EFDA–JET–CP(14)03/11

A. Murari, E. Peluso, M. Gelfusa, I. Lupelli, P. Gaudio
and JET EFDA contributors

New Data Analysis Methods to Maximize the Physics Information Which can be Derived from Diagnostic Measurements

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 It is made available on the understanding that it may not be further circulated and extracts or references may not be published prior to publication of the original when applicable, or without the consent of the Publications Officer, EFDA, Culham Science Centre, Abingdon, Oxon,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

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

The contents of this preprint and all other JET EFDA Preprints and Conference Papers are available to view online free at www.iop.org/Jet. This site has full search facilities and e-mail alert options. The diagrams contained within the PDFs on this site are hyperlinked from the year 1996 onwards.

New Data Analysis Methods to Maximize the Physics Information Which can be Derived from Diagnostic Measurements

A. Murari[1], E. Peluso[2], M. Gelfusa[2], I. Lupelli[2,3], P. Gaudio[2]
and JET EFDA contributors*

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

[1]Consorzio RFX-Associazione EURATOM-ENEA per la Fusione, I-35127 Padova, Italy.
[2]Associazione EURATOM-ENEA – University of Rome "Tor Vergata", Roma, Italy
[3]Culham Centre for Fusion Energy, Abingdon, Oxfordshire, OX14 3DB, United Kingdom
** See annex of F. Romanelli et al, "Overview of JET Results",*
(24th IAEA Fusion Energy Conference, San Diego, USA (2012)).

Preprint of Paper to be submitted for publication in Proceedings of the
41st EPS Conference on Plasma Physics, Berlin, Germany
23rd June 2014 – 27th June 2014

ABSTRACT

Many measurements are required to control thermonuclear plasmas and to fully exploit them scientifically. In the last years JET has shown the potential to generate about 50 Gbytes of data per shot. These amounts of data require more sophisticated data analysis methodologies to perform correct inference and various techniques have been recently developed in this respect. The present paper covers a new methodology to extract mathematical models directly from the data without any a priori assumption about their expression. The approach, based on symbolic regression via genetic programming, is exemplified using the data of the ITPA database for the energy confinement time. The best obtained scaling laws are not in power law form and suggest a revisiting of the extrapolation to ITER. On the other hand, more comprehensive and better databases are required to fully profit from the power of these new methods and to discriminate between the hundreds of thousands of models that they can generate.

1. INTRODUCTION TO THE DATA ANALYSIS PROBLEM IN FUSION

Thermonuclear plasmas are nonlinear, open systems, kept well out of equilibrium to maximize their efficiency and rate of energy production. They therefore present all the problems typical of open systems and living organisms, from the need of adequate inputs of energy and matter to stringent requirements in terms of exhaust, control and purity. They are also characterised by a very high level of complexity, which practically prevents the formulations of theories from basic principles. This leads to a hierarchy of descriptions of thermonuclear plasmas (particle, kinetic, fluid) and to a plethora of ad hoc models of limited applicability (see for example the L-H transition, for which an undisputed control parameter has not been found yet, or the MHD treatment with its various stability regions and the variety of instabilities). This plurality of models is not a fault per se but more a specific characteristic of complex systems; on the other hand, robust statistical techniques are required to assess the quality of the various alternative descriptions of the phenomena.

These challenges are to be tackled first by analysing the large amounts of data produced by present day diagnostics. In JET for examples, about 50 Gigabytes of data can be generated in a well diagnosed discharge and the whole database now exceeds 250 Terabytes. On the other hand, given the lack of a unifying theory and the large amounts of data available, sometimes important information remains hidden in the databases and it can be difficult to identify it manually or with traditional methods. To overcome this difficulty, in the last years a series of new data analysis tools have been developed to increase the physics output that can be derived from the measurements. These data mining tools consist of sophisticated techniques of correlation and pattern recognition to discover within the databases useful and understandable knowledge that was previously unknown. Their application ranges from exploration of the physics to pattern recognition and to prediction [1-5]. To illustrate the potential of the tools for the extraction of mathematical models directly from the data, in this paper the subject of the determination of scaling laws has been chosen. New regression tools, indicated collectively by the term symbolic regression, have been developed to determine the

mathematical form of various scaling laws for the energy confinement time in Tokamaks. These methods exploit genetic algorithms and do not assume a priori that the scaling laws are in power law monomial form.

2. SYMBOLIC REGRESSION VIA GENETIC PROGRAMMING

As mentioned in the previous section, this paper describes the application of advanced techniques of symbolic regression (SR) via genetic programming (GP) to the problem of deriving scaling laws for the confinement time from large databases. The main advantage of the proposed approach consists of practically eliminating any assumption about the mathematical form of the scaling laws. The methods developed indeed allow identifying the most appropriate mathematical form for the regression equations, the Best Unconstrained Empirical Model Structure (BUEMS), and to demonstrate that it has the potential to better interpret the present experimental data for the confinement time in comparison with traditional scalings. Solutions of varying levels of complexity can be generated and evaluated to obtain the best trade-off between accuracy and computational complexity. In this study, SR analysis has been implemented with a GP approach. SR via genetic programming is a non-parametric, non-linear technique that looks both for the appropriate model structure and the optimal model parameters simultaneously[6,7]. This approach provides a natural extension of the traditional linear and nonlinear regression methods that fit parameters to an equation of a given mathematical structure. The first step is the generation of the initial population of CPs (formulas in our case) and then the algorithm finds out how well an element of the population works evaluating its behaviour to some appropriate metrics. This assessment is quantified by a numeric value called fitness function (FF). In the second phase, as with most evolutionary algorithms, genetic operators (Reproduction, Crossover and Mutation) are applied to individuals that are probabilistically selected on the basis of the FF, in order to generate the new population. That is, better individuals are more likely to have more child elements than inferior individuals. When a stable and acceptable solution, in terms of complexity, is found or some other stopping condition is met (e.g., a maximum number of generations or acceptable error limits are reached), the algorithm provides the solution with best performance in terms of the FF.

In this work, CPs are composed of functions and terminal nodes and can be represented as a combination of syntax trees. The function nodes can be standard arithmetic operations and/or any mathematical functions, squashing terms as well as user-defined operators. The function nodes included in the analysis performed in this paper are reported in Table 1.

The fitness function is a crucial element of the genetic programming approach and it can be implemented in many ways. To derive the results presented in this paper, the AIC criterion has been adopted [8] for the FF. The AIC form used is:

$$AIC = 2k + n \cdot \ln(RMSE/n) \tag{1}$$

In equation (1), RMSE is the Root Mean Square Error, k is the number of nodes used for the model and n the number of y_{data} provided, so the number of entries in the database (DB). The FF parameterized above allows considering the goodness of the models, thanks to the RMSE, and at the same time their complexity is penalised by the dependence on the number of nodes. To assess the quality of the final models the well-known criteria of BIC and Kullback-Leibler divergence have been used.

3. THE MAIN CHARACTERISTICS OF THE ITPA DATABASE DB3v13f

To maximise the generality of the results obtained with the methodology described in the previous sections, an international database has been considered [9]. This database was explicitly conceived to support advanced studies of the confinement time and includes validated signals from the vast majority of the most relevant Tokamak machines ever operated in the world. In line with the previous literature on the subject, the following quantities have been considered good candidate regressors in the present work: $B[T]$, $I[MA]$, $n[10^{19}m^{-3}]$, $R[m]$, M, ε, k_a; $P[MW]$. In the previous lists, k_a indicates the volume elongation measurement, ε the inverse aspect ratio, q_{95} the plasma safety factor evaluated at the flux surface enclosing the 95% of the poloidal flux, M the effective atomic mass in a.m.u, n the central line average plasma density, B the toroidal magnetic field, R the plasma major radius, I the plasma current and finally P the estimated lost power [10]. All the selected quantities are generally known with accuracy better than ± 20% and they are routinely available in all the major Tokamaks, providing enough data for a sound statistical analysis. The entries of the database, for which all the variables used are available or computable, have been considered, resulting in a total of 3093 entries, corresponding to the DB3 dataset [9].

4. RESULTS IN TERMS OF DIMENSIONAL QUANTITIES

The best models found with symbolic regression via genetic programming in general are not in power law form but they present additional terms. In particular, the most performing models include multiplicative squashing terms in the plasma current and plasma density. It is interesting to note that the method by itself selects these two quantities as the ones responsible for the saturation of the confinement time. The effect of the plasma volume, represented by the major radius R, is not found to be affected as intuitively expected.

One of the best functional forms for the energy confinement time is reported in Table II. Also the power laws (PL1 and PL2) typically used as reference by the community are reported: PL1 is IPB98(y,2) and PL2 is EIV of [10]. The most important aspect of the non-power law (NPL) functional form is the presence of a squashing term for the density whose effect can be clearly seen in Figure 1.

The equations of the scaling laws have been used to generate their estimate corresponding to the values of the database and their pdfs have been compared with the experimental ones. The KL divergence, the MSE and the BIC and AIC criteria all show that the NPL scaling is better than the PLs in interpreting the experimental data available. The comparison between the traditional PL

and the NPL scalings, in terms of statistical indicators, is summarised in Table III, proving the best quality of the NPL regression.

Even if the superior properties of the NPL scaling are quite consolidated in statistical terms, its extrapolation capability could still be questioned and has therefore been checked with additional tests. A first test has been performed by fitting the NPL model on the database for currents below 2.5MA and then extrapolating this model to higher current data; all the criteria agree that the NPL scaling performs better than the PLs ($KL_{NPL} = 0.2347$, while $KL_{PL1} = 0.3210$ or $KL_{PL2} = 0.7478$).

A second more relevant test has been performed, by non-linearly fitting the various models using data of smaller devices and testing the results with JET data. Since the PL1,2 models have the same variables, only one fit has been performed and labelled as PLs. The non linear fits for the PLs and the NPL have been performed using the weights obtained by the percentiles method. This technique allows weighting data depending on their distribution in order to give more importance to the data falling in their tails since they carry more relevant of information for the determination of scaling laws (e.g. high plasma current values ($I > 2.5[MA]$) of JET are extremely valuable since they are the closest to ITER operational region of 15MA). For each physical quantity, five percentiles have been computed in order to divide the distribution function in six partitions, which can be independently weighted. In our case we have chosen the inverse of the cumulative probability, defining the percentiles themselves, as the weight for data falling in each different partition. This can be repeated for all the selected physical quantities to obtain a final weight for each entry of the DB.

The PLs perform slightly better on the small machines ($MSE_{PLs} = 3.322 \cdot 10^{-4} s^2$ and $KLD_{PLs} = 0.0516$; while $MSE_{NPL} = 3.554 \cdot 10^{-4} s^2$ and $KLD_{NPL} = 0.0630$, but when the scaling are applied to JET data, the superiority of the NPL model in terms of the indicators considered can be clearly seen; the fitted models are reported in Table IV and their statistical estimators for JET data in Tab.V.

The higher extrapolation capability of the scaling laws in NPL form motivate a revision of the expected performance in terms of confinement time for ITER. Using the equations of Table II, the predicted value of the confinement time for ITER ($n_e = 1.03[10^{20} m^{-3}]$ $\kappa_\alpha = 1.70$, Ip = 15[MA], R = 6.2 [m], P = 87 [MW]) is about $2.83^{3.31}_{2.42}$ seconds to be compared to then $3.6^{3.13}_{4.14}$ seconds of the traditional extrapolations obtained with power law scalings. There is therefore a basic agreement within the confidence intervals between the various estmates.

On the other hand, the solution of Table II is not the only candidate. Other models have quite good statistical performance a similar physics substance and credibility. One example is reported in Table VI. This time two squashing terms are present. The high quality of this model can be seen in table VII which summarises its statistical performance. The extrapolation to ITER of this model would give more pessimistic estimate for the confinement time ($1.8^{2.4}_{1.3}$ seconds).

5 CONCLUSIONS AND FURTHER DEVELOPMENTS

In this paper, symbolic regression via genetic programming has been applied to the derivation of empirical scaling laws for the energy confinement time in Tokamaks. The analysis has been

particularized for the H-mode of confinement. The examples used for the present investigation belong to the ITPA international database, which include all of the most relevant machines in the world.

Contrary to the main results reported in the literature, the obtained empirical scalings are not in the form of power laws. Indeed they present either multiplicative squashing terms. The superior quality of these new scalings, compared to the traditional power laws, has been demonstrated first of all with the help of a series of statistic indicators. To complement this analysis, the extrapolation capability of the new scalings has been verified by dedicated investigations of different group of devices (small and large machines). On the basis of the new found scalings, the confinement time to be expected in an ITER class device could be significantly lower than the predictions of the traditional power laws. This is due to the excess rigidity of the power law scalings, which probably tend to overestimate the confinement time in the case of large extrapolations. On the other hand, a specific value of the confinement time cannot be given now since the database is not of enough quality to discriminate between various equally satisfactory models. Better databases and/or specific experiments will have to be considered to narrow down the best estimate for the confinement time in ITER

REFERENCES

[1]. A. Murari et al, 2008 Nuclear Fusion **48** 035010 doi:10.1088/0029-5515/48/3/035010

[2]. S. Dormido-Canto et al, 2013 Nuclear Fusion **53** 113001 doi:10.1088/0029-5515/53/11/113001

[3]. A. Murari et al, Nuclear Fusion **48** (2008) 035010 (10pp).

[4]. A. Murari, et al, Nuclear Fusion **53** (2013) 033006 (9pp)

[5]. A. Murari et al, 2012 Nuclear Fusion **52** 063016 doi:10.1088/0029-5515/52/6/063016

[6]. M. Schmid, H. Lipson, Science, **Vol. 324**, April 2009

[7]. Koza J.R., Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, MA, USA (1992).

[8]. Kenneth P. Burnham, David R. Anderson (2002), Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach. Springer. (2nd ed)

[9]. http://efdasql.ipp.mpg.de/hmodepublic/DataDocumentation/Datainfo/DB3v13/db3v13.htm

[10]. McDonald D.C. at al, Nuclear Fusion **47** (2007).147–174

Function class	List
Arithmetic	c (real and integer constants),+,-,*,/
Exponential	$\exp(x_i)$,$\log(x_i)$,power(x_i, x_j), power(x_i,c)
Squashing	logistic(x_i),step(x_i),sign(x_i),gauss(x_i),tanh(x_i), erf(x_i),erfc(x_i)

Table I: Types of function nodes included in the symbolic regression used to derive the results presented in this paper, x_i andx_j are the generic independent variables.

PL1	$5.62 \cdot 10^{-2} I^{0.93} B^{0.15} n^{0.41} M^{0.19} R^{1.97} \epsilon^{0.58} \kappa_a^{0.78} P^{-069}$
PL2	$5.55 \cdot 10^{-2} I^{0.75} B^{0.32} n^{0.35} M^{0.06} R^{2.0} \epsilon^{0.76} \kappa_a^{1.14} P^{-062}$
NPL1	$0.070_{0.069}^{0.071} \cdot I^{1.071_{1.062}^{1.079}} R^{1.706_{1.685}^{1.727}} \kappa_a^{1.250_{1.211}^{1.290}} P^{-0.715_{-0.723}^{-0.707}} h(n)$

Table II: Power laws (PLs) and Non Power Law model (NPL). PL1 is the IPB98(y,2) scaling while PL2 is (EIV)[15]. The term h(n) is:

$$h(n) = n^{0.100_{0.091}^{0.109}} \cdot \left(1 + e^{-0.408_{-0.426}^{-0.390} \cdot n^{1.036_{0.996}^{1.108}}}\right)^{-1}$$

	k	AIC	BIC	MSE $[s^2]$	KLD
PL1	10	-19416.86	-19362.86	$1.866 \cdot 10^{-3}$	0.0337
PL2	10	-19084.36	-19203.68.	$2.077 \cdot 10^{-3}$	0.0802
NPL	9	-19610.81	-19556.55	$1.753 \cdot 10^{-3}$	0.0255

Table III: Statistical estimators used to qualify the scaling reported in Table II. The KLD has been computed in a range of ± 6σ around the mean value of the data.

PLs	$5.35_{5.30}^{5.39} \cdot 10^{-2} I^{0.67_{0.62}^{0.73}} B^{0.12_{0.07}^{0.17}} n^{0.38_{0.35}^{0.41}} M^{0.43_{0.35}^{0.51}} R^{1.69_{1.59}^{1.78}} \epsilon^{0.53_{0.44}^{0.63}} \kappa_a^{0.39_{0.29}^{0.48}} P^{-0.54_{-0.57}^{-0.52}}$
NPL	$0.074_{0.073}^{0.075} \cdot I^{0.954_{0.927}^{0.981}} R^{1.369_{1.323}^{1.415}} \kappa_a^{0.219_{0.202}^{0.236}} P^{-0.514_{-0.539}^{-0.489}} h(n)$

Table IV: Power laws (PLs) and Non Power Law model (NPL) fitted on the subset of data without the JET's entries. The term h(n) is:

$$h(n) = n^{0.139_{0.120}^{0.158}} \cdot \left(1 + e^{-0.350_{-0.378}^{-0.322} \cdot n^{1.363_{1.245}^{1.480}}}\right)^{-1}$$

	k	AIC	BIC	MSE $[s^2]$	KLD
PLs	10	-5842.20	-6720.79	$1.578 \cdot 10^{-2}$	5.895
NPL	9	-6052.17	-6758.98	$1.360 \cdot 10^{-2}$	2.831

Table V: Statistical estimators used to qualify the extrapolations to JET data using the equations derived for the smaller devices.

PL1	$5.62 \cdot 10^{-2} I^{0.93} B^{0.15} n^{0.41} M^{0.19} R^{1.97} \epsilon^{0.58} \kappa_a^{0.78} P^{-069}$
PL2	$5.55 \cdot 10^{-2} I^{0.75} B^{0.32} n^{0.35} M^{0.06} R^{2.0} \epsilon^{0.76} \kappa_a^{1.14} P^{-062}$
NPL	$0.101_{0.099}^{0.102} \cdot R^{1.666_{1.621}^{1.712}} \kappa_a^{1.270_{1.192}^{1.346}} P^{-0.716_{-0.732}^{-0.700}} h(n) g(I)$

Table VI: Power laws (PLs) and Non Power Law model (NPL). PL1 is the IPB98(y,2) scaling while PL2 is (EIV)[15]. The two terms h(n) and g(n) are:

$$h(n) = n^{0.130_{0.114}^{0.147}} \cdot \left(1 + e^{-0.501_{-0.535}^{-0.468} \cdot n^{1.055_{0.972}^{1.139}}}\right)^{-1}$$

$$g(I) = I^{0.760_{0.743}^{0.776}} \cdot \left(1 + e^{-0.450_{-0.480}^{-0.420} \cdot I^{1.701_{1.615}^{1.787}}}\right)^{-1}$$

	k	AIC	BIC	MSE	KLD
PL1	10	-19416.86	-19362.86	$1.866 \cdot 10^{-3}$	0.0037
PL2	10	-19084.36	-19203.68.	$2.077 \cdot 10^{-3}$	0.0802
NPL	11	-19719.08	-19653.64	$1.691 \cdot 10^{-3}$	0.0219

Table VII: Statistical estimators used to qualify the scaling reported in Table VI.

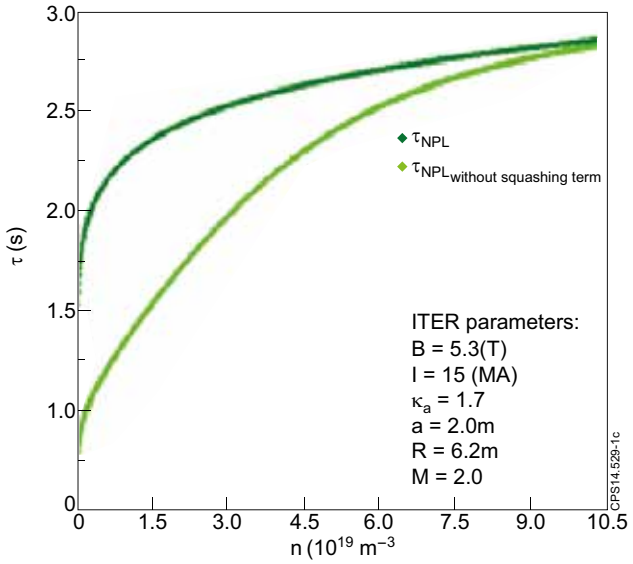

Figure 1: Comparison of the NPL scaling law behaviour with the plasma density at the parameters of ITER with or without the squashing term.