Identification of variables causing clustering in the global energy confinement data by use of discriminant analysis

A. Kus[1], A. Dinklage[1,] E. Ascasibar[2], C.D. Beidler[1], A.A. Beletskii[3], B.D. Blackwell[4], T. Estrada[2], H. Funaba[5], J. Geiger[1], J.H. Harris[4,6], C. Hidalgo[2], M. Hirsch[1], D. Lopez-Bruna[2], A. Lopez-Fraguas[2], H. Maaßberg[1], T. Minami[5], T. Mizuuchi[7], S. Murakami[7], N. Nakajima[5], S. Okamura[5], D. Pretty[4], M. Ramisch[8], S. Sakakibara[5], F. Sano[7], U. Stroth[1], Y. Suzuki[5], Y. Takeiri[5], J. Talmadge[9], K. Thomsen[10], Yu.A. Turkin[1], K.Y. Watanabe[5], A. Weller[1], R. Wolf[1], H. Yamada[5], M. Yokoyama[5]

[1]Max-Planck-Institut für Plasmaphysik, Euratom Assoc., Garching & Greifswald, Germany, [2]Laboratorio Nacional de Fusión, Asociación Euratom/CIEMAT, Madrid, Spain, [3]Institute of Plasma Physics, Kharkov, Ukraine, [4]Australian National University, Canberra, Australia, [5]National Institute for Fusion Science, Toki, Japan, [6]Oak Ridge National Laboratory, Oak Ridge, USA, [7]Kyoto University, Kyoto, Japan, [8]Universität Stuttgart, Stuttgart, Germany, [9]University of Wisconsin, Madison, USA, [10]European Commission, Brussels, Belgium

Motivation

Regression analyses of confinement databases may suffer from unrevealed dependencies on the energy confinement time τ_E. Missing regression parameters (e.g. due to a lack of experimental accessibility) or nominal dependencies (like the dependence on device/magnetic configuration [1,2] or physics regimes [3]) may become significant obstacles to the derivation of predictive scaling laws. These potential unknowns have a direct impact on the reliability of performance predictions for future fusion reactors.

This paper applies statistical techniques to address 1) if the set of regression variables is sufficient for data fitting, 2) the identification of nominal dependencies by cluster analysis, and 3) the assessment of the impact of regression variables. The latter issue is particularly interesting with regard to the question of which parameters are most important for predictions. Therefore such assessment may also provide feedback for proposals for parameter studies in running experiments.

The specific case investigated here is the ISS04 dataset. In the ISS04 scaling study [1], an essential step was the division of analyzed data into 14 subgroups inducing systematic offsets in the estimated confinement. Based on physical considerations, the allocation of data into subgroups was determined by devices and magnetic configurations. Later investigations of the internal database structure, using *cluster analysis* [4,5] showed the existence of some natural substructures (*clusters*) in the data that, in general, do not completely coincide with the ISS04 subgroups. However, when increasing the number of the clusters (say 12-15), one can observe a kind of saturation in the regression parameters towards the ISS04 scaling [4].

In general, a single cluster contains data from different devices (and also from different ISS04 subgroups). For further studies it is important to identify which parameters are responsible for the subgroup formation (cluster analysis recognizes and groups objects with similar properties in a dataset). The present paper uses the *discriminant function analysis* to identify the most important clustering parameters.

Basic ideas of discriminant function analysis

Discriminant function analysis [5] aims at determination of linear combinations of independent variables (*predictors*) that discriminate among the *categories* of the grouping variable. A single linear combination of predictor variables X_1, …, X_p, called a *discriminant function*, is constructed such that it assigns its values into two subgroups that differ as much as possible.

The maximal number of discriminant functions is equal to the number of subgroups minus one, or the number of predictor variables, whichever is smaller. In the most simple case, with two subgroups (e.g. L/H mode data), there exists only one discriminant function

$$D = b_1X_1 + b_2X_2 + \cdots + b_pX_p, \qquad \text{for data previously standardized.} \tag{1}$$

All discriminant functions are pairwise orthogonal (uncorrelated). Viewing the coefficients *b*'s one can see how the predictor variables contribute to the discrimination: the larger the *b* (in absolute value) the larger the contribution. The percentage, *pcp*, of the correctly predicted assignments (using the same data set as for model development) assesses the quality of the chosen discriminant model.

Figure 1 shows the 14 ISS04 subgroups in the plane spanned by the first two discriminant functions (denoted as *Canonical* 1/2). As predictors the set of the ISS04 *engineering* variables {LOG_TAU, LOG_A, LOG_R, LOG_PMW, LOG_N, LOG_B, LOG_IOTA} has been used. The names LOG_TAU, …, LOG_IOTA stand for logarithms of the confinement time, small and large plasma radii, absorbed power, density, magnetic field and iota, respectively.

Clusters in different subsets and dimensions

We analyze here two datasets: a) the current version ISHCDB_25 of the stellarator-heliotron confinement database [6] extended by 1200 new LHD high-beta observations presented in [7],

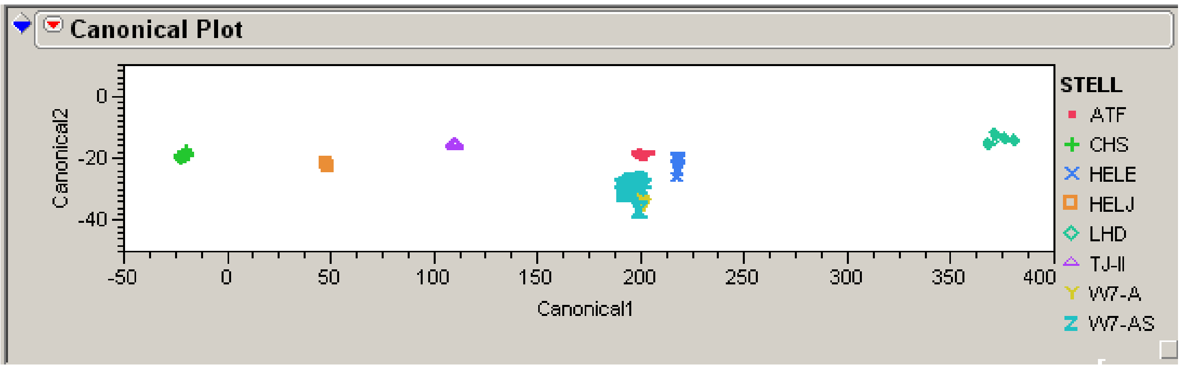

Figure 1. *ISS04 subgroups in the first two discriminant dimensions (in the upper right corner the five LHD subgroups can be seen). Discriminant analysis results in the function coefficients* D_1*= {2.6, -2.6, 658.6, 1.3, -1.4, 0.1, -1.7} and* D_2*= {-1.8, 30.0, -9.7, -0.4, -0.1, -0.3, 12.2}, with pcp=97.2. Hence, almost all the discrimination takes place along the horizontal axis, dominated by LOG_R.*

and b) the just mentioned dataset joined with the representative tokamak L-Mode subset of the ITER H-Mode database [8] as used in [9,10].

Table 1 gives an overview of cluster and discriminant analyses for different datasets in both engineering and dimensionless variables. The clusters in the present paper have been derived using the same method as in [4]. Cases 1-4 concern subsets of stellarator-heliotron data, while number 5 deals with the joined stellarator-heliotron and tokamak data. The column *ic#* indicates the possible number of main clusters suggested by the cluster analysis. Columns *Main predictors* and *pcp* represent the variables mainly causing clustering and the percentage of the correctly predicted assignments, respectively. Based on the *pcp* parameter, it is clearly visible that the ascertained clusters in the engineering variables are well discriminated, mainly by the geometrical parameters, while for clusters in the dimensionless variables no discriminant model can be determined satisfactorily (with the possible exception of case 2).

Some findings from other conducted analyses

In the W7-AS subset, high-beta data always form a separate cluster, in both engineering and dimensionless variables (the low pcp-value in Table 1, case 4, right-hand, is caused only by difficulties in separation inside of the non-high-beta group). High-beta data in both LHD and W7-AS subsets scale similar to each other but differently from ISS04 (in particular, with LOG_IOTA coefficient = -0.16 instead of +0.41).

Discriminant analysis has also been performed using fixed, defined subgroups (i.e. without automatic clustering in the prestage). In the case of high beta in engineering variables, for both LHD and W7-AS subsets, magnetic field and the injected power seem to be the significant determinators. In dimensionless variables no satisfactory model could be found (in the subset containing LHD and W7-AS data, there is the indication that RHOSTAR and BETA are significant predictors, with pcp=88).

The difference between stellarators-heliotrons and tokamak L-mode is observed in both engineering and dimensionless variables, and the tokamak data are clearly separated (pcp=99) with main separators LOG_R, LOG_A, LOG_IOTA, and TAU, RHOSTAR, BETA, resp.

Table 1. *Clusters in different datasets and dimensions.*

Discrimination		Engineering variables			Dimensionless variables		
Case	Dataset	ic#	Main predictors	pcp	ic#	Main predictors	pcp
1	all data	3	LOG_A, LOG_R	100	5	BETA, TAU	79
2	LHD + W7-AS	2	LOG_R	100	2	RHOSTAR, BETA	88
3	LHD	5	(*without LOG_A*, *LOG_R*) LOG_B, _PMW, _TAU, _N	94	5	BETA, TAU	57
4	W7-AS	4	(*without LOG_A*, *LOG_R*) LOG_B, _PMW, _IOTA	100	5	BETA, TAU, RHOSTAR	67
5	stell-hel-tok	6	LOG_A, _R	96	7	TAU, BETA, RHOSTAR	51

In summary, application of discriminant analysis in conjunction with cluster analysis on the ISS04 dataset identifies “anticipated” dependencies (geometry) indicating again the particular role of the scaling in size resulting due to the comparison of different devices. Because *a* and *R* separate ISS04 subgroups, a caveat must be raised for the predictive use of the scaling with regard to the geometrical parameters.

The applied technique is capable of separating major physics differences, e.g. low- and high-beta data, but so far, however, has failed to give a clear distinction in more elaborate groups such as shaping dependent subgroups. In a joint dataset, tokamak data are clearly separated from the stellarator-heliotron group. As expected [2], there is no joint tokamak-stellarator scaling. These findings confirm a previous conclusion, that the ISS scaling is a reference scaling rather than a predictive one, and underlines the necessity for 1-D predictive transport modelling.

This paper has been conducted within the International Stellarator-Heliotron Profile Database collaboration [6].

References

[1] H. Yamada, et al., Nucl. Fusion **45** 1684 (2005)
[2] A. Dinklage, et al., Nucl. Fusion **47** 1265 (2007)
[3] R. Preuss, et al., Phys. Rev. Lett. **99** 245001 (2007)
[4] A. Kus, et al., 35th EPS, Hersonissos, (2008)
[5] K.V. Mardia, J.T.Kent, J.M. Bibby, *Multivariate Analysis*, Academic Press, London (1979)
[6] http://www.ipp.mpg.de/ISHPDB/, http://ishpdb.nifs.ac.jp/ (2011)
[7] A. Weller, et al., Nucl. Fusion **49** (2009)
[8] http://efdasql.ipp.mpg.de/HmodePublic (2011)
[9] A. Dinklage, et al., Fusion Sci.Technol. **51** 1 (2007)
[10] A. Kus, et al., 25th EMS, Oslo, http://www.ipp.mpg.de/~kus/eps2011References/2005_Oslo (2005)