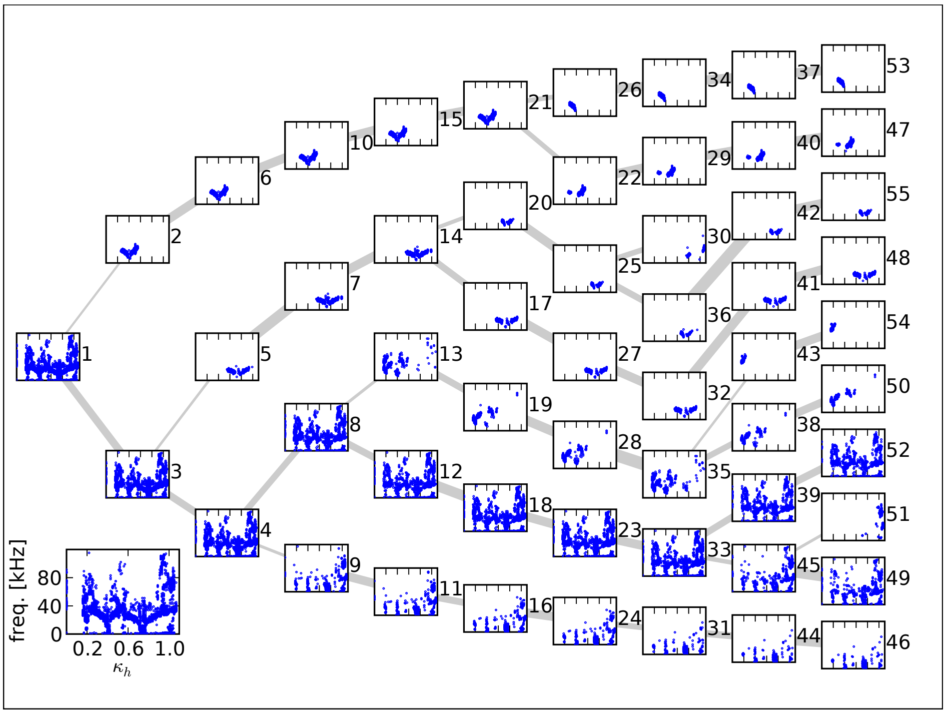


Figure 3. Cluster tree representation of the Mirnov data. The sum of clusters in each column is the entire data set.

this transform offset is due to error fields or the presence of toroidal currents which, though not measured during this campaign, have been observed at these configurations in lower field argon discharges.

Frequency in cluster 5 is shown to scale with ι and n_e in figure 7(a) with λ and δ_ι corrections applied. The structure of cluster 6 (figure 7(b)) is more complicated; it shows the same gradient $d(f\sqrt{n_e})/d|\iota - \delta_\iota - n/m| = mA$ as cluster 5, where $A = B/(4\pi^{3/2}Rm_i^{1/2})$, however frequency offsets Δ_a and Δ_b are also present. There is a coincidental change in the radial transform profile between the two slopes of cluster 6, with the $\kappa_h < 0.4$ (Δ_a) side having a monotonic profile (see figure 2) and the $\kappa_h > 0.4$ (Δ_b) side having a zero-shear region within the plasma volume. The difference in profile may account for the different values of Δ_a and Δ_b through different inaccuracies of the cylindrical model assumption or variation in toroidal current profile.

We have assumed throughout that a class of fluctuations can be described by a Gaussian distribution in $\Delta\psi-$space. While such an assumption is necessary for the EM clustering algorithm, there are alternative algorithms which do not include assumptions about the shape of clusters at the expense of additional computation. We have obtained the same qualitative results with an agglomerative hierarchical clustering algorithm which we will now describe in brief. As an initial condition each fluctuation structure defines a cluster so that $N_{Cl} = N_\alpha$, where N_α is the number of fluctuation structures. The two clusters with least distance between them in $\Delta\psi-$space are agglomerated into a single cluster; this process which maps $N_{Cl} \rightarrow N_{Cl} - 1$ is repeated until $N_{Cl} = 1$, with cluster definitions being recorded

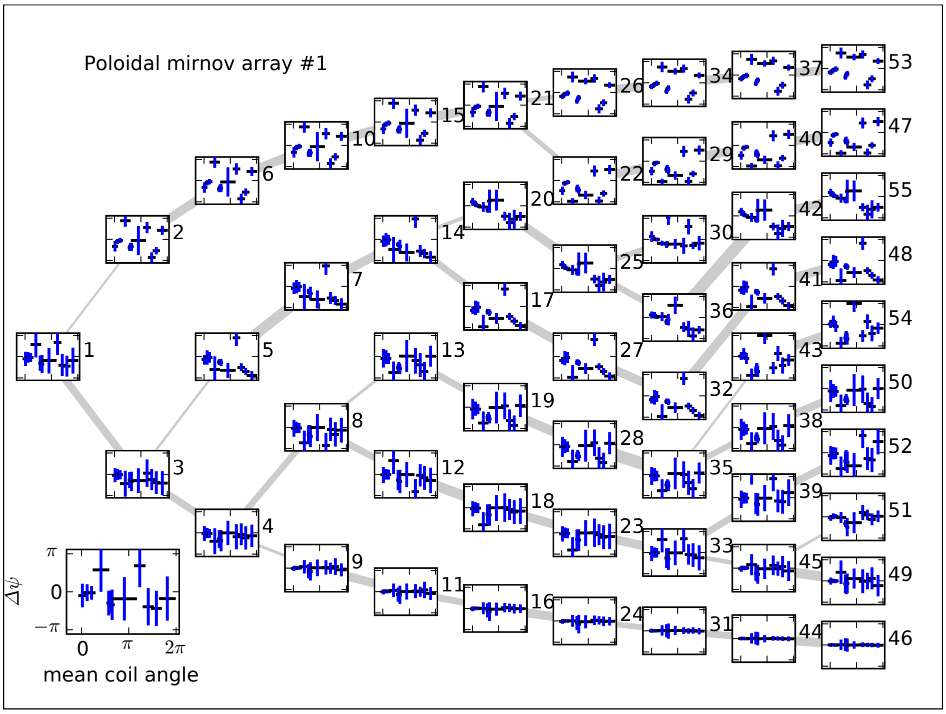
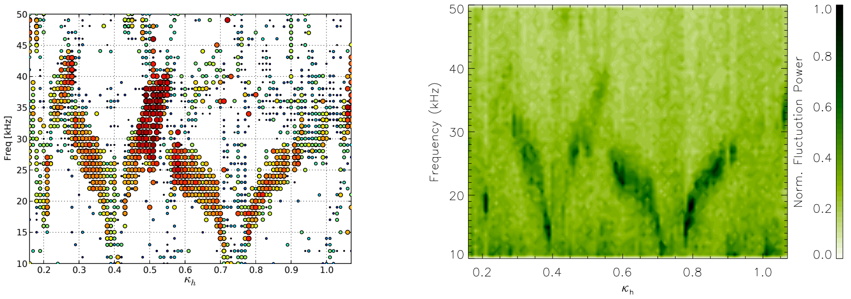


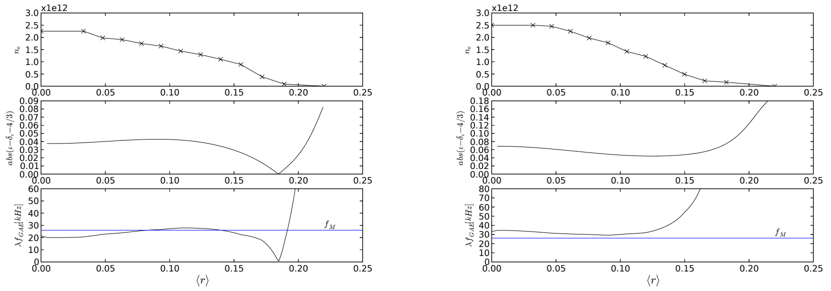
Figure 4. Cluster tree showing cluster definitions. Only poloidal phase differences for the poloidal Mirnov array at $\phi = 44°$ are shown, which is a projection of the cluster definition in the higher dimensional space of all nearest neighbour coil phases. The vertical error bars are standard deviations of the Gaussians which define the clusters.



(a) Magnetic fluctuations

(b) Electron density fluctuations

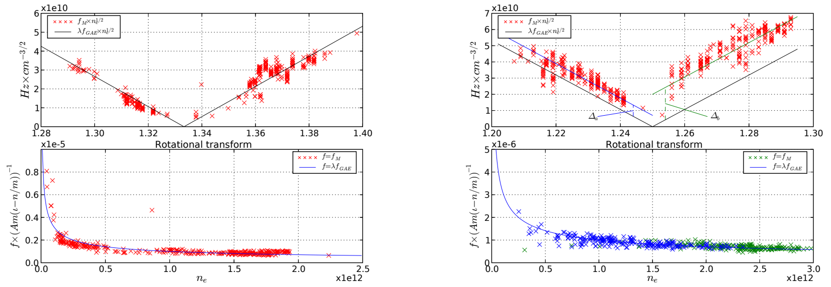
Figure 5. Common spectra are observed by Mirnov coils and the scanned interferometer throughout the configuration scan.



(a) Shot 58064, $\kappa_h = 0.56$, $t = 18\,\mathrm{ms}$.

(b) Shot 58082, $\kappa_h = 0.92$, $t = 30\,\mathrm{ms}$.

Figure 6. GAE scaling for configurations (a) with and (b) without the 4/3 rational surface in the plasma volume using the cylindrical model with $R = 1.0\,\mathrm{m}$, $B_0 = 0.46\,\mathrm{T}$ and $m_i/m_p = 4.0$. The top panel shows the electron density profile and the middle panel shows $|\iota - \delta_\iota - 4/3|$ with $\delta_\iota = 7.2 \times 10^{-3}$. The bottom panel has a frequency scaling of $\lambda = 1/3$ applied to f_{GAE} in both cases.



(a) Cluster 5.

(b) Cluster 6. The frequency offsets Δ_a and Δ_b are discussed in the text.

Figure 7. GAE scaling for clusters (a) 5 and (b) 6 located about the $\iota \sim 4/3$ and 5/4 configurations respectively. The top panel shows a comparison of $f_M\sqrt{n_e}$ and $\lambda f_{GAE}\sqrt{n_e}$ against ι; the bottom panel compares the density scaling where $A = B/(4\pi^{3/2}Rm_i^{1/2})$. The plasma parameters and frequency scaling are the same as for figure 6.

for each value of N_{Cl}. Unlike the EM clusterer, this algorithm requires a distance calculation between each pair of datapoints giving quadratic computational complexity $O(n^2)$ and making it unsuitable for very large datasets.

Work is beginning on the modelling of these modes in stellarator geometry. This should reduce the gap between experiment and theory. The spectrum as a whole appear to be a combination of resistive interchange and Alfvénic activity. Resistive interchanges have been observed in stellarators across a wide range of parameters [13] but it is remarkable to observe driven Alfvén eigenmodes in a device with $T_e < 100\,\mathrm{eV}$. These instabilities have generally been studied in devices with parameters approaching those of fusion reactors, however recent results indicate that Alfvén modes can be excited by thermal ions traveling well below the Alfvén velocity [14].

Finally, we note that although this experiment involves a large number of plasma discharges with parameters varied from shot to shot, this data mining technique is equally amenable to the analysis of steady state operations or long discharges with variation in plasma properties within the shot.

Acknowledgements

The authors would like to thank the H-1 team for continued support of experimental operations as well as R. Dewar, B. McMillan, T. Luce and M. Hegland for useful discussions. This work is supported in part by the U.S. Department of Energy under Contract No. DE-AC05-00OR22725 with UT-Battelle, LLC. This work was performed on the H-1NF National Plasma Fusion Research Facility established by the Australian Government, and operated by the Australian National University, with support from the Australian Research Council Grant DP0344361 and DP0451960.

References

[1] S M Hamberger, B D Blackwell, L E Sharp, and D B Shenton. *Fusion Technol.*, 17:123–30, 1990.

[2] 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. *Nucl. Fusion*, 44:279–286, 2004.

[3] U Fayyad, D Haussler, and P Stolorz. Mining scientific data. *Commun ACM*, 39(11):51–57, 1996.

[4] J Han and M Kamber. *Data Mining: Concepts and Techniques*, chapter 3. Morgan Kaufmann, 2001.

[5] T Dudok de Wit, A L Pecquet, J C Vallet, and R Lima. *Phys. Plasmas*, 1(10):3288–3300, 1994.

[6] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2nd edition, 2005.

[7] Santhosh T A Kumar et al. Accurate determination of the magnetic geometry of the h-1nf heliac. in preparation, 2006.

[8] G G Borg, B D Blackwell, and S M Hamberger et al. *Fusion Eng. and Design*, 26:191–201, 1995.

[9] J Howard and D Oliver. Appl. Opt. (Accepted for publication, 2006).

[10] K-L Wong. *Plasma Phys. Control. Fusion*, 41(1):R1–R56, 1999.

[11] A Weller et al. *Phys. of Plasmas*, 8(3):931–56, 2001.

[12] D A Spong, R Sanchez, and A Weller. Shear alfven continua in stellarators. *Phys. of Plasmas*, 10(8):3217–24, 2003.

[13] J H Harris et al. Magnetohydrodynamic activity in high-β, currentless plasmas in heliotron-e. *Phys. Rev. Lett.*, 53(23):2242–5, 1984.

[14] R Nazikian et al. Multitude of core-localized shear alfven waves in a high-temperature fusion plasma. *Phys. Rev. Lett.*, 96:105006, 2006.