


OPEN

A Model Falsification Approach to Learning in Non-Stationary Environments for Experimental Design

Andrea Murari¹, Michele Lungaroni ^{2*}, Emmanuele Peluso², Teddy Craciunescu³ & Michela Gelfusa²

The application of data driven machine learning and advanced statistical tools to complex physics experiments, such as Magnetic Confinement Nuclear Fusion, can be problematic, due the varying conditions of the systems to be studied. In particular, new experiments have to be planned in unexplored regions of the operational space. As a consequence, care must be taken because the input quantities used to train and test the performance of the analysis tools are not necessarily sampled by the same probability distribution as in the final applications. The regressors and dependent variables cannot therefore be assumed to verify the i.i.d. (independent and identical distribution) hypothesis and learning has therefore to take place under non stationary conditions. In the present paper, a new data driven methodology is proposed to guide planning of experiments, to explore the operational space and to optimise performance. The approach is based on the falsification of existing models. The deployment of Symbolic Regression via Genetic Programming to the available data is used to identify a set of candidate models, using the method of the Pareto Frontier. The confidence intervals for the predictions of such models are then used to find the best region of the parameter space for their falsification, where the next set of experiments can be most profitably carried out. Extensive numerical tests and applications to the scaling laws in Tokamaks prove the viability of the proposed methodology.

Experimental design in non stationary environments

In the last two decades, incessant technological developments and various breakthroughs in computer and automation have led to an increased complexity of the data to be processed, very often in real time. This complexity is also linked to the fact that nowadays automated tools are expected to monitor and control systems, whose dynamics can change due to their internal evolution or external disturbances. The consequent variations in the systems require sophisticated forms of adaptive learning under “concept shift”¹.

The problems posed by non-stationary conditions are particularly relevant in many scientific applications, especially for the planning of experiments and the design of new devices. These tasks inherently require extrapolation to uncharted territory. Indeed, when planning new experiments, it is essential to explore new regions of the operational space, to increase knowledge as much as possible. These issues, of planning new experiments and designing new devices, are particularly challenging in the case of complex, nonlinear systems difficult to model from first principles. For such systems, the only alternative consists of deriving empirical models directly from the data, using some forms of data mining such as genetic programming and machine learning. The non-stationarity of the phenomena, whose behavior the models have to predict, is often neglected but should not be underestimated, because the training set can be statistically quite different from the test set and the final application. On the contrary, the mathematics of practically all statistical and machine learning tools hinges on the i.i.d. (independent and identical distribution) assumption. The i.i.d conditions imply that the data are sampled independently from a stationary distribution function, which must be the same for the training set, for the test set and for the final

¹Consorzio RFX (CNR, ENEA, INFN, Universita' Padova, Acciaierie Venete SpA), Corso Stati Uniti 4, 35127, Padova, Italy. ²Department of Industrial Engineering, University of Rome “Tor Vergata”, via del Politecnico 1, 00100, Rome, Italy. ³National Institute for Laser, Plasma and Radiation Physics, Atomistilor Street 409, Magurele-Bucharest, RO-077125, Romania. *email: michele.lungaroni@uniroma2.it

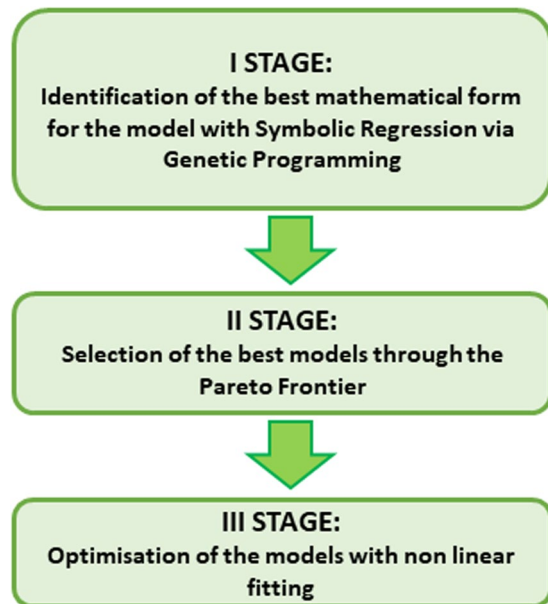


Figure 1. Block diagram of the steps required to perform Symbolic Regression via Genetic Programming for the data driven derivation of mathematical models.

deployment conditions. In many experiments in physics and chemistry, the i.i.d assumption is strongly violated, resulting in suboptimal performance of the models and predictors. This is certainly true for example in Magnetic Confinement Nuclear Fusion (MNCF). In MNCF the experiments, called discharges, are pulsed. On JET, the largest Tokamak in the world, they can last for tens of seconds and they can be repeated about every half an hour²⁻⁵. The parameter settings of each discharge are typically different, depending on the results of the previous experiments. The i.i.d. conditions are therefore violated and the models to plan future experiments have to be derived under conditions of “concept shift”.

Contrary to other approaches, the methodology presented in the following does not make any “a priori” assumption about the probability density functions (pdf) of the regressors; on the contrary it helps determining their most appropriate values for performing new experiments. On the other hand, it is predicated on the hypothesis that the concept shift is slow and therefore that experiments can be planned on the basis of previous data, which carry extrapolable information on the status of the system under investigation. To address the issue of slow “concept shift”, the models are derived using Symbolic Regression (SR) via Genetic Programming (GP)^{6,7}. The choice of this family of techniques is driven by the consideration that they can typically output more scientific meaningful equations than traditional machine learning methods. With this very powerful tools, a series of candidate models are selected, using the approach of the Pareto Frontier. Nonlinear fitting of the data allows determining the confidence intervals of the candidate models⁸. At this point, a numerical exploration of the operational space permits to identify the best combination of parameters to discriminate between the models. This typically means selecting a combination of the independent variables, for which the predictions of the models are different well outside their confidence intervals. Once the most advantageous operational region to investigate is selected, experiments are performed, new data are collected and the process can be repeated. SR via GP can be deployed again to identify a new set of candidate models to plan additional experiments. The process ends when only a single model, or a sufficiently limited number of candidate models, has been identified.

The paper is structured as follows. The technique of Symbolic Regression via Genetic Programming is briefly reviewed in the next Section. The description of the proposed method of learning in non-stationary conditions for experimental design is the subject of Section 3, together with examples of the numerical tests performed to verify the viability of the approach. A real life application, the design of experiments to investigate the scaling laws for the confinement time in Tokamaks, using the data of an international multimachine database, is discussed in Section 4. Discussion and future developments are the subject of the last Section 5.

Symbolic Regression via Genetic Programming for Data Driven Modelling

As mentioned in the introductory section, this paper presents a new methodology for experimental design. In this perspective, the main tool used is Symbolic regression via Genetic Programming. The technique consists basically of testing a large number of mathematical expressions to fit a given database. To keep the number of alternatives to be evaluated at a manageable level, various generations of models are built. The most performing ones of the previous generation are retained and used as the basis for producing a new generation of models with genetic programming techniques. The flow chart showing the main steps of the methodology is reported in Fig. 1.

In more detail, in terms of knowledge representation, the various candidate formulas are expressed as trees. In this context, the trees can be considered as consisting of functions and terminal nodes. The functional nodes can be arithmetic operators, any type of mathematical functions and saturation terms^{6,7}. This representation

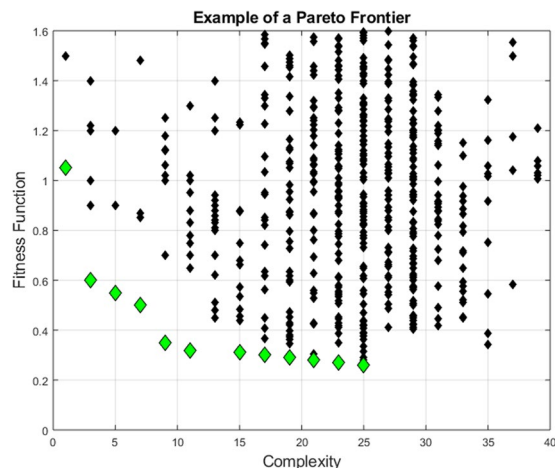


Figure 2. The models indicated with diamonds form the Pareto Frontier, since they have the best values of the FF for each level of complexity.

of the formulas is not unique but it is the preferred one because it permits an easy implementation of Genetic Programming (GP) operations. Genetic Programs are computational techniques, which have been explicitly developed to solve complex optimization problems, mimicking the genetic processes of living organisms. They operate on a population of individuals, e.g mathematical expressions in our application. Each individual represents a possible solution, a potential model of the phenomenon under investigation in our case. One of the crucial aspects of SR via GP is the qualification of these candidate models. Such an evaluation is based on specific indicators called fitness functions (FFs). The FF is selected to measure how good an individual is with respect to the database. Once the best individuals have been identified, on the basis of the FF, genetic operators (Reproduction, Crossover and Mutation) are applied to them to obtain the new population. Therefore SR via GP operates in such a way that better individuals are more likely to have more descendants than inferior individuals. The iteration is stopped when a stable and acceptable solution is identified. At this point, the algorithm provides the solution with best performance in terms of the FF^{9–12}.

The fitness function is probably the most crucial element of the genetic programming approach, because it determines the quality of the candidate solutions. Various indicators have been used in the past to implement the FF: the Akaike Information Criterion (AIC), the Takeuchi Information Criterion (TIC) and Bayesian Information Criterion (BIC)^{13,14}. It is worth mentioning that, in all numerical and experimental cases reported in the paper, BIC and TIC have provided the same results as AIC, which is discussed in more detail in the following as a representative example. The AIC is explicitly conceived to reduce the generalization error. Indeed it can be demonstrated that AIC provides an unbiased estimate of the predictive inaccuracy of a model. The most widely used form is:

$$AIC = n \cdot \ln\left(\frac{RMSE}{n}\right) + 2k \quad (1)$$

where RMSE is the Root Mean Square Error, errors indicate the residuals, the difference between the experimental values and the estimates of the scaling laws, k is the number of nodes in the model and n the number of entries in the database.

The AIC and the other mentioned criteria are indicators to be minimised, in the sense that better models have lower values. This property can be appreciated by inspection of the various indicators. Indeed, all of them consist basically of two parts. The first one depends on the quality of the fit (on the RMSE of the residuals for the case of the AIC). Models closer to the data have lower values of this term. The second addend implements a penalty for complexity, since it is proportional to the number of nodes in the trees representing the model equations. All the mathematical details to fully appreciate the relative merits of these criteria can be found in¹⁴.

In practical applications, given the limitations of the databases available, typically the Fitness Functions does not manage to converge on a single individual model, clearly outperforming all the others. Normally, SR via GP provides a series of models, which are good candidates for the interpretation of the data available. The main tool implemented, to select the final candidate models, is the Pareto Frontier (an example is shown in Fig. 2). The Pareto Frontier (PF) reports the best model for each level of complexity, according to the FF. As can be appreciated from Fig. 2, the PF presents an L shape, meaning that there is a tendency of lower returns: increasing the complexity above a certain level does not produce significant improvements in the fitting quality of the models. The models around the inflexion points of the PF are the ones, which require attention and are the best candidates for extrapolation.

The last step of the methodology consists of nonlinear fitting of the candidate models, identified with SR via GP⁸. This is an essential phase to associate confidence levels to the estimates of the models, which is an indispensable piece of information for scientific applications.

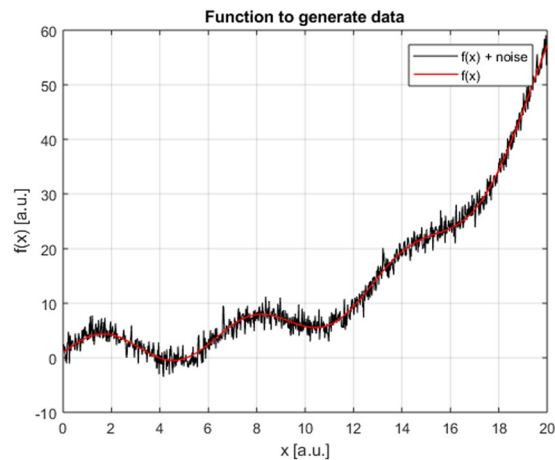


Figure 3. Function of one variable to illustrate the methodology. The red line is the plot of the function generating the data, Eq. (2). The black line reports the actual data including the noise.

Exploring the Operational Space with SR via GP

In this paper, a new alternative is proposed to complement the traditional approaches of training under concept shift. The main motivation is that, in the planning of future experiments, the pdf of the new experiments is unknown and the conditional probabilities of the dependent variables on the regressors are not necessarily the same. Therefore, the proposed technique is based on the selection of the parameter range more appropriate for the falsification of the available models. The database of past experiments is analyzed first with SR via GP. The main idea consists then of selecting the reasonable candidate models on the basis of the Pareto Frontier. After application of the nonlinear fitting procedures, it is possible to associate confidence intervals to the predictions of the models. With this information available, the developed algorithm traverses the unexplored operational space to identify the regions closest to the past data, where the predictions of the candidate models differ outside the confidence intervals. In other words, the technique determines the smallest variations in the operational parameters of the experiments to falsify the derived models. The range of parameters closest to the one already explored is selected for two main reasons: a) because typically this is the most accessible region for additional experiments b) because predictions of the previous models in this region are expected to be the most robust even in presence of concept shift. Once the experiments have explored the new region of the operational space and new data are collected, the process can be repeated until convergence on a sufficiently specific model for the interpretation of the phenomena under study.

The methodology just described has been subjected to a systematic series of numerical tests. Many families of mathematical functions have been used to generate synthetic data. Various levels of Gaussian noise have been added to the points generated by the numerical functions to simulate experimental conditions. The proposed method has then been iteratively applied to the data until the original function is identified. The results have always been positive and the proposed technique has always allowed recovering the original equations generating the data in a very efficient way. In the following, a quite challenging example is described in some detail to show the potential of the proposed approach.

For clarity's sake, a low dimensional case is illustrated, but it has been verified that the approach is equally valid for high dimensional cases, provided of course a sufficient number of good quality examples and adequate computational resources are available.

In Fig. 3, a simple example of a function depending on a single regressors is shown. The equation of the function used to generate the data is:

$$f(x) = 3 \sin x + \exp(x/5) \quad (2)$$

Gaussian noise, of zero mean and variance equal to 10% of the dependent variable absolute value, has been added to the individual points.

The first training has been performed by generating 100 points in the interval between 0 and 2. The best three solutions identified from the Pareto Frontier, after application of SR via GP, are:

$$\begin{aligned} y_{1,1} &= 2.28 \left(\sin x + \frac{1}{1 + \exp(-1.26x)} \right) \\ y_{2,1} &= 4.38 \sin x^{0.62} \\ y_{3,1} &= 3.53x^{0.41} \end{aligned}$$

The three functions are shown Fig. 4; from the confidence intervals, it appears clearly that a very advantageous range of x to falsify the three solutions is the one between 3 and 5. Synthetic points have therefore been generated in that interval simulating the results of additional experiments. SR via GP has then been applied including this new data to obtain additional candidate models. By repeating the process two more times, it has been possible to converge on the final set of equations:

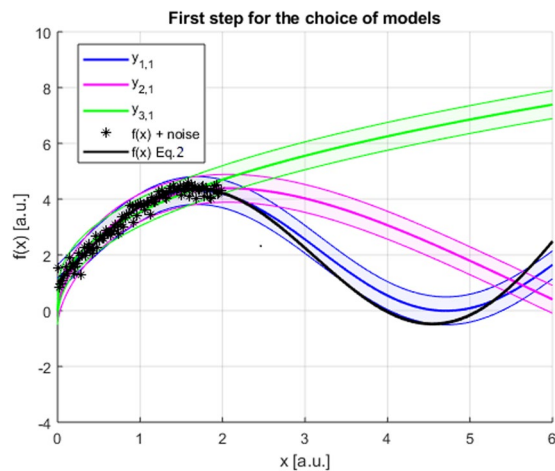


Figure 4. First step to converge on the final and correct solution. The data provided to SR via GP are 100 points in the interval between 0 and 2. The three best candidates derived from the Pareto Frontier are reported in green, red and blue with the relative confidence intervals. The input points are depicted in black. In black also the actual function generating the data.

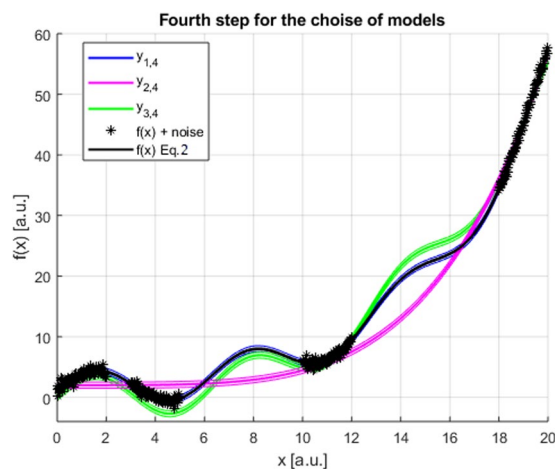


Figure 5. Third and final step to converge on the final and correct solution. The new data provided to SR via GP are 100 points in the interval in the interval between 16 and 18 (after two previous iteration with data between 3 and 5 and between 10 and 12). The three best candidates derived from the Pareto Front are reported in green, red and blue with the relative confidence intervals. The input points are depicted in black. In black also the actual function generating the data.

$$y_{1,4} = 3 \sin x + \exp(0.2x)$$

$$y_{2,4} = 3.40 \sin x + 0.0073 x^{2.96}$$

$$y_{3,4} = 0.00007 x^{4.48} + 2$$

The three candidate equations are plotted in Fig. 5. They all fit the input data acceptably but now the right solution can be clearly identified because it outperforms significantly all the others in terms of all statistical indicators. It should be mentioned that traditional methods of covariate shift¹⁵, to analyse the data in non-stationary conditions, would be significantly less efficient in solving the problem and converging on an acceptable solution, because they have problems handling the oscillating character of the function generating the data.

Real World Example: Scaling of Confinement Time in Tokamaks

As an example of real world applications, in this section the case of scaling laws is discussed. Scaling laws are particularly important not only for the understanding of the physics but also for the planning of new experiments and devices. In the Tokamak community, the extrapolation of the energy confinement time to the next generation of machines has been studied for several decades. Unfortunately, the problem is too complex to be attacked with models based on first principles. Therefore, various empirical scaling expressions have been proposed

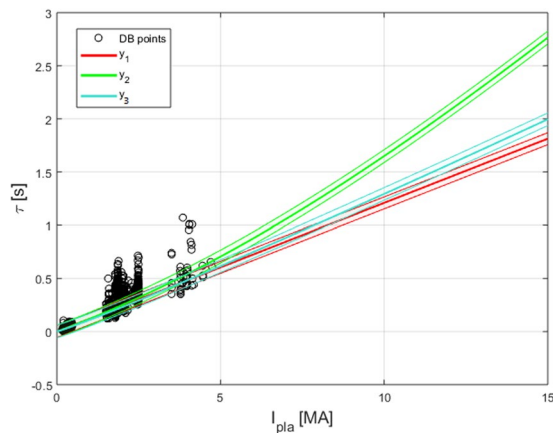


Figure 6. The three intervals of plasma current (x axis) sampled to identify the final scaling laws for the energy confinement time (y axis) reported in Table 1: below 0.5 MA, between 1.5 and 2.5 MA and between 3.5 and 5 MA. Again the black dots indicate the experimental data used as input to the three iterations of the procedure.

Model	τ_{ITER}
y_1	2.55 s
y_2	4.34 s
y_3	2.85 s

Table 1. The prediction for the confinement time of ITER for the three families of scaling laws identified after the third iteration of the proposed procedure.

based on the analysis of data collected in previous experiments. Their extrapolation to future experiments and devices is a typical example of learning under covariate shift. Since the objective is to apply the derived models to larger machines, and therefore in a completely different range of parameters, they are typically predicated on the assumption that the scalings are in power law form. To show the potential of the proposed approach, the devised method of learning under covariate shift is tested using an international experimental database for the confinement time in Tokamaks, DB3v13f¹⁶. This database was explicitly conceived to support advanced studies of the confinement time, including validated signals from most relevant Tokamak machines ever operated in the world. In line with the previous literature on the subject, the following quantities have been considered good candidate regressors in the present work:

$$B[T], I[\text{MA}], n[10^{19} \text{m}^{-3}], R[m], \varepsilon, k, P[\text{MW}]$$

In the previous list, k indicates the volume elongation, ε the inverse aspect ratio, n the central line average plasma density, B the toroidal magnetic field, R the plasma major radius, I the plasma current and finally P the estimated power losses. The main objective of the analysis is to identify the best range of plasma currents to obtain a good confinement. To exemplify the potential of the proposed approach, SR via GP has been applied first to the smaller devices in the database, i.e. all the devices except JET, which operated at a plasma current up to 0.5 MA. Two iterations of the proposed procedures identified as optimal the current intervals between 1.5 and 2.5 MA and between 3.5 and 5 MA to collect new data, as shown in Fig. 6.

Including the data of experiments carried out in these intervals allows reducing the candidate models to three main families of equations, reported in the following:

$$y_1 = 0.076 \cdot I \cdot R^2 \cdot k \cdot P^{-1} \cdot \frac{1}{1 + \exp(-0.201n^{2.004})} \cdot \left(\frac{1}{n \cdot P^{1.293}} \cdot \frac{1}{1 + \exp(0.715M)} \right)^{-0.1761}$$

$$y_2 = 0.1831 \cdot I \cdot R^2 \cdot P^{-0.662} \cdot \frac{1}{1 + \exp(-0.094 \cdot I)} \cdot \frac{1}{1 + \exp(-0.408 \cdot n)}$$

$$y_3 = 0.070 \cdot I^{1.071} \cdot R^{1.706} \cdot k^{1.250} \cdot P^{-0.715} n^{0.100} \cdot \frac{1}{1 + \exp(-0.408 \cdot n^{1.036})}$$

The new equations narrow down the extrapolations to the next international device, ITER, to a reasonable interval between 4.35 and 2.85 s (see Table 1). Also ITER plasma current, 15 MA, seems to be more than adequate to discriminate between even the closest scaling laws y_2 and y_3 (see Fig. 6).

It is important to notice that the equations, derived with the proposed approach, are practically the same as the ones already extracted using the entire ITPA database but have been obtained using only about 50% of the entries¹⁰. Of course, a reduction of such a factor in the experimental investments and efforts would be extremely significant, given the cost of running present day large fusion devices.

Conclusions

In this paper, an original methodology is presented to guide scientists in the design of experiments in new regions of the operational space. This is a very important step in the scientific process, since only exploration of new regions of the parameters space can provide real new information and knowledge. On the other hand, planning experiments in an unexplored range of parameters is delicate both conceptually and practically, because the extrapolation of previous knowledge is uncertain. Since the available models have been derived in conditions different from the ones of the final applications, the i.i.d assumptions cannot be invoked. Consequently, care must be taken because the models obtained with old data can perform sub optimally and provide even wrong answers.

To alleviate this problem, the methodology proposed in the paper is based on the falsification of the models with experiments in the range of parameters as close as possible to the previous experiments. The approach, based on SR via GP, has been successfully tested with a variety of numerical tests. The application to a multimachine international database of Tokamak devices has also provided very encouraging results. It is worth mentioning that practically the same methodology can be adopted to fine tune the analysis of existing databases. In various scientific fields, large data sets are indeed available but only a small subset of the entries can be analysed in detail due to the lack of man power. The suggested approach, of falsifying the models obtained with SR via GP, can also provide clear indications about the most important entries, on which to concentrate further investigations. On the other hand, the proposed methodologies can help only addressing “soft nonlinearities” and not catastrophic events, not having left any signature in the available data. This is of course a limitation of all data driven techniques, which are designed to learn what is in the data and cannot provide guidance about information not available in their inputs. Nonetheless it should be considered that the proposed approach of SR via GP can be advantageous, compared to other methodologies, at least in two respects. First, by better modelling soft nonlinearities, they can help in approaching regions at risk of catastrophic changes. Moreover, if “a priori” information about the nature and the possible localisation in the parameter space of strong nonlinearities is available, the models can be steered to favour solutions including this knowledge (for example by selecting appropriate basis functions).

With regard to future developments, more sophisticated forms of the fitness function, such as the Widely Applicable Information Criteria¹⁷ and Generalised information criteria¹⁸, can be certainly implemented without any particular conceptual difficulty, provided of course the quality and quantity of the data and the computational resources are adequate to take advantage of them. Another important topic to be further refined is the treatment of the errors. Methods of Information Geometry, particularly the Geodesic Distance between probability density functions, are expected to have strong potential to improve the capability of the proposed methodology¹⁹. On a longer time scale, it should be remembered that the proposed techniques are purely statistical. Of course, the availability of actual causal models would improve the confidence in the design of new experiments and/or machines. Identifying good candidate causal models from databases is therefore one of the main lines of research for the future research on learning under systemic concept shift. Comparing the results of the proposed tools with Graphical and Bayesian Network Models^{20,21} is also expected to generate very interesting new insights.

Data availability

The ITPA database (DB3v13f) used for this study: https://urldefense.proofpoint.com/v2/url?u=http-3A__efdasql.ipp.mpg.de_threshold_Public_ThresholdDbInfo_ThresholdvarDB2-5F3.htm&d=DwIBaQ&c=vh6FgFnduejNhPPD0fl_yRaSFZy8CWbWnIf4XJhSqx8&r=W8RE-88OJ6HkLx5-vnSwGMbECZfnybKmwCXImBggWgpaE5ASEwQeYcLKZnw2tz-i&m=gnXLSpCZeP2D39tup-cwv45ohkSr48CVU0tFO2UmW4&s=MtDspa2m7fXrriEAMABTQ5R6ElmLaECxYmdho4lzD_c&e. If the link is offline, contact the corresponding author.

Received: 2 May 2019; Accepted: 9 November 2019;

Published online: 29 November 2019

References

1. Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J. & Wozniak, M. Ensemble Learning for Data Stream Analysis: a survey. *Information Fusion* **37**, 132–156, <https://doi.org/10.1016/j.inffus.2017.02.004> (2017).
2. Wesson, J. Tokamaks. Clarendon Press Oxford (Oxford Third edition 2004).
3. Romanelli, F. *et al.* Overview of JET results. *Nuclear Fusion* **49** (10), Article number 104006, <https://doi.org/10.1088/0029-5515/49/10/104006> (2009).
4. Ongena, J. *et al.* Towards the realization on JET of an integrated H-mode scenario for ITER. *Nuclear Fusion* **44**(1), 124–133, <https://doi.org/10.1088/0029-5515/44/1/015> (2004).
5. Fasoli, A. *et al.* Computational challenges in magnetic-confinement fusion physics. *Nature Physics* **12**, 411–423, <https://doi.org/10.1038/NPHYS3744> (2016).
6. Schmid, M. & Lipson, H. Distilling Free-Form Natural Laws from Experimental Data. *Science*, **324**, <https://doi.org/10.1126/science.1165893> (2009).
7. Koza, J. R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. (MIT Press, Cambridge, MA, USA, 1992).
8. Dielman T.E. *Applied Regression Analysis*. South Western Cengage Learning, Mason, USA (2005).
9. Murari, A., Peluso, E., Gelfusa, M., Lupelli, I. & Gaudio, P. A new approach to the formulation and validation of scaling expressions for plasma confinement in tokamaks. *Nuclear Fusion* **55**(7), 073009, <https://doi.org/10.1088/0029-5515/55/7/073009> (2015).
10. Murari, A. *et al.* Symbolic regression via genetic programming for data driven derivation of confinement scaling laws without any assumption on their mathematical form. *Plasma Phys. Control. Fusion*. **57**(1), 014008, <https://doi.org/10.1088/0741-3335/57/1/014008> (2015).
11. Peluso, E., Murari, A., Gelfusa, M. & Gaudio, P. A statistical method for model extraction and model selection applied to the temperature scaling of the L–H transition. *Plasma Phys. Control. Fusion* **56**, 114001, <https://doi.org/10.1088/0741-3335/56/11/114001> (2014).
12. Murari, A., Peluso, E., Lungaroni, M., Gelfusa, M. & Gaudio, P. Application of symbolic regression to the derivation of scaling laws for tokamak energy confinement time in terms of dimensionless quantities. *Nuclear Fusion* **56**, 026005, <https://doi.org/10.1088/0029-5515/56/2/026005> (2016).
13. Hirotsugu, A. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6), 716–723 (1974).

14. Kenneth, P. B & Anderson, D. R. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. Springer. (2nd ed) (2002).
15. Sugiyama, M. & Kawanabe, M *Machine Learning in Non-Stationary Environments Introduction to Covariate Shift Adaptation*. MIT press ISBN: 9780262017091 (2006).
16. McDonald, D. *et al.* Recent progress on the development and analysis of the ITPA global H-mode confinement database. *Plasma Phys. Control. Fusion*. **46**, 519–34, <https://doi.org/10.1088/0029-5515/47/3/001> (2004).
17. Konishi, S. & Kitagawa, G. “*Information Criteria and Statistical Modelling*” *Springer Series in Statistics*. (Springer, NY, 2008).
18. S.Watanabe *Journal of Machine Learning Research* **14** 867–897??? (2013).
19. Murari, A. *et al.* Clustering based on the geodesic distance on Gaussian manifolds for the automatic classification of disruptions. *Nuclear Fusion* **53** (3), <https://doi.org/10.1088/0029-5515/53/3/033006S>. (2013).
20. Koller, D. & Friedman, N. “*Probabilistic Graphical Models. Principles and Techniques*”. (The MIT Press, Massachusetts, 2009).
21. Darwiche, A. “*Bayesian Networks: an Introduction*”. (Cambridge University Press, Cambridge, 2009).

Acknowledgements

This research received no external funding

Author contributions

Data curation, M.L.; Formal analysis, A.M. and E.P.; Investigation, E.P. and M.G.; Methodology, T.C.; Project administration, A.M.; Software, M.L.; Validation, M.G. and T.C.; Writing—original draft, A.M.; Writing—review & editing, A.M.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019