Data Driven Theory: how to derive mathematical models directly from data

EUROfusion Consortium, JET, Culham Science Centre, Abingdon, OX14 3DB, UK

by A.Murari[1], E.Peluso[2], M.Lungaroni[2], P.Gaudio[2], J.Vega[3] and M.Gelfusa[2] and JET Contributors*

1) Consorzio RFX (CNR, ENEA, INFN, Universita' di Padova, Acciaierie Venete SpA), Corso Stati Uniti 4, 35127 Padova, Italy.
2) Department of Industrial Engineering, University of Rome "Tor Vergata", via del Politecnico 1, Roma, Italy
3) Laboratorio Nacional de Fusión, CIEMAT. Av. Complutense 40. 28040 Madrid. Spain

Formulating general theories, based on first principles, for thermonuclear plasmas is quite difficult and numerical simulations are computational demanding, because these physical objects are highly nonlinear and open systems, involving phenomena on many spatial and time scales. The poor accessibility for measurement results often in data with quite high error bars. The challenges posed by the complex nature of high temperature plasmas are therefore compounded by the difficulties inherent in interpreting the large amounts of uncertain data produced by the diagnostics. On JET for example, in a well diagnosed discharge, more than 50 Gigabytes of data can be generated and the whole database now approaches half a Terabyte but some measurements can have error bars reaching 30% (even if 10 % is more common). Due to the large amounts of data available and the significant uncertainties in the measurements, important information can remain buried in the databases if they are analysed manually or with traditional methods. To overcome these difficulties, in the last years a new methodology to extract mathematical models directly from the databases has been developed. The technique is based on Symbolic Regression via Genetic Programming and imposes no limitation on the complexity of the derived equations, except those due to the quality of the data available. To illustrate the potential of the approach, the following topics have been selected for a detailed discussion: scaling laws of the energy confinement time and the models of the boundary between safe and disruptive regions of the operational space.
With regard to scalings, it is important to note that the proposed approach does not assume a priori that the scaling laws are in power law monomial form. Indeed in various applications, such as the L to H transition and the confinement time, it can be demonstrated that different mathematical expressions are much more adequate to interpret the experimental data. In relation to disruptions, the deployment of Symbolic Regression via Genetic Programming has allowed reformulating the models of machine learning tools in more interpretable and physically meaningful form.

*See the author list of "X. Litaudon ***et al*** 2017 ***Nucl. Fusion*** 57 102001