Cluster Analysis of the International Stellarator Confinement Database

A. Kus [a], A. Dinklage [a], R. Preuss [a], E. Ascasibar [b], J. H. Harris [c,d], S. Okamura [e], , F. Sano [f], U. Stroth [g], J. Talmadge [h], H. Yamada [e]

[a] *Max-Planck-Institut für Plasmaphysik, EURATOM-IPPAssociation, Greifswald, Germany*
[b] *Laboratorio Nacional de Fusión, EURATOM-CIEMAT Association, Madrid, Spain*
[c] *Oak Ridge National Laboratory, Oak Ridge, USA*
[d] *Australian National University, Canberra, Australia*
[e] *National Institute for Fusion Science, Toki, Japan*
[f] *Kyoto University, Kyoto, Japan*
[g] *University of Stuttgart, Stuttgart, Germany*
[h] *University of Wisconsin, Madison, USA*

Abstract. Heterogeneous structure of collected data is one of the problems that occur during derivation of scalings for energy confinement time, and whose analysis tourns out to be wide and complicated matter. The International Stellarator Confinement Database [1], shortly ISCDB, comprises in its latest version 21 a total of 3647 observations from 8 experimental devices, 2067 therefrom beeing so far completed for upcoming analyses. For confinement scaling studies 1933 observation were chosen as the standard dataset. Here we describe a statistical method of cluster analysis for identification of possible cohesive substructures in ISDCB and present some preliminary results.

Keywords: ISCDB, ISS95, ISS04, cluster analysis, hierarchical clustering.
PACS: 52.55.HC, 52.55.Dy.

INTRODUCTION

Inhomogeneous statistical population, poor dispersion of data in the multidimensional space spanned by regression variables, and correlation between these variables belong to the most important detrimental effects in the least squares regression analysis [2]. During analyses for derivation of existing ISS scalings all these problems were revealed and outlined in [3, 4, 5].

In particular the problem of subgroups was already recognized and handled in the ISS95 scaling [3] by introduction of a new *S* parameter to distinguish between stellarators with/without shear, and in the ISS04 scaling [4] by defining subgroups of devices and using renormalization.

Cluster structures if existing in the space spanned by regression variables, more precisely if existing as subsets in the mulitdimensional ellipsoid built by observed data, may strongly affect the regression results. Here, we make an attempt to use a statistical exploratory technique called *cluster analysis* to discover possible clumping structures in the collected data.

Complementary material available in [1] contains some tables with parameter distributions, single and multivariate correlations (including collinearity checks to detect possible multivariate correlations that cannot be discovered when analysing only pairwise correlations), and details of performed cluster analysis.

CLUSTER ANALYSIS

Clustering [6] is a technique of grouping rows together that share similar values across a number of variables. This procedure simply discovers structures in data without explaining why they exist. For our purposes a method called *hierarchical* has been applied, using the *JMP* statistical package [7, 8].

The Principle of Hierachical Clustering

Hierarchical clustering is the most straightforward method of clustring. Its principle is the following. We start with $x_1,\ldots,x_n$ observations of *p* variables (using ISS scaling parameters *a*, *R*, *P*, n_e, *B*, *iota*, is *p*=6) and consider each observation as its own cluster. Our investigations aim at identification of N_c homogeneous groups (clusters) of observations, that means the observations in each cluster are all close to each other. At every of the subsequent steps the distance between each cluster is calculated and two closest clusters are combined together. The process continues until all the observations are in a single final cluster.

One problem here is to choose a correct definition of the distance between two observations. The most commonly used type is the Euclidean distance between standardized data, but other definitions may be used as well [9]. Another question is to define a proper rule to determine the closeness between clusters, so we need a *linkage* formula to determine when two clusters are sufficiently similar to be linked together. Refer to [9] for short overiew of possible solutions. The present work uses the Euclidean formula for calculations of distances between variables and the *Ward's* rule for linking clusters. The Ward's rule minimizes the sum of squares of any two possible clusters that can be formed at each step. Usually the clustering process is illustrated graphically in form of a *dendrogram* together with a curve that represents the increase in the distances between new composed clusters. From this curve and the clustering history supplied by the software one can determine a reasonable number of end-clusters. A theoretical way to determine the right number of clusters does not exist.

For large data sets, hovewer, hierarchical clustering is not practical due to the huge amount of memory required to store the distance matrix used in finding the clusters. Therefore other, faster, methods, like *k-means* clustering (starting with a predetermined number of clusters, k) are recommended for future analyses.

Analysis of ISCDB Version 21

The choice of clustering variables depends on the objectives of a study. From the regression analysis point of view it is important to descry which parameters are primarily responsible for clustering. Such a study is presented in [10], even though on other terms and conditions. The concept, adapted for our purposes, is as follows. Let us assume we pay attention to the variable LOG_V. First , all observations are divided into *nCL* clusters (on the basis of previous studies using dendrogram and further checks). Then, the ISS regression model, Eq. 1,

$$LOG\_TAU = a_0 + a_a\, LOG\_A + a_R\, LOG\_R + a_P\, LOG\_P + a_n\, LOG\_N\ n + a_B\, LOG\_B + a_\iota\, LOG\_I \qquad (1)$$

is fitted, for *i*=1, …, *nCL*, using only observations contained in the *i*th cluster in each case. (In the formula above a_0 is the intercept, and LOG_TAU,…, LOG_I are common logarithms of energy confinement time, small and large plasma radii, absorbed power, density, magnetic field and iota, respectively). Regressions on some of the clusters, say on *r* clusters, are expected to yield better fits than a regression on the all data.

The ratio $p_c=r/nCL$ allows an information about the importance of LOG_V for clustering. For example p_c>>0.5 means that there exist significant subgroups in ISCDB, primarily determined by LOG_V, as the most fits on clusters are qualitative much better than the fit on all data. As a measure of the goodness of fit the *R2* parameter is used. The value of R2 says how much variation in the response variable is represented by the used model. An R2 value of 0.94 as obtained from regression on ISCDB standard subset without clustering, cf. table 1, is relatively very high – it states that 94 precent of variation is caused by the model and only 6 perecent are randomly variations. For comparison: in [10] the reference value of 0.29 was used.

Clustering investigations on ISCDB were not confined to single variables only, but also impacts of more variables, including crossproducts, on the clustering were tested. A review of regression analysis results shows that for both the standard subset and its extension primarily LOG_P builds two and three clusters, and the cross product of LOG_P and LOG_R forms two clusters (see the complementary material in [1] for details). The highest values of R2 results for two clusters and LOG_P as the clustering variable, so we present here results only for this case.

For the analysis two subsets of ISCDB database version 21 have been used, the standard subset and the standard subset extended by 144 observations from W7-AS experiment. Results of regression analyses for two clusters caused by LOG_P are shown in table 1. In comparison with ISS95 and ISS04 scalings the greatest differences are in ai, aB, and aR. An ultimate statement about the impact of described clustering cannot be made in this preliminary work. The results may be affected by the high R2 reference value (in other words, by the high grade of the ISS fit quality). In the further studies one will have to try make checks probably also with smaller R2 values, may be 0.8 or something like this.

Interesting is the question which devices belong to the individual clusters. Figure 1 presents distributions by some predefined device subgroups, where in the both cases the same grouping formulas have been used. In the further clustering studies one will have to go a step lower and analyze single observations in the clusters.

ISS_DB07_21_stdset_gg_c2

Device	NOBS	C1	C2
	.	.	.
ISS_DB07_21_stdset	1933	1027	906
	.	.	.
ATF	229	130	99
CHS	196	125	71
HELE	120	80	40
HELJ	54	0	54
LHD	162	162	0
TJ-II	316	0	316
W7-A	13	0	13
W7-AS	843	530	313
	.	.	.
ATF	229	130	99
CHS	196	125	71
HELE	120	80	40
HELJ	54	0	54
LHD inw.obl.	26	26	0
LHD inw.prol.	17	17	0
LHD-in	67	67	0
LHD-out	16	16	0
LHD-STD	36	36	0
TJ-II	316	0	316
W7-A	13	0	13
W7-AS iota<0.48	554	348	206
W7-AS iota>=0.48	289	182	107
	.	.	.
ECRH	761	26	735
mixed	137	108	29
NBI	1033	893	140
	.	.	.
W7-AS ECRH	263	23	240
W7-AS mixed	85	84	1
W7-AS NBI	493	423	70

a)

ISS_DB07_21_allData_gg_c2

Device	NOBS	C1	C2
	.	.	.
ISS_DB07_21_allData	2067	1085	982
	.	.	.
ATF	229	130	99
CHS	196	125	71
HELE	120	80	40
HELJ	54	0	54
HSX	0	0	0
ITER	0	0	0
LHD	162	162	0
TJ-II	316	0	316
W7-A	13	0	13
W7-AS	977	588	389
	.	.	.
ATF	229	130	99
CHS	196	125	71
HELE	120	80	40
HELJ	54	0	54
LHD in	67	67	0
LHD inw.obl.	26	26	0
LHD inw.prol.	17	17	0
LHD out	16	16	0
LHD std	36	36	0
TJ-II	316	0	316
W7-A	13	0	13
W7-AS iota<0.48	603	348	255
W7-AS iota>=0.48	374	240	134
	.	.	.
ECRH	835	26	809
mixed	137	108	29
NBI	1093	951	142
	.	.	.
W7-AS ECRH	337	23	314
W7-AS mixed	85	84	1
W7-AS NBI	553	481	72

b)

FIGURE 1. Device distributions per cluster for the standard subset (a) and its extension (b) in various subgroups. The upper part contains single stellarators, the next part contains subgroups as defined for derivation of ISS04 scaling. The following two subsequent segments show distribution of heating groups.

TABLE 1. Regression coefficients obtained by fitting several datasets. Data are divided in two clusters with LOG_P as the clustering variable. Nobs shows the number of observations used in the respective regression analysis. R2 values for ISS95 and ISS04 were not published as in both cases nonlinear fitting procedures (not calculating R2) were used. RMSE is the estimate of the error standard deviation.

Dataset	Nobs	R2	a0	aa	aR	aP	an	aB	ai	rmse
ISS95	812	-	0.08	2.21	0.65	-0.59	0.51	0.53	0.40	0.0910
ISS04	1721	-	0.13	2.28	0.64	-0.61	0.54	0.84	0.41	0.2366
ISS_DB07_21 all Data	2067	0.93	0.03	2.23	1.14	-0.72	0.69	0.98	-0.16	0.1192
ISS_DB07_21 all Data, cluster 1	1085	0.95	0.02	1.97	1.02	-0.44	0.52	1.07	-0.30	0.0979
ISS_DB07_21 all Data, cluster 2	982	0.94	0.02	2.05	0.91	-0.89	0.79	0.79	-0.17	0.1073
ISS_DB07_21 stdset	1933	0.94	0.03	2.22	1.38	-0.72	0.68	1.02	-0.19	0.1168
ISS_DB07_21 stdset, cluster 1	1027	0.96	0.02	1.94	1.03	-0.42	0.49	1.03	-0.33	0.0954
ISS_DB07_21 stdset, cluster 2	906	0.95	0.02	2.06	0.89	-0.88	0.77	0.88	-0.18	0.1070

CONCLUSION

A preliminary cluster analysis of the International Stellarator Confinement Database version 21 have shown that there exist cohesive structures in the collected data with the LOG_P having the primary meaning for the clustering.

Furher cluster analysis, in connection with collinearity studies, is necessary. By reason of increasing data in ISCDB (a new version 22 with currently 4913 observations is already in preparation) the use of k-means clustering method is recommended. Continuative analyses are indispensable as the data collected in ISCDB have not been prepared using a statistically designed experiment, but combined solely according to physical considerations.

REFERENCES

1. http://www.ipp.mpg.de/ISS and http://iscdb.nifs.ac.jp.
2. O. Kardaun and A. Kus, Basic probability theory and statistics for experimental plasma physics,Technical Report IPP5/68, Max-Planck-Institut für Plasmaphysik, Garching (1996).
3. U. Stroth, et al., Nucl. Fusion **36** 1063 (1996).
4. H. Yamada, et al., , Nucl. Fusion **45** 1684 (2005).
5. A. Dinklage, et al., Physical model assessment of the energy confinement time scaling in stellarators, *Nucl. Fusion*, **47**, 1265 (2007).
6. K.V. Mardia, J.T. Kent, and J.M. Bibby, *Multivariate Analysis*, Academic Press, London (1979).
7. http://www.jmp.com.
8. JMP Statistics and Graphics Guide, Release 7, SAS Institute Inc., Cary, NC, USA (2007).
9 http://www.statsoft.com/textbook/stcluan.html
10. B. Bishop and G. Veerender, Forecasting water demand using cluster and regression analysis, Utah State University, http://www2.bren.ucsb.edu/~keller/energy-water/6-8%20Bruce%20Bishop%20& %20Veerender%20Garg.pdf (2007).