*35th EPS Conference on Plasma Phys. Hersonissos, 9 - 13 June 2008 ECA Vol.**32D**, P-2.113 (2008)*

Status of the International Stellarator/Heliotron Profile Database

A. Kus[1], D. Pretty[2], E. Ascasibar[3], C.D. Beidler[1], B. D. Blackwell[2], R. Brakel[1], R. Burhenn[1], F. Castejon[3], A. Dinklage[1], T. Estrada[3], Y. Feng[1], A. Fujisawa[4], H. Funaba[4], J. Geiger[1], J.H. Harris[2,5], C. Hidalgo[3], K. Ida[4], M. Kobayashi[4], R. König[1], G. Kühner[1], D. Lopez Bruna[3], H. Maaßberg[1], K. McCarthy[3], D. Mikkelsen[6], T. Minami[4], T. Mizuuchi[7], S. Murakami[7], N. Nakajima[4], S. Okamura[4], R. Preuss[1], S. Sakakibara[4], F. Sano[7], F. Sardei[1], T. Shimozuma[4], U. Stroth[8], Y. Suzuki[4], Y. Takeiri[4], J. Talmadge[9], H. Thomsen[1], Yu. A. Turkin[1], J. Vega[3], K.Y. Watanabe[4], A. Weller[1], A. Werner[1], R. Wolf[1], H. Yamada[4], M. Yokoyama[4]

[1] *Max-Planck-Institut für Plasmaphysik, Euratom Association, Greifswald, Germany*

[2] *Australian National University, Canberra, Australia*

[3] *Laboratorio Nacional de Fusión, Asociación Euratom/CIEMAT, Madrid, Spain*

[4] *National Institute for Fusion Science, Toki, Japan*

[5] *Oak Ridge National Laboratory, Oak Ridge, USA*

[6] *Princeton Plasma Physics Laboratory, Princeton, USA*

[7] *Kyoto University, Kyoto, Japan*

[8] *Universität Stuttgart, Stuttgart, Germany*

[9] *University of Wisconsin, Madison, USA*

Introduction

The International Stellarator/Heliotron Profile Database, ISHPDB, jointly hosted by IPP and NIFS [1] aims at an inter-machine comparison of energy confinement and transport in 3D devices. ISHPDB is an extension of the former International Stellarator Confinement Database established in 1994, which was the basis for the first widely acknowledged ISS95 energy confinement scaling [2]. In its current version a total of 4049 observations from nine devices are comprised. The international cooperation addresses issues of energy confinement scalings, core electron root confinement, edge physics, high beta, high performance, impurity transport, and density limit. The most advanced issue is the assessment of the energy confinement time [3].

To address a specific data analysis problem in a such large data collection sophisticated statistical methods such as classification, nonlinear regression, Bayesian inference, and data

mining are required. This paper focuses on some aspects of the energy confinement scaling derivation. Firstly, a scaling formula derived from the recent database version is presented. Secondly, some results of the cluster analysis are demonstrated.

Regression analysis of the new database

The use of meta-datasets such as ISHPDB is necessary for comparative studies, however difficulties arise when the data lacks homogeneity in the multidimensional space defined by the regression parameters. In the ISS95 scaling study [2] the problem of device-specific data subsets was handled by the introduction a new S parameter to the regression procedure in order to distinguish between devices with and without shear. The ISS04 scaling [4], based on a larger database, tackles this problem by splitting all data into several subgroups of devices, renormalizing using ISS95 as the reference scaling, weighting according to the number of observations in subgroups, and applying a standard regression with a collisional high beta constraint. Also, Bayesian inference methods have been applied for model comparison for different subgroups of devices [5].

The procedure used for the derivation of the ISS04 scaling formula has been applied to the new database release, ISS_DB07_22, to fit the model given in Eq. 1,

$$LOG\_TAU = a_0 + a_a LOG\_A + a_R LOG\_R + a_P LOG\_P + a_n LOG\_N + a_B LOG\_B + a_i LOG\_I \quad (1)$$

where a_0 is the intercept, and LOG_TAU, …, LOG_I are common logarithms of the confinement time, small and large plasma radii, absorbed power, density, magnetic field and iota, respectively. A subset of 2465 observations has been used for calculations. The defined subgroups are shown in Fig. 2, where items marked by "x_" denote the newest data. A regression on this dataset conforms well with the ISS04 scaling, see Fig. 3.

Cluster analysis

Cluster analysis is a statistical technique for identyfing observations with similar properties, contained in a (large) data collection [6]. Clustering is a type of *classification method* and is often applied in *data mining* [7]. In which space the cluster analysis is performed depends on the objective of a study. In this work we are interested in regression aspects of the database structure concerning energy confinement time scaling, hence the space spanned by the regression variables in Eq. 1 has been investigated.

There are two main methods of clustering: the *hierarchical* method for smaller datasets and *k-means* algorithm that is used for large data collections. Both approaches have their

advantages and disadvantages. Here we are using a mixed method, proposed in [8], first starting with the k-means method to identify a reasonable number of pre-clusters, and then applying the hierarchical procedure to obtain the final clusters. The results of the cluster analysis also depend strongly on the definition of the distances between observations and between clusters. In the present analysis the Euclidean and Ward distances have been used as a measures for the closeness between two observations and two clusters, respectively. The Ward's method minimizes the sum of squares of any two possible clusters that can be formed at each step. After several trials, cf. complementary material available in [1], k=20 pre-clusters have been used in the present analysis. Fig. 1 illustrates the final hierarchical clustering. The *dendrogram* shows how the single cluster are created, starting with 20 pre-clusters until all data are in one single final cluster. The curve underneath the dendrogram increases relatively evenly at the beginning of the clustering process. One can observe significant breaks in the distance between six and five and three and two clusters. These breaks suggest the number of the final clusters. A general algorithm to define the right number of clusters does not exist. By choosing three clusters all subgroups defined for the regression analysis are assigned to exactly one cluster. Between six and four clusters only ATF, HELE, and W7AS-low-iota subgroups are divided into different clusters, while all other subgroups remain stable in the same clusters. This fact can indicate that the subgroups definition may be further refined. In case of W7AS-low-iota only nine of 319 observation tend toward another cluster as the remaining 310. Fig. 2 shows how the subgroups defined for regression are assigned to the single clusters. Fig. 3 demonstrates a comparison of some fits where division of data by subgroups has been replaced by a division by clusters. Naturally, one cannot expect the same results for all cases. However, this

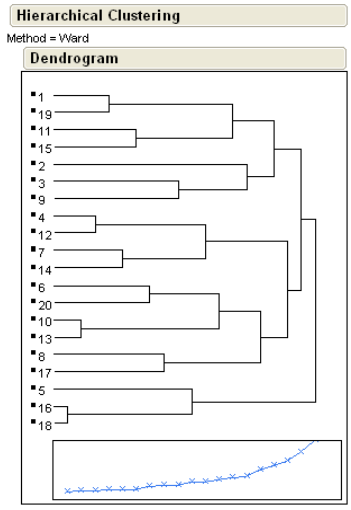

Figure 1. *The dendrogram with a scree plot presenting distances between clusters.*

Partitioning of Devices per Cluster

No.	Device	Nobs	C1	C2	C3
1	ATF	229	105	0	124
2	CHS	196	196	0	0
3	HELE	120	46	0	74
4	HELJ	54	54	0	0
5	LHD in	67	0	67	0
6	LHD inw.obl.	26	0	26	0
7	LHD inw.prol.	17	0	17	0
8	LHD out	16	0	16	0
9	LHD std	36	0	36	0
10	TJ-II	316	316	0	0
11	W7-A	13	0	0	13
12	W7-AS high beta	86	0	0	86
13	W7-AS high iota	124	0	0	124
14	W7-AS low iota	319	0	0	319
15	x_ LHD transp. data, g=1.22	221	0	221	0
16	x_ LHD transp. data, g=1.25	173	0	173	0
17	x_ W7-AS density scan	8	0	0	8
18	x_ W7-AS HDH attached	30	0	0	30
19	x_ W7-AS HDH detached	16	0	0	16
20	x_ W7-AS HDH normal confinement	12	0	0	12
21	x_ W7-AS iota scan	74	0	0	74
22	x_ W7-AS Preuss high beta /high iota	133	0	0	133
23	x_ W7-AS Preuss high beta /low iota	179	0	0	179
		2465	717	556	1192

Figure 2. *Subgroups assigned to the single clusters.*

diagram and Fig. 2 show that one can start with a division of data by clusters first, and then define subgroups according physical reasons.

Regression results (ISS04 scaling procedure)

Case	RMSE	a0	log a0	log a0 E	aa	aa E	aR	aR E	aP	aP E	an	an E	aB	aB E	ai	ai E
1	0.0267	0.134	-0.87	0.02	2.28	0.02	0.64	0.02	-0.61	.	0.54	0.01	0.84	0.01	0.41	0.01
2	0.0251	0.126	-0.90	0.01	2.28	0.02	0.66	0.01	-0.62	.	0.54	0.01	0.86	0.01	0.41	0.01
3	0.0218	0.080	-1.10	0.03	2.28	0.04	0.89	0.04	-0.57	.	0.55	0.01	1.01	0.02	0.14	0.04
4	0.0233	0.132	-0.88	0.02	2.32	0.03	0.52	0.02	-0.60	.	0.49	0.01	0.84	0.01	0.02	0.02
5	0.0255	0.125	-0.90	0.02	2.29	0.03	0.54	0.02	-0.58	.	0.52	0.01	0.78	0.01	0.09	0.02
6	0.0268	0.101	-0.99	0.02	2.16	0.03	0.71	0.02	-0.58	.	0.50	0.01	0.84	0.01	0.24	0.02
7	0.0239	0.133	-0.88	0.01	2.34	0.02	0.64	0.02	-0.64	.	0.55	0.01	0.89	0.01	0.30	0.01

Regression results, continued

Case	Dataset	Nobs	arho	abeta	anu	aiota	aeps
1	ISS_DB07_19 (ISS04 dataset)	1721	-0.79	-0.19	0.00	1.06	-0.07
2	ISS_DB07_22	2465	-0.83	-0.21	0.01	1.10	-0.09
3	Recently added data ('x_' subgroups)	846	-1.58	0.12	-0.17	0.14	-1.18
4	ISS_DB07_22 grouped into 3 clusters	2465	-0.62	-0.24	-0.02	0.02	0.15
5	ISS_DB07_22 grouped into 4 clusters	2465	-0.62	-0.12	-0.03	0.18	0.02
6	ISS_DB07_22 grouped into 5 clusters	2465	-0.76	-0.11	-0.06	0.51	-0.48
7	ISS_DB07_22 grouped into 6 clusters	2465	-0.93	-0.25	0.02	0.85	0.02

Figure 3. *ISS04 scaling procedure applied to different datasets. Columns with suffix E denote errors in estimated parameters. RMSE is the estimate of the error standard deviation. In the right five columns of the lower part the corresponding dimensionless variables are listed, see ref. [4].*

Summary and conclusions

Regression analysis of the newest ISHPDB database version agrees well with the ISS04 scaling. The performed cluster analysis shows the existence of significant cohesive subgroups of data and thereby a necessity for grouping data in the scaling procedures. The gathered experiences can be useful for further ISHPDB studies (not only scalings), the more so as ISHPDB has not been prepared using a statistically designed experiment, but combined solely according to physical considerations. Further database analyses should also involve collinearity checks and data mining techniques.

References

[1] http://www.ipp.mpg.de/ISS, and http://iscdb.nifs.ac.jp
[2] U. Stroth, et al., Nucl. Fusion 36, 1063 (1996)
[3] A. Dinklage, et al., Nucl. Fusion 47, 1265 (2007)
[4] H. Yamada, et al., Nucl. Fusion 45, 1684 (2005)
[5] R. Preuss, et al., PRL 99, 245001 (2007)
[6] A. Kus, et al., AIP Conf. Proc. 993, 47 (2008)
[7] D. Pretty, et al., A data mining algorithm for automated characterisation of fluctuations in multichannel timeseries, submitted to Comp. Phys. Comm. (2008)
[8] http://www.jmp.com/about/newsletters/jmpercable/pdf/17_spring_2005.pdf (2005)