ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 23 МАТЕМАТИКИ Выпуск 2
2016

# XXXIII INTERNATIONAL SEMINAR ON STABILITY PROBLEMS FOR STOCHASTIC MODELS

**D. Malakhov** [1], **A. Korchagin** [2], **A. Pupykina** [3], **N. Skvortsova** [4] (Prokhorov General Physics Institute of RAS, Moscow State University, Togliatti State University, Prokhorov General Physics Institute of RAS & National Research Nuclear University "MEPhI"). **Data preprocessing and statistical analysis techniques for toroidal fusion devices.**

Modern fusion devices include many components that collect multi-channel data related to high-temperature plasma state. The result of these data collection experiments are multi-million signal samples registered on various diagnostics and technical sub-systems of toroidal fusion devices. The biggest data samples are collected when studying "extremely non-linear" plasma states. Such samples contain data related to plasma turbulence required for understanding the processes of anomalous transport of particles and energy transfer in such environments.

Fusion devices have dozens of multi-channel diagnostics aimed to study multiple non-linear processes taking place in plasma (e. g., ITER reactor diagnostics — www.iter.org). In general, data set for a single diagnostic can be stored in a form of matrix with size of $N_c \times N_s$ :

$$S = \begin{pmatrix} s_0(t_0) & s_0(t_0+\tau) & s_0(t_0+2\tau) & \ldots & s_0(t_0+N_s\tau) \\ s_1(t_0) & s_1(t_0+\tau) & s_1(t_0+2\tau) & \ldots & s_1(t_0+N_s\tau) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{N_c}(t_0) & s_{N_c}(t_0+\tau) & s_{N_c}(t_0+2\tau) & \ldots & s_{N_c}(t_0+N_s\tau) \end{pmatrix}$$

where $\tau$ is the inverse of the sampling frequency $F_s$, $N_c$ is the number of channels and $N_s$ is the number of samples [1,2].

When processing turbulence-related data, one has to remember that time series registration occurs in huge-scale live environments. This means that the diagnostic channels and the registering hardware can evolve/change over time.

A group of diagnostics can be presented as $G = \{G_0, G_1, \ldots, G_M\}$, where $M$ is the number of diagnostic. Each such diagnostic has it's own finite set of parameters $P_M^L$ at any given (limited) time period $[t_a, t_b] = \{T_M^L \in \mathbf{R} \colon t_a \leqslant T_M^L \leqslant t_b\}$, where $L$ — experiment number. Diagnostic parameters are: $N_c$ number of channels, $N_s$ number of samples, $N_b$ bit of resolution, $T_M^L$ — diagnostic time related to the local time of the environment, signal amplification coefficients and many others.

It's worth mentioning that time series are better stored in databases in "raw" format, i. e. without any modifications. In the case of any signal registering errors the relevant data are also saved to the database, but doesn't take part in the analysis afterwards.



[1] e-mail: malakhov@fpl.gpi.ru
[2] e-mail: sasha.korchagin@gmail.com
[3] e-mail: anna.pupikina@gmail.com
[4] e-mail: nina@fpl.gpi.ru

Pre-processing procedures for matrix $S$ have a number of operators $F_M^L$ for each diagnostic $G_M$ and the index of the experiment $L$. For each group $\{M, L\}$ there can be defined a dedicated set of operands, but in typically pre-processing procedures include: preliminary grouping of data, noise cancellation, physical values calculation (if required), calculation of parameters dynamic evolution [3] (statistical moments, FFT, energy characteristics, etc.), statistical analysis [4] (EM, SEM, hybrid methods, etc.) and results validation.

The result of pre-processing are data frames $\varphi\colon S \times F \to D$ for the new discrete time line. Selection of such time line is dictated by one of the conditions in order to obtain optimal and valid result, on the other hand, data interpolation can be applied. Such data frames $D$ are saved in the timeseries database and used as source data for deep learning and analysis algorithms.

This paper present a methodology for data pre-processing and statistical analysis for data collected on diagnostic system for turbulence study on stellator «L-2M» [5].

This research was supported by the Russian Federation Presidential Grant (No.MK-5298.2016.8).

## REFERENCES

1. *Haskey S. R.*, *Blackwell B. D.*, *Pretty D. G.* Clustering of periodic multichannel timeseries data with application to plasma fluctuations. — Comput. Phys. Commun., 2014, v. 185, is. 6, p. 1669–1680.
2. *Haskey S. R.*, *Blackwell B. D.* A data mining algorithm for automated characterisation of fluctuations in multichannel timeseries. — Comput. Phys. Commun., 2009, v. 180, is. 10, p. 1768–1776.
3. *Skvortsova N. N.*, *Borzosekov V. D.*, *Konchekov E. M.*, *Tedtoev B. A.*, *Vasilkov D. G.*, *Malakhov D. V.*, Multiparametric data processing on fusion device stellarator L-2M. — J. Phys.: Conference Ser., 2014, v. 490, conf. 1, p. 12073-1–12077-5. doi:10.1088/1742-6596/490/1/012073
4. *Korolev V. Yu.*, *Korchagin A. Yu.* A modified gridmethod for statistical separation of normal variance-mean mixtures. — Informatics and Its Applications, 2014, v. 8, is. 4, p. 11–19. (In Russian.)
5. *Skvortsova N. N.*, *Chirkov A. Y.*, *Kharchevsky A. A.*, *Malakhov D. V.*, *Gorshenin A. K.*, *Korolev V. Yu.* Doppler reflectometry studies of plasma gradient instabilities in L-2M stellarator. — J. Phys.: Conference Ser., 2016, v. 666, p. 12007-1–12007-7. doi:10.1088/1742-6596/666/1/012007