Singularization of data subgroups in the International Stellarator-Heliotron Confinement Database

A. Kus[1], A. Dinklage[1], E. Ascasibar[2], C.D. Beidler[1], T. Estrada[2], H. Funaba[3], J. Geiger[1], J.H. Harris[4,5], C. Hidalgo[2], D. Lopez-Bruna[2], A. Lopez-Fraguas[2], H. Maaßberg[1], T. Minami[6], T. Mizuuchi[6], S. Murakami[6], N. Nakajima[3], S. Okamura[3], D. Pretty[4], F. Sano[6], Y. Suzuki[3], Y. Takeiri[3], J. Talmadge[7], K. Thomsen[8], Yu. A. Turkin[1], K.Y. Watanabe[3], R. Wolf[1], Y. Xu[9], H. Yamada[3], M. Yokoyama[3]

[1]Max-Planck-Institut für Plasmaphysik, Euratom Assoc., Greifswald, Germany, [2]Laboratorio Nacional de Fusión, Asociación Euratom/Ciemat, Madrid, Spain, [3]National Institute for Fusion Science, Toki, Japan, [4]Australian National University, Canberra, Australia, [5]ORNL, Oak Ridge, USA, [6]Kyoto University, Kyoto, Japan, [7]University of Wisconsin, Madison, USA, [8]European Commission, DG Research & Innovation, Unit K6, Brussels, Belgium, [9]Southwestern Institute of Physics, ChengDu City, China

Introduction

The ISS04 scaling [1], based on the International Stellarator-Heliotron Confinement Database (ISHCDB) [2], provides a reference for the energy confinement τ_E of the existing stellarator-heliotron devices in figures of volume averaged *engineering variables* a, R, P, n, B, iota (standing respectively for minor and major plasma radii, absorbed power, density, magnetic field and rotational transform at r/a=2/3). An assessment of the energy confinement data for future performance prediction, such as a unified scaling law, is addressed in this paper.

Here, we focus on the principles of *similarity* and *scale invariance* of confinement [3,4,5], and, as a consequence, work with dimensionless variables. Such variables also establish a connection to the transport theory, while regression models calculated in dimensional (*engineering*) variables are not theoretically justified. A transformation between both kinds of variables is possible [6]. If the regression is dimensionally correct, the scaling law can be fully expressed in both dimensionless and engineering variables.

In a first step, the statistical methods of cluster and discriminant analysis [7,8] are applied to check whether there exist any natural subgroups in the data (which scale possibly differently) and to identify variables causing the cluster formation.

Based on the extrapolation of the available confinement data it is to be defined whether a unified scaling is possible at all. If so, the correct set of scaling parameters and whether the current ISHCDB dataset is sufficient for the required analyses needs to be figured out.

Cluster analysis in different spaces

In this paper the ISHCDB_26 dataset [2, *Confinement Data*] has been analyzed and we use the nomenclature as documented in [2]. A multicollinearity check was done as the first step in each analyses, sometimes leading to limitations in the number of variables used in the analyses. During the calculations a couple of very conspicuous datapoints have been identified and temporarily hidden for later checks. Another important issue are errors in variables (δB, δW, …) [9] which could not be implemented in the analyses, but needs to be incorporated in future studies.

There exist many different sets of possible dimensionless variables, see [5]. In this work three groups of variables are used. The first group consists of the common {ρ*, ν*, β} [3], denoted here as RHOSTAR, NUSTAR, BETA. The second category are called *dimensionless engineering* variables [10]

$$B^* \sim B_t\, a^{5/4},\quad P^* \sim P_{heat}\, a^{3/4},\quad n^* \sim n\, a^{3/4} B_t^{-1},$$

denoted as BSTAR, PSTAR, NSTAR, which can be viewed as dimensionless if we ignore the dimensions of natural constants including the mass of the ion. These variables can be transformed to {RHOSTAR, NUSTAR, BETA}. The advantage of using these variables is that they can be specified prior to the operation of a device from envisaged settings, while {RHOSTAR, NUSTAR, BETA} are determined only thereafter from measured parameters.

The third group form again three variables, but derived from the so-called Connor-Taylor constraints [4]. They may be incorporated in the power law ansatz:

$$CT_1 = \frac{P}{n\, a^4 B^3},\quad CT_2 = \frac{a^3 B^4}{n},\quad CT_3 = \frac{1}{n\, a^2},\qquad W \sim (CT)_1^{x1}\, (CT_2)^{x2}\, (CT_3)^{x3}.$$

For these three spaces, cluster analysis has been performed. To check which variables are significant for clustering, discrimination analysis was done on clustered data. Some results are shown in Tab. 1. For example (all data), five clusters may be identified in the space {TAU, RHOSTAR, NUSTAR, BETA}, there is no serious multicollinearites between the four used variables, the main factors causing discrimination between different devices are BETA and TAU. The number 65 is the percentage of the data points correctly assigned to the original clusters (after the discriminant model was used for prediction as a standard check of the model). A value like 65 suggest a relatively poor discrimination caused here by some overlaps between assessed clusters. The last column shows the discrimination of the high- and low-beta data. E.g., in the space spanned by {TAU, BSTAR, PSTAR, NSTAR} in the LHD subset, there exist two clusters, and mainly the BSTAR is responsible for data splitting with regards to high/low beta reflecting the difference of operational magnetic field as expected.

An important point is to know, which data belong to the single clusters. For example, in the space {TAU, RHOSTAR, NUSTAR, BETA} all data are divided into five clusters. One cluster consists exclusively of LHD high-beta data, another cluster contains all kind of LHD data (also high-beta), a different cluster exclusively consists of W7-AS data (mostly high-beta), and the remaining clusters contain data from several devices (also LHD and W7-AS). It should be emphasized here again that all these clusters overlap in the analyzed space. It is notable that this statistical analysis shows these differences in the confinement data which are related to physically interpretable differences.

Dataset	Nobs	Space	Correlations	nCL	Device discriminator	Beta discriminator
all Data	3375	{TAU, RHOSTAR, NUSTAR, BETA}	---	5	BETA, TAU (65)	---
LHD	1362	{TAU, RHOSTAR, NUSTAR, BETA}	RHOSTAR, BETA	2	---	RHOSTAR, TAU (85)
TJ-II	318	{TAU, RHOSTAR, NUSTAR, BETA}	---	4	---	---
W7AS	1098	{TAU, RHOSTAR, NUSTAR, BETA}	---	4	---	BETA (99)
	•			•		
all Data	3446	{TAU, CT1, CT2, CT3}	---	3	TAU, CT1, CT3 (63)	
LHD	1362	{TAU, CT1, CT2, CT3}	---	5	---	CT2 (93)
TJ-II	318	{TAU, CT1, CT2, CT3}	CT2, CT3	3	---	---
W7AS	1098	{TAU, CT1, CT2, CT3}		7	---	CT3 (96)
	•			•		
all Data	3446	{TAU, BSTAR, PSTAR, NSTAR}	---	3	PSTAR, TAU (71)	---
LHD	1362	{TAU, BSTAR, PSTAR, NSTAR}	---	2	---	BSTAR (93)
TJ-II	318	{TAU, BSTAR, PSTAR, NSTAR}	---	4	---	---
W7AS	1083	{TAU, BSTAR, PSTAR, NSTAR}	PSTAR, NSTAR	4	---	NSTAR, PSTAR (99)

Table 1. *Clusters identified in different spaces. Column nCL contains the number of recognized clusters.*

In the next step, dimensionally correct regression analysis for each cluster has been performed. To compare with ISS04, the same variables were used as well as the collisional high-beta constraint. But, unlike the ISS04 scaling procedure, there was neither renormalization nor weighting. Results are presented in Tab. 2. In the preliminary analysis, in clusters dominated by one device regression coefficients for both minor and major radii were fixed to the ISS04 values. In general, single clusters scale differently from the ISS04 as well as differently among each other. But, except for the iota coefficient, the ISS04 trend remains.

Summary

Clusters of data can be identified in different spaces defined by dimensionless variables and the clustering is found to depend on the choice of variables. Clusters in single devices are more distinct than in all the data together and uncertainties in the cluster identification have significant consequences in the subsequent analyses.

In the dataset containing all data, the clusters of data scale differently. The coefficients for absorbed power, plasma density and magnetic field are partly in accordance with the ISS04 scaling. By contrast, the rational transform coefficient shows a large variation, which may suggests that this parameter should be removed from the model.

Case	Dataset	Nobs	RMSE	aa	ar	ap	an	ab	ai
1	ISS04 scaling	1721	0.0267	2.28	0.64	-0.61	0.54	0.84	0.41
2	ISHCDB_26xg_allData_DIML_c_5_0_CL_1	894	0.0782	2.27	1.16	-0.73	0.70	1.06	-0.20
3	ISHCDB_26xg_allData_DIML_c_5_0_CL_2	669	0.1099	2.30	1.17	-0.64	0.62	1.16	-0.17
4	ISHCDB_26xg_allData_DIML_c_5_0_CL_3	342	0.0808	•	•	-0.64	0.50	0.92	1.60
5	ISHCDB_26xg_allData_DIML_c_5_0_CL_4	533	0.0663	•	•	-0.49	0.50	0.83	0.00
6	ISHCDB_26xg_allData_DIML_c_5_0_CL_5	937	0.0503	•	•	-0.79	0.45	1.09	0.85
7	ISHCDB_26xg_allData_CT_c_3_0_CL_1	2819	0.1120	2.21	1.02	-0.44	0.52	1.01	-0.07
8	ISHCDB_26xg_allData_CT_c_3_0_CL_2	501	0.0815	2.08	1.01	-0.58	0.75	0.63	-0.21
9	ISHCDB_26xg_allData_CT_c_3_0_CL_3	126	0.1032	•	•	-0.48	0.36	1.05	3.27
10	ISHCDB_26xg_allData_KL_c_3_0_CL_1	2013	0.0978	1.78	1.20	-0.61	0.58	0.83	-0.21
11	ISHCDB_26xg_allData_KL_c_3_0_CL_2	682	0.0126	•	•	-0.62	0.48	0.94	2.26
12	ISHCDB_26xg_allData_KL_c_3_0_CL_3	751	0.0500	•	•	-0.77	0.50	1.00	0.54

Table 2. *Regressions performed on single clusters identified in different spaces . All available data are used. Case1 shows the ISS04 scaling coefficients, cases 2-6 denote clusters in {RHOSTAR, NUSTAR, BETA}, cases 7-9 clusters in {CT1, CT2,CT2}, cases 10-12 clusters in {BSTAR, PSTAR, NSTAR}.*

The results of this study indicate the ISHCDB dataset to be statistically inhomogeneous. The ad-hoc separation in subsets as for the ISS04 scaling is not reflected by statistical cluster analysis for different formulations of dimensionless variables. Therefore, a justification for a unified scaling law based on similarity arguments cannot be drawn on the existent database. As a next step, a specification of parameters allowing scalings with predictive qualities needs to be done.

References

[1] H. Yamada et al. Nucl. Fusion 45, 1684 (2005)

[2] https://ishpdb.ipp-hgw.mpg.de/ISS/public/

[3] B.B. Kadomtsev, Sov. J. Plasma Phys. 1, 295 (1975)

[4] J.W. Connor and J.B. Taylor, Nucl. Fusion 17, 1047 (1977)

[5] C.C.Petty, Phys. Plasmas 15, 080501 (2008)

[6] T.C. Luce, et al., PPCF 50, 043001 (2008)

[7] A. Kus, et al., 38th EPS, Strasbourg (2011), http://ocs.ciemat.es/EPS2011PAP/pdf/P4.120.pdf

[8] K.V. Mardia, J.T.Kent, J.M. Bibby, *Multivariate Analysis*, Academic Press, London (1979)

[9] Rawlings, J.O., *Applied Regression Analysis*, Wadsworth &Brooks/Cole (1988)

[10] K. Lackner, Fusion Sci. Technol. 54, 4, 989 (2008)