AN EXAMPLE FOR COMPLEMENTARITY BETWEEN PLASMA PHYSICS AND DATA-DRIVEN RESEARCH

M. YOKOYAMA and H. YAMAGUCHI
National Institute for Fusion Science, National Institutes of Natural Sciences,
The Graduate University for Advanced Studies, SOKENDAI
Toki, Gifu, Japan
Email: yokoyama@nifs.ac.jp

Abstract

An example for complementarity between plasma physic and data-driven research will be reported. It is the application of the information criterion (either Akaike or Bayesian) in the field of the statistics to the data obtained in the fusion research. A particular example described in the paper is the trials being conducted by utilizing the thermal diffusivity database in the Large Helical Device (LHD), Japan. The paper will reveal that the information criterion can be a powerful tool to unravel complicated entangled phenomena in fusion plasmas, from a viewpoint different from plasma physics. By efficiently extracting the information contained in the data, which could not always be achieved only by the variables of interest based on plasma physics (physicist's view), the convincingness supported by plasma physics and/or new discoveries and awareness from the perspective of the data-driven approach can be achieved. By considering complementarity between plasma physics and data-driven approach, fusion research should be qualitatively strengthened.

As a conventional way of thinking in fusion research, data is typically analysed/interpreted with the variables of interest based on plasma physics. For example, in the case of heat transport problems, variables relevant to plasma physics, such as temperature gradient length, temperature ratio, collision frequency, and so forth, are often used as the axis of the graphs. However, with this approach, the viewpoint is rather fixed to a small number of dimensions (two, [Y] vs [X], for plane graphs, or three ([Z] vs ([X] and [Y])), and it means that only a small portion of all the information contained in the data can be extracted. In other words, the information obtained with huge efforts (experimentally or numerically) cannot be "fully" utilized. Therefore, it has been tried to introduce a new perspective such as "selection of "statistically" important variables based on the information criterion". Such a clue has been already discovered in the paper by the authors [M. Yokoyama and H. Yamaguchi, Nuclear Fusion **60** (2020) 106024]. A relevant figure and table (modified from those shown in the NF2020 paper) is shown below for concrete explanation.

Akaike's information criterion with correction (AICc) for small sample sizes (in this case, 404 data) is employed for the thermal diffusivity database created by the integrated transport analysis suite, TASK3D-a [M. Yokoyama et al., Nuclear Fusion 57 (2017) 126016], utilizing so-called high ion temperature discharges in LHD. It should be noted that Bayesian information criterion (BIC) gives almost the same values as AICc for this small data size. Nine variables (plasma parameters and magnetic configurations) are prepared as plausible explanatory variables for the objective variable: the thermal diffusivity. The exhaustive search (all possible combinations) on AICc with utilizing these nine variables has been conducted, and the result is shown in Figure. 1. The bars for each number of variables (NV) indicates that AICc vary for different combinations with given NV. The AICc minimum decreases up to NV = 5, and then almost unchanged beyond. Thus, it is considered that the combination of five variables giving the minimum AICc corresponds to the "statistically" optimal model. This is different approach to obtain a relevant model based on statistical consideration, other than that based on plasma physics. Then, the appearance of variables and those exponents (in log-linear multivariate regression) up to NV = 5 are summarized in Table 1. It is surprising that physically convincing variables, such as the temperature gradient length and the temperature ratio, and even normalized ion Larmor radius, are "statistically" selected as an important variables, and furthermore, those exponents seem to be converged (in the regression). In this way, it can be possible to find "important" variables (dimension in organizing data) to describe and interpret available data.

By examining data from the perspectives of statistical/data-driven approach, it can be possible to extract much information from the available data. This is complementary approach to plasma physics, and then may create convincingness by plasma physics and even "data-oriented" new discoveries.

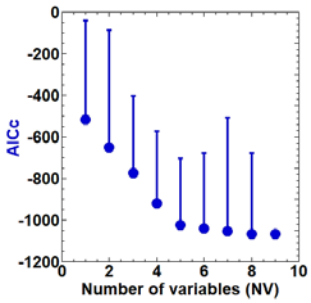

Figure 1. The evolution of AICc values for the ion thermal diffusivity as a function of number of variables ("NV").

NV	ρ_i^*	R/L_{Ti}	T_e/T_i	ϵ_h	ϵ_t	$\iota/(2\pi)$	AICc
1		-1.43					-516.6
2		-1.08			-1.00		-650.05
3		-0.92	0.55	-1.08			-772.83
4	3.64	-0.89	2.63		-1.06		-918.43
5	3.79	-0.84	2.88		-1.37	1.97	-1023.9

Table 1. Appearance of variables and those exponents as a function of NV up to NV = 5 (at the lowest AICc for each NV). Notations should be referred to NF2020 paper indicated in the text.