WPJET1-PR(18) 21512

A Murari et al.

A New Approach to the Planning of New Experiments based on Learning in Non-Stationary Conditions

Preprint of Paper to be submitted for publication in
Nuclear Fusion

This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014-2018 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

A New Approach to the Planning of New Experiments based on Learning in Non-Stationary Conditions

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

by A.Murari[1], M.Lungaroni[2], E.Peluso[2], M.Gelfusa[2], T.Craciunescu[3] 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) National Institute for Laser, Plasma and Radiation Physics, Magurele-Bucharest, Romania; teddy.craciunescu@inflpr.ro

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

Abstract

In the last decades, advanced statistical and machine-learning tools have made enormous progress and they find applications in many fields. On the other hand, their penetration in the scientific domain is delayed by various factors, among which one fundamental limitation is that they assume stationary conditions. This is due to the fact that traditional machine learning tools guarantee their results only if the data in the training set, the test set and the final application are sampled from the same probability distribution function. On the contrary, in most scientific applications, the main objective of new experiments consists precisely of exploring uncharted regions of the parameter space to acquire new knowledge. Traditional methods of covariate shift to address this issue are clearly insufficient. In this paper, a completely new method is proposed, which is based on the falsification of data driven models. The technique is based on symbol manipulation with evolutionary programmes. The performance of the approach has been extensively tested numerically, proving its competitive advantages. The capability of the methodology, to handle practical and experimental cases, has been shown with the example of determining scaling laws for the design of new experiments, a typical issue violating the assumptions of stationarity. The same methodology can be adopted also to investigate large databases or the outputs of complex simulations, to focus the analysis efforts on the most promising entries.

Keywords: Symbolic regression, Genetic programming, Planning of experiments, Concept shift, Covariate shift, Scaling laws

Corresponding author: gelfusa@ing.uniroma2.it

1 Learning in non-stationary conditions and experimental design

For about two decades, new technologies have allowed collecting unprecedented quantities of data about society. This data deluge has been experienced also by the sciences, in particular Big Physics experiments. For example, among the experiments coordinated by Eiroforum, at CERN the main detector ATLAS has shown the capability of producing 25 Petabytes of data per year and in its prime the Hubble space telescope was able of sending to earth Gigabytes of data per day. Coming to fusion, the data warehouse of the Joint European Torus (JET) is now approaching 0.5 Petabytes. These amounts of data challenge human understanding and manual analysis. Machine learning tools and advanced statistical techniques have therefore been extensively used to derive useful knowledge from increasingly large datasets. The great performance of these new tools has motivated also an increase in the ambitions, resulting in new challenges, ranging from the analysis of more complex phenomena to increasing demands in terms of interpretability and reliability of the results. Progress on these fronts is very substantial but, in the perspective of scientific applications, a basic limitation of present day tools is becoming very relevant. This is the fact that practically all machine-learning tools are based on the i.i.d. assumption [1]. Assuming that the data are independent and identically distributed means that the examples in the training set and test set are sampled randomly from the same probability distribution function as the final application. Such an assumption is clearly violated in experimental planning for the exact sciences. Indeed, the vast majority of new experiments in these fields are meant to explore new regions of the parameter space. Typically, a right trade-off has to be found. Extrapolating excessively in the unexplored regions of the parameter space can be too risky and cause the entire experiments to fail. On the other hand, too conservative choices can result in modest increases in knowledge, insufficient to justify the efforts. In any case, for experimental planning the i.i.d assumption is clearly untenable.

New methods are therefore needed to address situations, in which significant extrapolations are required, as in the case of planning of new experiments or of designing new devices. In probabilistic terms, the problem can be illustrated in following way. Let us assume that the goal of the experiments consists of studying the relation between the regressors x_i and the dependent variable y. According to the Bayesian theory, one can write:

$$y_i = \mathrm{P}(y|x=x_i^{pe})\mathrm{P}(x_i^{pe}) \qquad (1)$$

where the superscript *pe* indicates previous experiments, $\mathrm{P}(y|x=x_i^{pe})$ is the conditional probability and $\mathrm{P}(x_i^{pe})$ the prior. In new experiments, in general both the conditional and the prior probabilities could be different. This is a situation of learning in non-stationary conditions or under "*concept shift*" [2,3].

In many applications, to solve the issues inherent in learning in non-stationary environments, the approach of Learning under Covariate Shift (LCS) is adopted [4]. This methodology consists of a series of techniques for supervised learning, aimed at addressing the situation when the input points for the training follow a different probability density function as the input points of the test and of the final application. On the other hand, it is assumed that the conditional distribution of the output values, given the input points, remains unchanged:

$$\mathrm{P}(y|x=x_i^{pe}) = \mathrm{P}(y|x=x_i^{ne}) \quad (2)$$

where the superscript *ne* indicates new experiments. The traditional way of addressing the issue of covariate shift consists therefore of increasing the weight of the training points close to the one of the final application. The weights can be calculated on the basis of the probability ratio of the various inputs in the training set and in the test set. An academic example, similar to the one introduced in [4], is shown in Figure 1, in which linear fits are shown for different weights of the data. If the actual experiments are planned to take place in the region of the red squares, the fit obtained by weighting more those points has higher predictive power but can be misleading in other parts of the *x* axis.

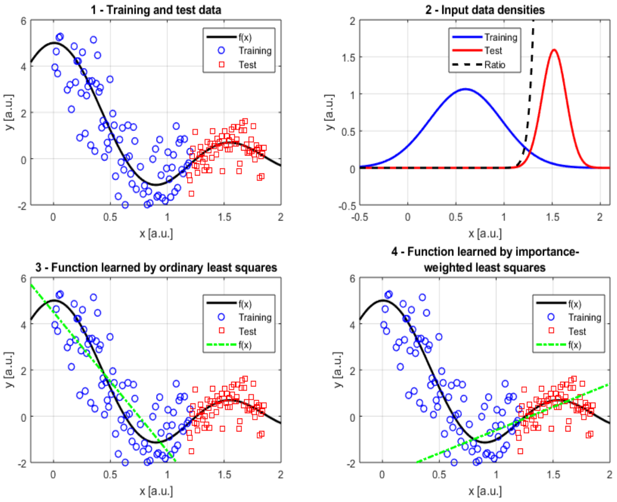


Figure 1 Example of learning under covariate shift for a function of one variable. Bottom left: linear fit for equal weights. Bottom right: linear fit for increased weights of the red squares.

Learning under covariate shift is insufficient in the case of the design of new experiments for various reasons. First of all, the methodology assumes and does not derive the new values of the regressors and therefore cannot guide in the planning of new experiments. Moreover, LCS needs also an a priori definition of the mathematical form of the models to fit the data (the linear fit in the academic example of Figure 1). In addition, LCS provides only a partial interpretation of the experimental evidence and not holistic models, which take into account all the available data.

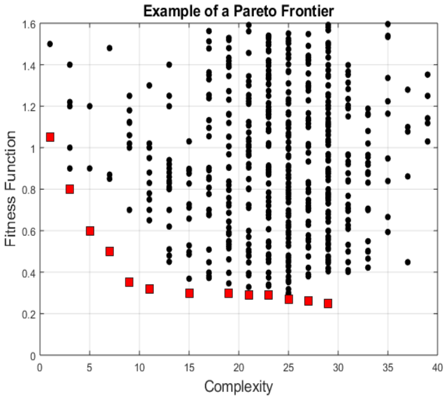
Finally, there is no reason to assume a priori that equation (2) is respected in the regions of the parameters explored by new experiments.

On the contrary, the new methodology proposed in this paper is not affected by any of the previous limitations. The approach is based on the falsification of the candidate models, extracted from the results of the already executed experiments, with the help of data driven tools. The analysis of the available data is performed with manipulation of symbols (formulas) using evolutionary programming (see next section for the details). The best models obtained from the available data are compared to identify the values of the parameters for which they provide clearly distinguishable predictions. New experiments can therefore be planned in these parts of the operational space to falsify the models. The process can be iterated until a satisfactory solution is found. The technique has been tested successfully with a systematic series of numerical tests, of which some examples are reported in Section 3. An application to the definition of the scaling laws for the energy confinement time in Tokamaks, on the basis of the ITPA database, is reported in Section 4. It is worth mentioning that the developed methodology is absolutely general and therefore can be applied also to guide in the investigation of large databases, which typically cannot be studied systematically for lack of resources; the proposed approach can therefore be very useful in identifying the most relevant entries on which to concentrate the further analysis efforts. The same can be said for the results of complex simulation. These points are discussed in the last section of the paper together with the conclusions.

2 Symbolic Regression via Genetic Programming for experimental design

As mentioned in the introductory section, this paper presents a new methodology for experimental design. The objective consists of determining the most important regions of the operational space to plan experiments, in order to identify the best models to describe the phenomenon under study. In this perspective, the main tool used is Symbolic regression via Genetic Programming. The main characteristics of these tools are summarised in the next subsection. Their application to the problem of extrapolation for experimental design is the subject of the following subsection.

2.1 Short overview of Symbolic Regression via Genetic Programming



SR via GP permits to identify the most appropriate mathematical expressions for modelling the process under investigation. The approach consists basically of testing a large number of mathematical expressions to fit a given database. To keep the number of alternatives to be fitted at a manageable level, various generations of models are tested. 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.

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 constituted of functions and terminal nodes. The function nodes can be arithmetic operators, any type of mathematical functions and squashing terms [5,6]. This representation of the formulas is not unique but it is the preferred one because permits an easy implementation of Genetic Programming (GP) operations. Genetic Programs are computational techniques, which have been explicitly developed to help addressing complex optimization problems [5,6]. They are designed to emulate the evolution of living beings through the interplay of mutation and selection. They operate on a population of individuals, e.g. mathematical expressions in our case. Each individual represents a possible solution, a potential model of the experiment 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 a metric 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 generate 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 or some halting condition is met (e.g., a maximum number of iterations or sufficiently small errors in subsequent iterations). At this point, the algorithm provides the solution with best performance in terms of the FF [7-10].

The fitness function is probably the most crucial element of the genetic programming approach, because it is the indicator that measures the quality of the candidate solutions. Various quantities have been used in the past to implement the FF: the Akaike Information Criterion (AIC), the Takeuchi Information Criterion (TIC) and the Bayesian Information Criterion (BIC) [11-13]. The AIC is based on the Kullback-Leibler divergence and it can be demonstrated that it minimises the generalisation error. The AIC can therefore be considered an unbiased estimate of the predictive inaccuracy of a model. The most widely used form of AIC is:

$$AIC = n \cdot \ln\left(\frac{RMSE}{n}\right) + 2k, \qquad (3)$$

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 y_{data} provided, so the number of entries in the database (DB).

A variation of the AIC criterion has been developed to improve its discrimination capability when the “right model” is not included in the list of candidate models under investigation. This more robust indicator is called the Takeuchi’s Information Criterion (TIC), which in practice can be calculated with the formula:

$$TIC = n \cdot \ln\left(\frac{RSS}{n}\right) + \left(\frac{\tau}{\sigma_0}\right)^2 \cdot \left(k + 1 - \left(\frac{\tau}{\sigma_0}\right)^2\right), \qquad (4)$$

where RSS is the residual sum of squares, σ_0 the standard deviation of the models uncertainties and τ the standard deviation of the measurements, both assumed to present a Gaussian probability density function.

An alternative criterion, the BIC, is an unbiased estimator of the likelihood of a model. The form of the BIC indicator used in this paper is:

$$BIC = n \cdot \ln\left(\sigma^2_{(\epsilon)}\right) + k \cdot \ln(n), \qquad (5)$$

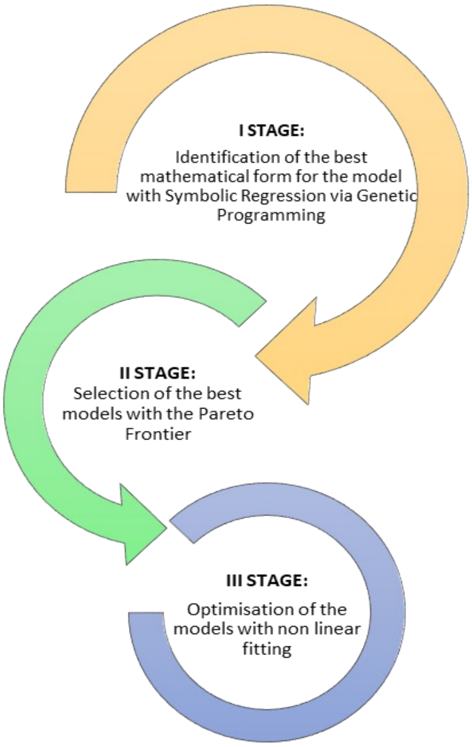

I STAGE:
Identification of the best mathematical form for the model with Symbolic Regression via Genetic Programming
II STAGE:
Selection of the best models with the Pareto Frontier
III STAGE:
Optimisation of the models with non linear fitting

Figure 3 Block diagram of the steps required to perform Symbolic Regression via Genetic Programming for the data driven derivation of mathematical models.

where $\epsilon = y_{data} - y_{model}$ are the residuals, $\sigma^2_{(\epsilon)}$ their variance and the others symbols are defined in analogy with the AIC expression.

All three criteria are indicators to be minimised, in the sense that better models have lower values of these metrics. This can be appreciated by inspection of the three indicators. Indeed, all of them consist basically of two parts. The first one depends on the quality of the fit. 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 three representing the model equations. All the mathematical background to fully appreciate the relative merits of these criteria can be found in [14]. To derive the results presented in this paper, the AIC criterion has been adopted for the FF.

In practical applications, given the limitations of the databases available, the Fitness Functions do not necessarily manage to identify a single individual model, clearly outperforming all the others. Normally, SR via GP converges on a series of models, which are good candidates for the interpretation of the data available. The main tool implemented to select the most performing candidate models is the Pareto Frontier (an example is shown in Figure 2). The Pareto Frontier (PF) reports the best models according to the FF for each level of complexity. As can be appreciated from Figure 2, the PF present an L shape form, meaning that there is a tendency of lower returns: increasing the complexity above a certain level does not produces 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 GP. This is an essential phase to associate confidence levels to the estimates of the models, which is an indispensable piece of information for the applications considered in this paper. A graphic overview of the methodology is shown in Figure 3.

2.2 Application of Symbolic Regression via Genetic Programming to Experimental Design

The tools described in Subsection 2.1 can be applied to the planning of future experiments. Indeed, the selection of the most profitable operational region, where to perform new experiments, can be considered an essential task of the data analysis process. In this perspective, the crucial aspect is the identification of the best parameter space region where to plan new experiments, which, from a statistical point of view, is equivalent to determining the new range of the regressors. Therefore, the technique should be refined to derive

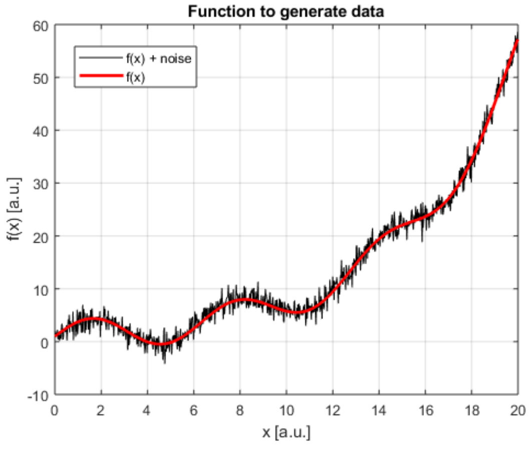
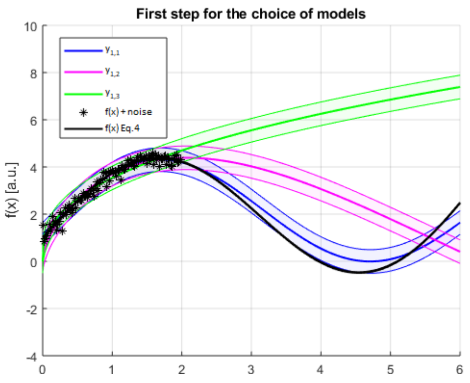


Figure 4 Function of one variable to illustrate the

the parameter range more appropriate for the falsification of the available models. In this perspective, the database of past experiments is analysed first with SR via GP. The main idea consists of selecting a pool of reasonable candidate models on the basis of the Pareto Frontier. The crucial point to appreciate is that at this stage, after application of nonlinear fitting, it is possible to obtain confidence intervals for the models predictions. With this information available, the algorithm, implemented to obtain the results reported in this paper, explores the operational space to identify the regions closest to the past data, where the candidate models differ sufficiently for their the predictions to be outside the confidence intervals. In practice, 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, because typically this is the most accessible region for additional experiments. Moreover, because predictions of the previous models in this close neighbourhood are expected to be the soundest even in presence of concept shift, it is wise not to extrapolate too much. In any case, the level of extrapolation for the following experiments is a parameter which can be tuned to best suit the needs of each experiment. At this point, once the experiments have explored the new region of the operational space and new data are collected, the process can be repeated. The best model, according to the FF, are selected using the Pf and a new region where they can be falsified is identified to perform the successive experiments. The process terminates when convergence on a sufficiently specific model, for the interpretation of the phenomena under study, is reached. The potential advantages of SR via GP for experimental design are various. The technique allows deriving directly from the data the most suited form of the models. A purely exploratory version of the algorithms can be implemented, with minimal a priori assumptions about the mathematical expression of the models (contrary to traditional fitting). On the other hand, if relevant a priori information is available, the solutions can be influenced to converge on certain specific classes of functions (by selecting appropriately the basis functions or the structure of the trees). The method also does not impose constraints neither on the type of errors affecting the data nor on the collinearity between the regressors. The proposed methodology is described in detail with the help of a simple example in the next section.

3 The falsification approach to experimental design: numerical tests

The methodology proposed in the last section 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. This statistics of the noise has been chosen because typically many measurements are affected by various sources of disturbance and therefore they satisfy the conditions of the central limit theorem. 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. The proposed methodology is also much more efficient than LCS, the more so the more complicated the problem. In the following, a quite challenging example is described in some detail to show the potential of the proposed approach. For clarity's sake, mainly 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 Figure 4 a simple example of a function of a single variable is shown. The equation of the function used to generate the data is:

$$f(x)=3\sin x+\exp\left(x/5\right) \qquad (4)$$

Gaussian noise, of zero mean and variance equal to 10% of the average 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 are:

$$y_{1,1}=2.28\left(\sin x+\frac{1}{1+\exp\left(-1.26\,x\right)}\right)$$

$$y_{1,2}=4.38\sin x^{0.62}$$

$$y_{1,3}=3.53\,x^{0.41}$$

The three functions are shown in Figure 5; from the confidence intervals it appears clearly that a very advantageous interval to falsify the three solutions is the one between 4 and 6. Therefore 100 additional points have been generated in this interval. The best solutions identified by SR via GP are shown in Figure 6 and their equations are:

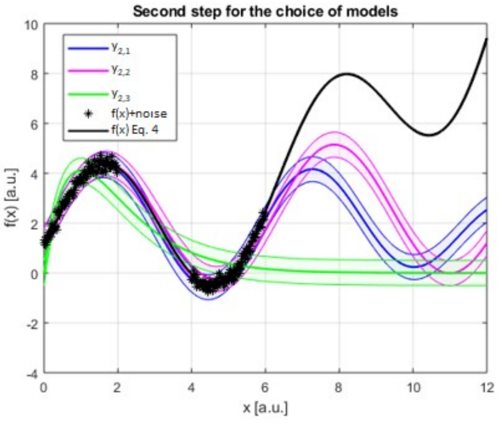


Figure 6 Second step of the proposed methodology to identify equation (4). The data provided to SR via

$$y_{2,1}=1.28\left[\sin\left(x^{0.67}\right)+\sin x+\frac{1}{1+\exp\left(-1.14\,x\right)}\right]+0.65$$

$$y_{2,2}=2.60\left(\sin x+\frac{1}{1+\exp\left(-0.48\,x\right)}\right)$$

$$y_{2,3}=11.14\,x\exp\left(-x\right)$$

At this point, a suitable interval to discriminate between the models is the one between 8 and 12. To progress the identification of the most suitable solution, therefore 100 points have been generated in this interval. The best outputs of SR via GP, considering also these additional entries (100 more points), are the following three equations:

$$y_{3,1}=3.27\sin x+0.43\,x^{1.31}-3.69\frac{1}{1+\exp\left(-0.70\,x\right)}+2.98$$

$$y_{3,2}=0.83\,x+3.47\sin x$$

$$y_{3,3}=0.79\exp\left(0.19\,x\right)$$

These equations provide different predictions in the interval of the x axis between 16 and 18. Taking into account these additional inputs, the right solutions emerges quite clearly among the competing models, as shown graphically in Figure 7. A part from the pictorial view, the right equation can be identified on the basis of the statistical indicators, which are all clearly better for the right model.

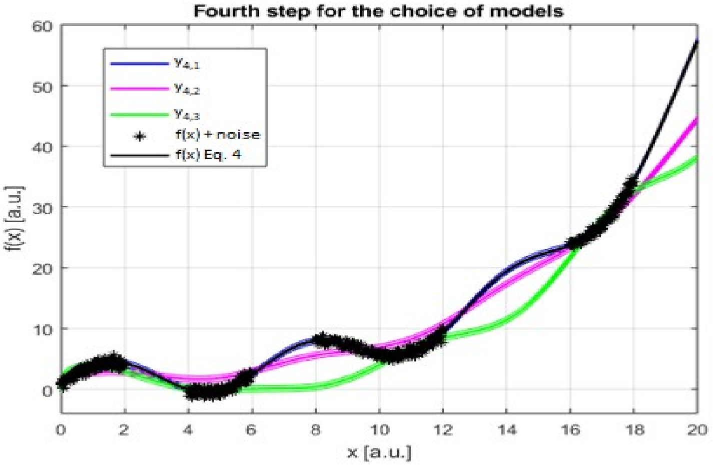


Figure 7 Fourth step to converge on the final and correct solution: eqution (4). The best data provided to SR via GP are 100 points in the interval in the interval between 16 and 18. The three best candidates derived from the Pareto Front are reported in green, red and blue with the relative confidence intervals. The input points are depicted in black. In black also the actual function generating the data.

4 The falsification approach to experimental design: energy confinement time

To exemplify the potential of the proposed methodology with a concrete example from Tokamak physics, in the following the scaling law of the energy confinement time is investigated. The importance of this parameter is obvious since it quantifies the rate at which energy is lost from the plasma. To cover a large range of regressors, an international database has been considered [14], which was explicitly built to support advanced studies of the confinement time. This ITPA database indeed includes validated signals from the vast majority of the most relevant Tokamak machines ever operated in the world. Coherently with the proposed procedure, for the next steps, only the intervals of plasma current indicated by symbolic regression have been considered. For the sake of direct comparison with previous scalings reported in the literature, the following quantities have been considered good candidates for the independent variables:

$$B[T], I[MA], n[10^{19} m^{-3}], R[m], M, \varepsilon, k; P[MW], M[a.m.u.].$$

In the previous lists k indicates the volume elongation, ε the inverse aspect ratio, q_{95} the plasma safety factor evaluated at the flux surface enclosing the 95% of the poloidal flux, 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 lost power [15]. The selection of the discharges included in the following analysis obeys also the selection rules of the DB3 dataset [15,16], used to obtain the famous IPB98y scaling law.

Since the main scaling parameter for this kind of studies is the plasma current I, at the first step of the procedure it is assumed than only data from devices with I less than 0.5 MA was available. This subset includes a total of 702 entries and with this data symbolic regression identifies the following 5 scaling laws as the most performing in terms of the model selection criteria:

$$y_{1,1}^{\tau} = 0.1434 \cdot I \cdot R^2 \cdot \frac{1}{1+\exp\left(-0.0363\, P^{-7.626}\right)} \cdot \frac{1}{1+\exp\left(-1.015\, n^2 \varepsilon\right)} \cdot \frac{1}{1+\exp\left(-0.5112\left(\frac{n}{P^{1.486}}\right)^{1.697}\right)}$$

$$y_{1,2}^{\tau} = 0.0239 \cdot I^{0.8469} \cdot R \cdot \exp(R) \cdot \frac{1}{1+\exp\left(-2.5222 n^{3.759} M \varepsilon^4 k\right)} \cdot P^{-0.604}$$

$$y_{1,3}^{\tau} = 0.0272 \cdot I \cdot R \cdot \exp(R) \cdot \frac{1}{1+\exp\left(-1.370 \frac{n^3 \varepsilon^2}{B}\right)} \cdot P^{-0.576}$$

$$y_{1,4}^{\tau} = 0.1515 \cdot I \cdot R^{3.540} \cdot \exp\left(\frac{I \cdot R}{k^{11.62}}\right) \cdot \frac{1}{1+\exp\left(-1.025\, R\right)} \cdot \frac{1}{1+\exp\left(-2.519 n^3 \varepsilon^2 k\right)} \cdot P^{-0.730}$$

$$y_{1,5}^{\tau} = 0.091 \cdot I \cdot R^2 \cdot \frac{1}{1+\exp\left(-0.548\, n\right)} \cdot \frac{1}{1+\exp\left(-0.557\, n\right)} \cdot P^{-0.539}$$

Table I Prediction for the τ_{ITER} based on the examples up to 0.5 MA.

Model	τ_{ITER}
$y^{\tau}_{1,1}$	20.6747
$y^{\tau}_{1,2}$	48.7173
$y^{\tau}_{1,3}$	94.9609
$y^{\tau}_{1,4}$	0.1175
$y^{\tau}_{1,5}$	4.7231

The values

indicators for these

of the variable ra

Appendix 1. As

interval is too sma

results. The scalin

mathematical for

unrealistic range

ITER, as reported in Table I. On the other hand, the m

first step belong basically to three families, which differ well outside the confidence intervals already in the current range 1.5 MA < I < 2.5 MA, as shown in Figure 8. Including the entries in this interval of currents increases the number of entries to 1428. Performing another iteration of the methodology allows identifying the following set of equations as again the most performing according to the Pareto Frontier:

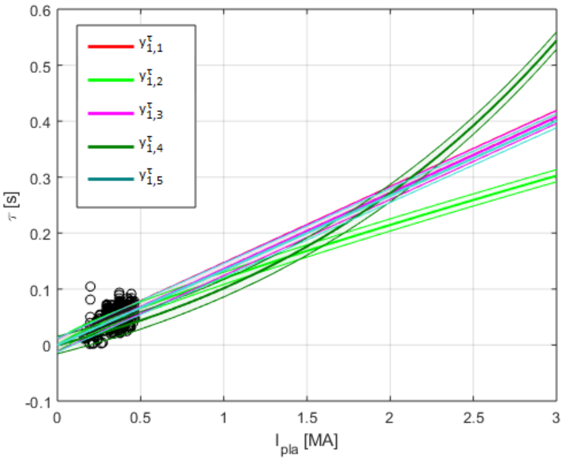


Figure 8 Main families of scaling laws which can be derived on the basis of the experiments with plasma current blow 0.5 MA.

$$y^{\tau}_{2,1}=0.1162\cdot I^{2}\cdot R^{1.953}\cdot\left(\frac{k^{0.553}}{P^{1.2008}}\right)^{3.622}\cdot\frac{1}{1+\exp\left(-0.782\frac{n^{1.465}}{I}\right)}\cdot\frac{1}{1+\exp\left(-0.553\frac{n}{M}\right)}\cdot$$

$$y^{\tau}_{2,2}=0.113\cdot I\cdot R\cdot\frac{1}{1+\exp\left(-7.284\frac{n}{P^{2}}\right)}\cdot\frac{1}{1+\exp\left(-7.381\frac{n}{P}\right)}$$

$$y^{\tau}_{2,3}=0.115\cdot I\cdot R\cdot\frac{1}{1+\exp(-0.998\,n)}\cdot\frac{1}{1+\exp\left(-1.705\frac{n}{P}\right)}\cdot\frac{1}{1+\exp\left(-5.650\frac{I\,R\,k^{2}}{P^{2}}\right)}$$

$$y^{\tau}_{2,4}=0.084\cdot I\cdot R\cdot\frac{1}{1+\exp\left(-26.774\frac{n\,R\,k}{P^{3}}\right)}$$

Even if the obtained scalings show again a wide range of different dependencies, the range of extrapolations to ITER is significantly narrowed to the interval between 2.9 s to about 4 s. On the other hand, the models still vary significantly in their form, suggesting that more data would be required. The details, including the extrapolations to ITER, the plots of the scalings and ranges of the variable are reported in Appendix 2. From the material in the Appendix, it can be seen how the estimates of the various models differ outside of the confidence intervals in the plasma current range between 3.5 and 5 MA. Fortunately in the database there are additionally examples also in this high current range for a final complete set of 1480 inputs. Repeating the procedure including these new entries identifies the following three high quality models:

$$y_{3,1}^{\tau}=0.076\cdot I\cdot R^{2}\cdot k\cdot P^{-1}\cdot\frac{1}{1+\exp\left(-0.201\,n^{2.004}\right)}\cdot\left(\frac{1}{n\cdot P^{1.293}}\cdot\frac{1}{1+\exp\left(0.715\,M\right)}\right)^{-0.1761}$$

$$y_{3,2}^{\tau}=0.1831\cdot I\cdot R^{2}\cdot P^{-0.662}\cdot\frac{1}{1+\exp\left(-0.094\cdot I\right)}\cdot\frac{1}{1+\exp\left(-0.408\cdot n\right)}$$

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

These are to be compared with the pure power law monomials reported in [15] and obtained using the whole of the database and not only the entries the ranges identified by the proposed technique:

$$y_{PL1}=5.55\cdot 10^{-2}I^{0.75}B^{0.32}n^{0.35}M^{0.06}R^{2.0}\epsilon^{0.76}\kappa_{a}^{1.14}P^{-0.62}$$

$$y_{PL2}=5.62\cdot 10^{-2}I^{0.93}B^{0.15}n^{0.41}M^{0.19}R^{1.97}\epsilon^{0.58}\kappa_{a}^{0.78}P^{-0.69}$$

Table II Prediction for the τ_{ITER} obtained at the last iteration of the developed methodology.

Model	τ_{ITER}
$y_{3,1}^{\tau}$	2.55
$y_{3,2}^{\tau}$	4.34
y_{NPL}	2.85
y_{PL1}	3.64
y_{PL2}	3.22

Of course at this stage the procedure has to be stopped because there are no more examples at higher plasma current.

The scaling law y_{NPL} is the one already identified in [7]. From the statistical parameters provided in Appendix 3, it appears very clearly that the exponential and squashing factors improve the scalings significantly compared to the power law monomials. It is also worth noticing that, with the developed tools, much less data is needed to converge on competitive if not superior scaling laws: 1480 instead of 3093. The values of the estimates for ITER are reported in Table II, showing that

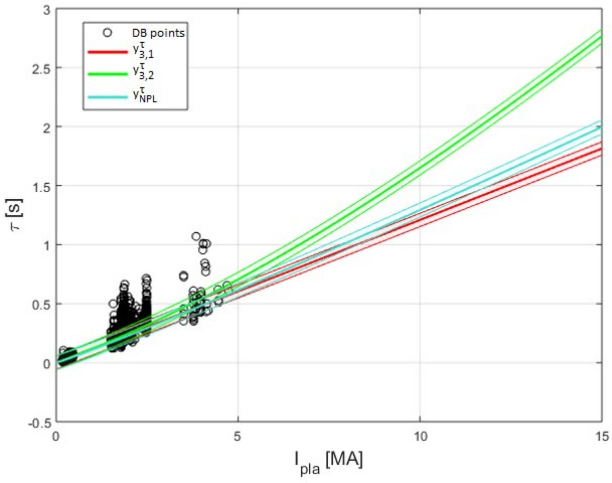


Figure 9 Plots of the scaling laws obtained with all the three current intervals: I<0.5 MA, 1.5 MA< I<2.5 MA, 3.5 MA< I<5 MA.

expecting τ well above 3 s could be on the optimistic side. In any case, as can be appreciated from Figure 9, ITER plasma current will be more than sufficient to clearly discriminate between the proposed models.

5 Conclusions and further lines of investigation

In this paper, an original methodology has been 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 a new range of the regressors can provide real additional information and knowledge. On the other hand, planning experiments in an unexplored zone of the 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 conditions cannot be invoked. As a consequence, care must be taken because the models trained with old data can perform sub optimally and provide even wrong answers. Moreover, contrary to other domains, no obvious assumption can be made to pin down the best model. To operate in such a situation of concept shift, it is important to find a good trade-off in exploring the operational space. The parameter region for the new experiments should not be too close to the previous cases, otherwise their added value would be limited. On the other hand, the extrapolation cannot be too aggressive, penalty the failure of the experiments, because the available models are completely at the loss to provide guidance in the new region of the parameter space; obviously, if there is no relation between the data in the training set and the final applications there is nothing to learn from the past. The methodology proposed in this paper is based on the falsification of the models with experiments in the range of parameters as close as possible to the previous experiments. Of course, this condition can be relaxed or substitute with others if appropriate for the studies to be performed. 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. The procedure is more efficient and reliable than previous approaches such as LCS. Moreover the proposed approach provides better results, at least from a statistical point of view, even with a significant smaller set of examples. On the other hand, it should be emphasized that the example of the energy confinement time is just meant to illustrate the potential of the proposed methodology not to propose a final form of the scaling, because the database is insufficient and also because the issue is to be addressed again with data of metallic devices.

It is worth pointing out that the developed technique can be used not only for experimental design. The approach can also be deployed to focus the analysis on existing databases, which are typically too large for exhaustive investigations. Indeed in many large devices, and particularly at JET given the large warehouse, only a small fraction of the data is actually analysed in detail. The tools developed can be utilised to identify the most relevant entries in the DB to be analysed with specific attention. The same applies to computer simulations, which are nowadays sometimes so complex that a lot of information remains untapped because for lack of resources to investigate the details of their outputs.

With regard to future developments, an important topic to be further developed 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 methodology [17]. Application to scenario integration, with specific attention to the effects of the impurities, is also expected to provide interesting results [18-22].

Aknowledgements

This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014-2018 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission. This work was also partially supported by the Spanish Ministry of Economy and Competitiveness under the Projects Nos. ENE2015-64914-C3-1-R and ENE2015-64914-C3-2-R. The authors would also like to acknowledge the ITPA Transport and Confinement group for their work in preparing the database http://efdasql.ipp.mpg.de /hmodepublic/DataDocumentation/ Datainfo /DB3v13/db3v13.html

References

[1] Vladimir Vovk, Alex Gammerman, and Glenn Shafer "*Algorithmic learning in a random world*" Springer, 2005**,** New York

[2] J.Gama, "*A Survey on Concept Drift Adaptation*" ACM Computing Surveys, Vol. 1, No. 1, Article 1, Publication date: January 2013.

[3] **Sayed-Mouchaweh**, M. "*Learning from Data Streams in Dynamic Environments*"
Springer Briefs in Applied Sciences and Technology 2016

[4] M.Sugiyama and M Kawanabe, "*Machine learning in non-stationary environments*" MIT press, London UK (2012)

[5] Schmid M and Lipson H, Science, Vol 324, April 2009

[6] Koza J.R., "*Genetic Programming: On the Programming of Computers by Means of Natural Selection"*. MIT Press, Cambridge, MA, USA (1992).

[7] Murari, A. et al. "*Non-power law scaling for access to the H-mode in tokamaks via symbolic regression*" (2013) Nuclear Fusion, 53 (4), art. no. 043001, DOI: 10.1088/0029-5515/53/4/043001

[8] Murari A. et al., P. "*Symbolic regression via genetic programming for data driven derivation of confinement scaling laws without any assumption on their mathematical form*" (2015) Plasma Physics and Controlled Fusion, 57 (1), art. no. 014008, DOI: 10.1088/0741-3335/57/1/014008

[9] Murari A. et al "*A new approach to the formulation and validation of scaling expressions for plasma confinement in tokamaks*" 2015 Nuclear Fusion, Volume 55, Number 7 https://doi.org/10.1088/0029-5515/55/7/073009

[10] Murari A. et al, "*Application of symbolic regression to the derivation of scaling laws for tokamak energy confinement time in terms of dimensionless quantities*" Nuclear Fusion **56**, 2, 26005, DOI: 10.1088/0029-5515/56/2/026005

[11] *Akaike, H.* (1973), *"Information theory and an extension of the maximum likelihood principle",* in Petrov, B. N.; Csáki, F., 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: *Akadémiai Kiadó*, pp. 267–281.

[12] Hirotugu A.1974, IEEE Transactions on Automatic Control 19 (6): 716–723, 1974

[13] Schwarz, Gideon E. (1978), *"Estimating the dimension of a model",* Annals of Statistics*,* ***6*** *(2): 461–464,* doi*:*10.1214/aos/1176344136*,* MR 0468014 .

[14] Kenneth P. B and Anderson D. R., 2002, *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. Springer. (2nd ed)

[15] McDonald D. *et al* 2004 *Plasma Phys. Control. Fusion* **46** 519–34

[16] http://efdasql.ipp.mpg.de/hmodepublic/DataDocumentation/ Datainfo /DB3v13/db3v13.html

[17] Murari *et al "Clustering based on the geodesic distance on Gaussian manifolds for the automatic classification of disruptions"* 2013 Nucl. Fusion **53** 033006, doi.org/10.1088/0029-5515/53/3/033006

[18] Peluso E.et al. "*A Statistical Analysis of the Scaling Laws for the Confinement Time Distinguishing between Core and Edge*", Physics Procedia, Volume 62, 113-117, (2015), doi: dx.doi.org/10.1016/j.phpro.2015.02.020

[19] Peluso E. et al, *"A statistical method for model extraction and model selection applied to the temperature scaling of the L-H transition"* (2014) Plasma Physics and Controlled Fusion, 56 (11), art. no. 114001

[20] Murari A. et al*, "Robust scaling laws for energy confinement time, including radiated fraction, in Tokamaks",* Nucl.Fusion Vol.57, num.12 (2017)

[21] Ongena, J. et al "*Towards the realization on JET of an integrated H-mode scenario for ITER*" (2004) Nuclear Fusion, 44 (1), pp. 124-133. DOI: 10.1088/0029-5515/44/1/015

[22] Puiatti, M.E. et al, "*Radiation pattern and impurity transport in argon seeded ELMy H-mode discharges in JET*" (2002) Plasma Physics and Controlled Fusion, 44 (9), pp. 1863-1878. DOI: 10.1088/0741-3335/44/9/305

.

Appendix 1 Data for plasma current below 0.5 MA

This set of discharges includes **702 input**s. The histograms of the main variables are reported in Figure 1.1 and their averages in Table 1.2. The values of the AIC and BIC of the various candidate models are reported in Table 1.1.

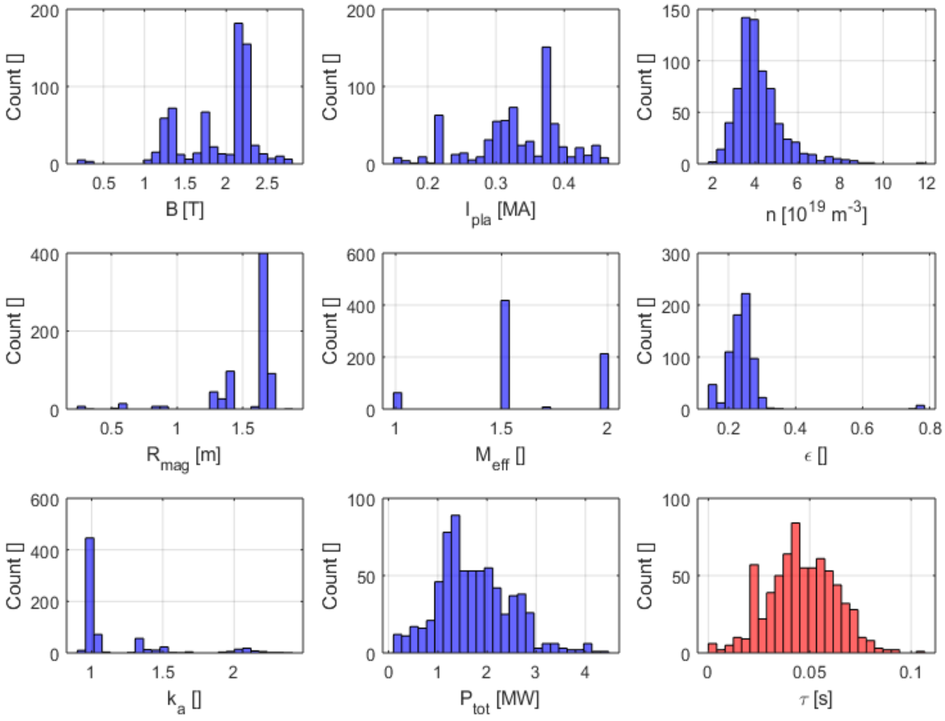


Figure 1.1

The values of AIC and BIC are reported in Table 1.1

Tabella 1.1

Model	k	AIC [10^3]	BIC [10^3]
$y^{\tau}_{1,1}$	32	-6.9789	-6.8328
$y^{\tau}_{1,2}$	26	-6.9264	-6.8071
$y^{\tau}_{1,3}$	29	-6.8704	-6.7374
$y^{\tau}_{1,4}$	17	-6.8248	-6.7467
$y^{\tau}_{1,5}$	42	-6.3828	-6.2714

Tabella 2.2

Variable	Average
B	1.9144
n	4.1996
R	1.5379
M	1.6091
ε	0.2419
k	1.1545
P	1.7319

Appendix 2 Data for plasma current in the intervals 0<Ip< 0.5 MA and 1.5 MA<Ip<2.5 MA

This set of discharges includes **1428 input**s. The histograms of the main variables are reported in Figure 2.1 and their averages in Table 2.3. The values of the AIC and BIC of the various candidate models and their extrapolations to ITER are reported in Table 2.1 and 2.2.

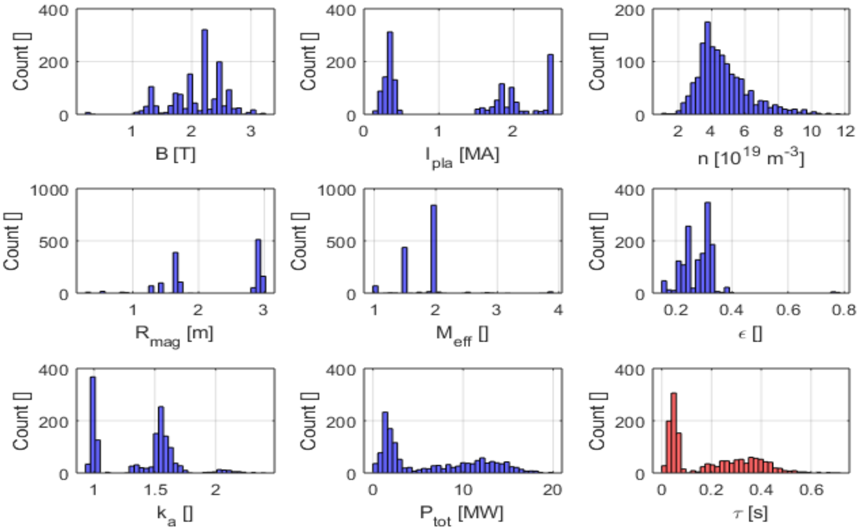


Figura 2.1

Table 2.1

Model	k	AIC [10^3]	BIC [10^3]
$y^{\tau}_{2,1}$	35	-9.7962	-9.6115
$y^{\tau}_{2,2}$	33	-9.6692	-9.4945
$y^{\tau}_{2,3}$	17	-9.457	-9.3699
$y^{\tau}_{2,4}$	19	-9.3558	-9.2554

Table 2.2

Model	τ_{ITER}
$y^{\tau}_{2,1}$	2.8903
$y^{\tau}_{2,2}$	4.1469
$y^{\tau}_{2,3}$	3.2341
$y^{\tau}_{2,4}$	3.9519

Table 2.3

Variable	Average value
B	2.0854
n	4.7019
R	2.2421
M	1.8143
Ɛ	0.2796

k	1.3660
P	6.5544

Appendix 3 Data for plasma current in the intervals 0<Ip<0.5 MA 1.5 MA<Ip<2.5 MA and 3.5 MA<Ip<5 MA

This set of discharges includes **1480 input**s. The histograms of the main variables are reported in Figure 3.1 and their averages in Table 3.2. The values of the AIC and BIC of the various candidate models are reported in Table 3.1.

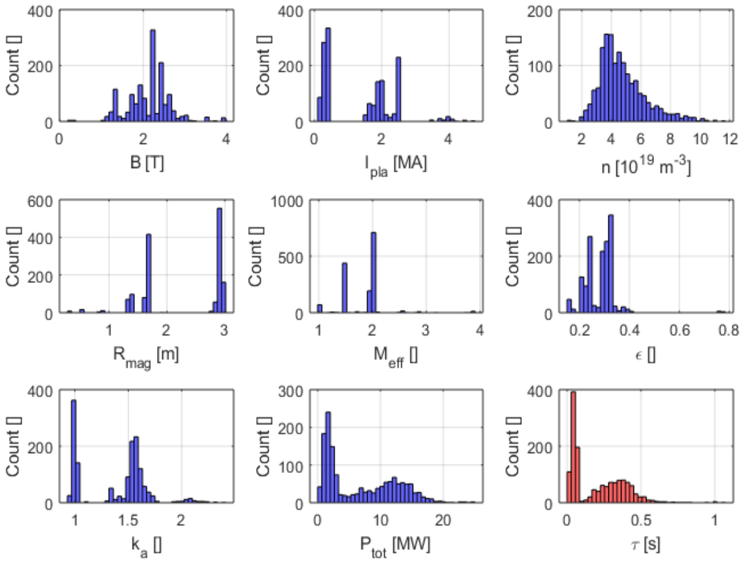


Figura 3.1

Table 3.1

Model	k	AIC	BIC
$y_{3,1}^{\tau}$	29	- 9935.54	- 9781.27
$y_{3,2}^{\tau}$	17	- 9752.85	- 9662.69
y_{NPL}	9	- 9778.58	- 9760.42
y_{PL1}	10	- 9628.72	- 9599.26
y_{PL2}	10	- 9379.43	- 9506.43

Table 3.2

Variable	Average
B	2.1313
n	4.7314
R	2.2644
M	1.8241
Ɛ	0.2818
k	1.3750
P	6.8464