Neural Computing and Applications (2023) 35:469–486
https://doi.org/10.1007/s00521-022-07768-3

ORIGINAL ARTICLE

Information theoretic and neural computational tools for meta-analysis of cumulative databases in the age of Big Physics experiments

A. Murari[1] · M. Lungaroni[2] · L. Spolladore[2] · E. Peluso[2] · R. Rossi[2] · M. Gelfusa[2]

Received: 13 September 2021 / Accepted: 30 August 2022 / Published online: 14 September 2022

Abstract

In the era of Big Data, many scientific disciplines and engineering activities rely on cumulative databases, consisting of many entries derived from different experiments and studies, to investigate complex problems. Their contents can be analysed with much finer granularity than with the usual meta-analytic tools, based on summary statistics such as means and standard deviations. At the same time, not being primary studies, also traditional statistical techniques are not adequate to investigate them. New meta-analysis methods have therefore been adapted to study these cumulative databases and to ensure their validity and consistency. Information theoretic and neural computational tools represent a series of complementary techniques, which can be deployed to identify the most important variables to analyse the problem at hand, to detect whether quantities are missing and to determine the coherence between the entries provided by the individual experiments and studies. The performances of the developed methodologies are verified with a systematic series of tests with synthetic data. An application to thermonuclear fusion proves the capability of the tools to handle real data, in one of the most complex fields of modern physics.

Keywords Meta-analysis · Information theory · Neural computation · Thermonuclear fusion

1 Introduction: meta-analysis and the era of Big Data

The progress of scientific knowledge is based on peer review and reproducibility of experiments. Repetition of studies is therefore intrinsic to the scientific process. Indeed, professional researchers know very well that it is very difficult, if not impossible, to reach definitive conclusions and address complex problems with a single investigation. Consequently, collective knowledge in science is consolidated in two major steps: (a) cumulating individual studies and experiments and (b) the organisation of facts and data in models and theories. The present work aims at contributing to the transition between these two phases in the era of Big Data. It is therefore a contribution to research synthesis [1] and in particular to Meta-Analysis (MA) [2, 3] but with a specific angle (see later).

Meta-analysis can be defined as the statistical synthesis of results from a series of different studies. From a historical perspective, the first meta-analytic publication can be considered the collation of data from several studies of typhoid inoculation by K. Pearson [4]. The 1940 paper *Extrasensory Perception After Sixty Years*, by Duke University psychologists J. G. Pratt and J. B. Rhine, is the first meta-analysis of all conceptually identical experiments, carried out by independent researchers on a specific issue [5]. Even if today the main applications of systematic reviews are probably in the health sciences (particularly epidemiology in the last years), meta-analysis papers on medical treatments were not published before the second half of the fifties. More sophisticated treatments were indeed motivated by problems in education, and the term meta-analysis was coined in 1976 by the statistician G.V Glass [6].

Coming to technical aspects and motivation, when multiple scientific studies have been performed to address

✉ L. Spolladore
luca.spolladore@uniroma2.it

[1] Consorzio RFX (CNR, ENEA, INFN, Università Di Padova, Acciaierie Venete SpA), Corso Stati Uniti 4, 35127 Padua, Italy

[2] Department of Industrial Engineering, University of Rome “Tor Vergata”, via del Politecnico 1, Roma, Italy

the same question, meta-analysis is a statistical set of techniques aimed at combining their results, to reach more solid conclusions. The inadequacies of the original investigations are limited statistical evidence and uncertainties in the data or any form of artefact effects, ranging from poor design to sampling errors. The main idea behind meta-analysis consists of taking advantage of statistical methods to derive pooled estimates, better approximating the underlying common truth.

Existing methods for meta-analysis have been conceived to provide a weighted average of the summary statistics, such as means and standard deviations, already derived in the individual studies. The main differences between these techniques relate to the methods for allocating the weights and to the procedures to associate the confidence intervals to the obtained point estimates [7]. The key benefit of meta-analysis is the aggregation of information, which typically results in higher statistical powers and more reliable point estimates, compared to those that are possible to derive from any individual study. In addition to this main objective, MA techniques can also provide other interesting outputs, such as contrasting different studies and identifying file drawer problems [8]. The importance of meta-analysis today can be appreciated by the scheme of ranking evidence GRADE, constructed by the Grading of Recommendations Assessment, Development and Evaluation Working Group, nowadays used by more than 50 organisations worldwide [9]. Meta-analysis and systematic reviews are ranked even above Randomised Control Trials (RCT), once considered the gold standards of scientific investigations.

The incredible developments in sensors, data acquisition, storage and computational power allow nowadays to build databases of different nature than the ones combined by traditional MA. In particular, in many fields, namely the physical sciences and engineering, it is possible to have access to the original entries, the individual results of the experiments, and not only to their summary statistics. These collections of data are not primary studies, because they contain data from different experiments. At the same time, they are not the traditional subject of meta-analysis either, because they are not limited to the summary statistics of the individual studies. For lack of a better name, these databases, consisting of individual entries, coming from different primary experiments or studies without any statistical manipulation, will be called "*cumulative databases*" in the rest of the paper. It is possible to argue that some of these cumulative databases were used at the beginning of modern science in the seventeenth century, particularly in the development of the heliocentric view of the solar system and the formulation of the Kepler laws.

Since the results of individual experiments, and not summary statistics, are available, these cumulative databases have been typically analysed with the techniques of traditional statistics and not meta-analysis. It is the basic contention of this work that such an approach is insufficient and can result in completing misleading conclusions. On the contrary, a series of checks can be applied to the results of the individual experiments, to improve the entire inference process. Such a task requires developing new techniques to perform meta-analysis analysis at the lower level than traditional methods: at the level of the individual entries and not summary statistics. This additional level of meta-analysis is indispensable, to take full advantage of the information available at the level of individual entries without being misled.

In more detail, the checks to be performed, when dealing with cumulative databases, relate mainly to three aspects: the selection of the most suitable quantities to analyse a certain phenomenon, the assessment of the data consistency and the evaluation of the final results. The main objective of this work is indeed to propose a methodology and a series of tools, to perform quality checks on the quantities included in the databases available and their coherence. None of these aspects is properly addressed neither by traditional statistics nor by usual meta-analysis techniques. Only psychometric meta-analysis deals with some of these issues but with a much more limited scope, as will be discussed in more detail in Sects. 6 and 7 [10]. The developed techniques can be considered a preliminary step to any form of actual scientific exploitation, be it meta-analysis, modelling or causality detection. They basically apply a meta-analytic mind-set to the entries derived from primary but different experiments. Neglecting this preliminary phase can often compromise the entire inference process. To fix the ideas, the discussion in the following is particularised for the case of regression, probably the most interesting from a scientific perspective.

With regard to the structure of the paper, next section presents the main rationale behind the importance of and the issues posed by cumulative databases for sound research synthesis. The main families of techniques, refined and deployed for their analysis, are described in Sect. 3 and Sect. 4; they are information theoretic tools and neural networks respectively. The potential of the proposed approach is substantiated in Sect. 5, with examples from a battery of systematic numerical tests performed with synthetic data. Section 6 is devoted to discussing the very important practical issue of overfitting due to spurious regressors. This problem can compromise the conclusions derived from a naïve application of traditional model selection criteria, but can be efficiently addressed with the methodology proposed in this paper. A quite challenging real time example from Big Physics, namely thermonuclear fusion, is reported in Sect. 7. This example shows the importance of performing meta-analysis at the level of the individual experiments, when dealing with cumulative databases. The conclusions are drawn in the last section of the paper.

2 The issues posed by cumulative databases: objectives of the analysis and overview of the main tools

In this work, it is assumed that cumulative databases, consisting of the collection of entries from different experiments, have been built. This means that the data are assumed to be already validated and that the basic information about the various quantities, such as error bars and calibration factors, are available for the entries corresponding to each study or experiment. The difficulties and subtleties related to the collection and validation of the individual entries, though extremely important, are therefore not within the scope of the present discussion and should be left to the scientists conducting the individual experiments. It is also assumed that the objective of the investigation is to determine the dependence of a quantity Y, called the dependent variable, on a series of independent quantities $X_1 \ldots X_n$, also called regressors. The number of individual studies, aggregated in the database (DB), is indicated with the letter k.

At this level of database validation, the main aspects to be addressed, for proceeding efficiently with the scientific exploitation of the data, relate to the relevance and consistency of the entries. In more detail, intuitively, prior to the modelling and further analysis of the data, the following questions should be answered:

(1) Which are the most important quantities to analyse the database?
(2) Are the quantities available sufficient to draw the conclusions of the study or some essential information is missing?
(3) Are the data from the various experiments or studies sufficiently coherent to grant the aggregation of the data?

In the rest of the paper, the tasks made explicit by these questions will be called *regressor selection*, *sufficiency* and *global consistency*, respectively. The first two have to be assessed individually for each of the k studies included in the cumulative database. Global consistency refers to the coherence between the different studies.

The importance of *regressor selection* is probably obvious. Among the quantities belonging to each experiment or study in the database, it is essential to determine which ones are really important to model the dependent variable. Excluding quantities, which have spurious correlations with the dependent variable, is essential to avoid overfitting, as will be shown in more detail later. *Sufficiency* consists of determining that the regressors, identified with the help of the feature selection tools, can properly explain the trends of the dependent variable. It should be noted that in many modern applications, linear regression is not an option, because the phenomena to be studied are typically nonlinear. Nonlinear versions of the unexplained variance have therefore to be implemented. *Global consistency* refers to the fact that criteria are needed to avoid mixing apples with pears. Basically, it must be verified that the same essential mechanisms are at play in all the studies included in the cumulative database, prior to extracting global models. To this end, it must first be verified that the regressors in the individual studies are sufficient to describe the dependent variable and that they are the same in all studies. It should also be checked that the dependence between the dependent and the independent variables is not too heterogeneous (according to quantitative criteria that will be specified later).

The proposed techniques, to answer the previously mentioned fundamental questions, are covered in the next sections. Profiting from the information in the original entries, they basically perform a forensic activity preliminary to the starting point of traditional meta-analysis, which assumes independent variables and effects sizes of the primary studies as given. The tools developed belong to two main classes; information theory and neural computation. Information theoretic quantities, such as Shannon entropy and mutual information, are quite consolidated and have been used in the past in many fields for the investigation of nonlinear correlations between quantities. Neural networks are powerful tools, which have witnessed many successes in the last years and can be used profitably also for the tasks identified in this work. It is worth pointing out that the two families of tools are mathematically completely different. Their combination is therefore particularly important because, if they provide coherent results, the investigators are authorised to have much stronger confidence in the conclusions.

3 Information theoretic tools for the investigation of cumulative databases

In this section, the main information theoretic techniques implemented, to analyse large cumulative databases, are introduced. First, purely information theoretic indicators, which can be deployed directly on the data, are overviewed (Sect. 3.1). They are first utilised for feature extraction, to select the most important independent variables in a database to regress a certain dependent quantity; two different approaches are introduced in Sects. 3.2 and 3.3. Then they can also be adapted to determine the sufficiency of the quantities in the DB to study the independent variable (Sect. 3.4). The techniques proposed to assess the global consistency are described (Sect. 3.5). The issue of assessing the statistical significance of the indicators is addressed in Sect. 3.6

3.1 Basic information theoretic quantities

The first information theoretic quantity [11], required to implement the meta-analytic techniques proposed in this work, is the Gibbs-Shannon entropy H:

$$H(X) = -\sum\nolimits_{x} P_x \log P_x \tag{1}$$

The higher the value of H, the higher the uniformity of the corresponding probability distribution function, whose values are indicated with P_x.

The other essential and well understood quantity is the Mutual Information (MI) [11]:

$$\mathrm{MI}(X, Y) = -\sum\nolimits_{x}\sum\nolimits_{y} P_{xy} \ln\left(\frac{P_{xy}}{P_x P_y}\right) \tag{2}$$

where P_{xy} is the joint probability distribution function (pdf) of the random variables X and Y. Being fully nonlinear, contrary to the Pearson correlation coefficient, the MI is well suited to extract, from a given database, the best features, i.e. the best regressors X_i to reproduce the desired dependent variable Y.

The third important information theoretic indicator, used in the rest of the paper, is the conditional mutual information, defined as:

$$\mathrm{CMI}(X, Y|Z) = -\sum_{z} P_z \sum\nolimits_{x}\sum\nolimits_{y} P_{xy|z} \ln\left(\frac{P_{xy|z}}{P_{xz} P_{yz}}\right) \tag{3}$$

Alternative continuous versions of the just defined quantities are available and do not entail any major difficulty for the applications discussed in this work. In any case, a detailed mathematical introduction and intuitive explanation of these information theoretical concepts can be found in [11].

3.2 Regressors selection for each individual study in the DB: the conditional mutual information approach

To investigate a database, the first step is the identification of the most important regressors, i.e. the most important quantities to regress the dependent variable. The first regressor X_1 to be selected is the one maximising the mutual information between itself and the dependent variable, MI (X_1, Y). The following X_i regressors have to be chosen with much more care, because they could have a very high MI with the dependent variable only because they are highly correlated with one or more previously selected regressors. They could therefore be very redundant and bring no additional information, even if they present a high MI(Y, X_i). The quantity to maximise is therefore the mutual information between the additional candidate variable X_{ac}, conditional on the regressors already selected. Assuming that the n quantities $X_1 \ldots X_n$ have already been identified as the most important independent variables, the indicator to maximise to select the next one is:

$$\mathrm{CMI}(Y, X_{\mathrm{ac}}|X_1 \ldots\ldots\ldots X_n) \tag{4}$$

The stopping criterion is easy to define; when $\mathrm{MI}(Y, X_{\mathrm{ac}}|X_1 \ldots\ldots\ldots X_n)$ becomes negligible, or negative, for all the remaining variables of the database, only the previously selected quantities have to be retained.

3.3 Regressors selection for each individual study in the DB: the redundancy approach

The solution proposed in the previous subsection is quite elegant, but it requires the evaluation of the conditional mutual information, which is based on the estimate of higher order probability distribution functions. When the number of variables to condition on increases, this estimate is affected by the problem typically called the curse of dimensionality. The number of entries, required for reliable density estimation, tends to diverge. The alterative proposed in this subsection needs only the evaluation of the mutual information between two quantities and can therefore become computationally preferable [12].

The first regressor X_1 can be selected again simply by maximising the mutual information between itself and the dependent variable, MI (X_1, Y). Again the following X_i regressors have to be chosen with greater attention, because they could be highly correlated with one or more previously selected regressors.

The technique, to choose the independent variables after the first one, is based on the definition of redundancy RD between a variable X_i and the set S_{ps} of previously selected ones X_j:

$$\mathrm{RD}(Xi, \mathrm{Sps}) = \sum_{X_j \in S_{\mathrm{PS}}} \mathrm{MI}(X_i, X_j) \tag{5}$$

Therefore after the first, all the next regressors X_j selected are the ones, which maximise the relevance RL defined as:

$$\begin{aligned}\mathrm{RL}(Xi, Y) &= \mathrm{MI}(Xi, \mathrm{Y}) - \mathrm{RD}(Xi, Sps)\\ &= \mathrm{MI}(Xi, Y) - \sum_{X_{j\in S_{\mathrm{PS}}}} \mathrm{MI}(X_i, X_j)\end{aligned} \tag{6}$$

When the relevance of a certain X_i is negligible or negative, that quantity does not bring any additional information and should not be included in the set of regressors [12]. The relations between RD, MI and RL are shown graphically in Fig. 1. The procedure to select the best regressors can therefore be written in pseudo-code as follows:

(1) For the first regressor, select the X_i with highest mutual information with the dependent variable $MI(X_i,Y)$

(2) For each of the following candidate regressors, select the X_i with highest relevance RL (X_i,Y)

(3) Stop when the relevance becomes negligible or negative for all the remaining candidate regressors

It is important to notice that the algorithm just described requires only the calculation of binary mutual information indicators and it is therefore numerically much more efficient than the approach introduced in the previous subsection.

3.4 Sufficiency of the regressors in each individual study in the DB

The stopping criteria, introduced at the end of the last two subsections, have been conceived to ensure that variables, which do not contribute any additional information, are not included in the list of regressors. This is very important, as discussed in Sect. 6, because including "spurious" quantities in the models can result in significant overfitting, difficult to detect later in the analysis. On the other hand, ensuring that redundant quantities are excluded does not imply that the selected regressors are sufficient to fit the dependent variable. Important information could be missing, due to the fact that additional quantities influence Y but are not in the DB. To assess whether the candidate regressors contain enough information to fit the dependent variable, the algorithms proposed in Sect. 3.2 would suggest to determine whether CMI($Y|X_1,..X_n$) is compatible with the uncertainties of the DB entries. This means basically to verify that:

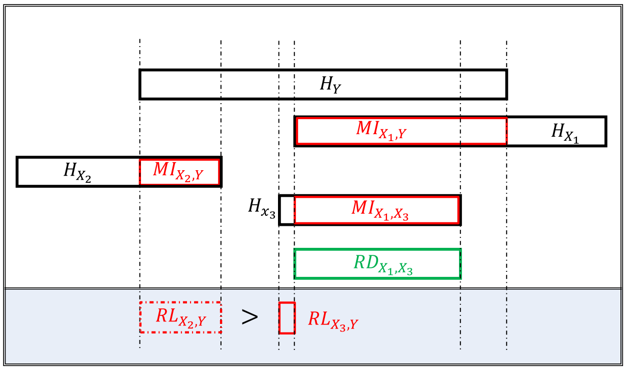

Fig. 1 A visual representation of mutual information and redundancy. X_1 is chosen first due to its high mutual information MI with Y. Given the high redundancy RD between X_3 and X_1, X_2 is chosen next, even if its MI with Y is smaller, because its relevance RL is higher (as shown in the last row)

$$H(Y) - \mathrm{CMI}(Y|X_1, ..X_n) \approx H_{\mathrm{Res}}(Y) \tag{7}$$

where the *residual entropy* H_{Res} *(Y)* quantifies the remaining entropy due to the intrinsic uncertainty of the dependent variable, once all the relevant regressors have been taken into account.

In the case of the approach introduced in Sect. 3.3, one should ascertain whether the following equality is valid:

$$\begin{aligned}H(y) - \sum_{X_{i\in S_{\mathrm{Reg}}}} \mathrm{RL}(X_i, Y) = H(y) &- \sum_{X_{i\in S_{\mathrm{Reg}}}} (\mathrm{MI}(X_i, Y)\\ &+ \sum_{X_{j\in S_{\mathrm{Reg}}}} \mathrm{MI}(X_i, X_j))\\ &\approx H\mathrm{Res}(Y)\end{aligned} \tag{8}$$

where S_{Reg} is the set of the quantities selected as regressors with the procedures described in the previous section. Indeed, if all and only the important regressors have been selected, and there are no quantities missing in the database, the entropy left after considering the regressors should be equal to the *residual entropy* $H_{\mathrm{Res}}(Y)$, compatible with the level of intrinsic uncertainty affecting the dependent variable Y.

3.5 Global consistency

In the case of cumulative databases, in addition to the relevance and sufficiency analysis, it is very important to assess also the consistency of the dependencies in the various experiments or studies. To this end, first the analysis described in the previous subsections should be particularised for each individual study, to verify that the set of regressors is sufficient to model the dependent variables in all of them. Moreover, it should also be checked that the most relevant variables are the same, in the same order of priority, in all the individual studies. Lastly, the coherency between the dependencies is also to be considered. In this perspective, it is very informative to calculate the mutual information between the regressors and the dependent variable, to assess whether they are consistent. To this end, the mutual information indicators MI (X_i,Y) are expected to be the same, within the confidence intervals, for all the k individual studies included in the cumulative database. The individual studies, which present significantly different dependencies between the regressors and the dependent variable, should indeed be considered separately or at least adequate motivations should be adduced to keep them in the global study.

3.6 Statistical significance of the indicators

The procedures, proposed in the previous sections, are well established, being based on consolidated information theoretic quantities. Their properties are typically guaranteed in the limit of an infinite number of examples. However, in practice, the indicators cannot be calculated in ideal conditions. Two aspects have to be taken into account when dealing with experimental databases, namely that the number of entries is limited and that the data can be affected by various forms of noise. In the perspective of the present work, the main consequence of these effects is that they render potentially difficult the interpretation of the numerical values of the calculated quantities (entropy, mutual information etc.). To overcome this difficulty, two different measures can be adopted, one for the calculation of the entropy and one for the mutual information.

For the residual entropy, one has to assume that the distribution function of the uncertainties is known. In this hypothesis, the residual value of the entropy H_{Res} is the one due only to the intrinsic uncertainties affecting the dependent variable (which can be due to noise, thermal fluctuations or any other source). Therefore only the difference, between the entropy of the dependent variable and the sum of the relevancies of the selected regressors, is the quantity to be considered to assess sufficiency.

In the case of MI and CMI, a good practice consists of calculating a sort of baseline value for the indicators, by randomising one of the two quantities involved. The obtained values are the reference $\mathrm{MI}_{\mathrm{Res}}$ and $\mathrm{CMI}_{\mathrm{Res}}$, against which to test the actual *MI* and *CMI* of the original data. Only the difference between the mutual information indicators and their residual values is statistically relevant. For the quantification of the statistical significance, traditional goodness-of-fit tests can be deployed; the most widely used are the Chi-squared, Anderson Darling and Kolmogorov–Smirnov [13].

Examples and a more detailed discussion about the inference process with information theoretic indicators, in the presence of the typical practical limitations of the data, are provided in Sects. 5 and 6.

4 Neural computational tools for the investigation of cumulative databases

Artificial Neural Networks (ANN) constitute a powerful alternative to the checks based on the information theoretic criteria presented in the previous section. In particular, multilayer perceptrons, with at least one layer of nonlinear activation functions, are known to be universal approximators (see Fig. 2). They can therefore be deployed to fit the dependent quantity, using the potential regressors as inputs. For the task in hand of assessing the consistency of cumulative databases, they can be trained, in a supervised way with traditional back propagation, to reproduce the dependent quantity on the basis of the potential regressors provided as inputs. The rest of this section is devoted to describing in detail how they can be deployed, to address the three issues of regressor selection, sufficiency and global consistency. In Sect. 4.4, the practical aspects of dealing with real data are discussed.

4.1 Regressors selection for each individual study in the DB

To identify the best regressors, by minimising the redundancy between them, the ANN can be trained to reproduce the dependent variables with increasing number of inputs, until adding regressors does not improve performance. In detail, the following procedure can be implemented:

(1) For the first regressor, select the $\mathrm{X_i}$ which minimises the output error of the networks, when they are trained with only one input

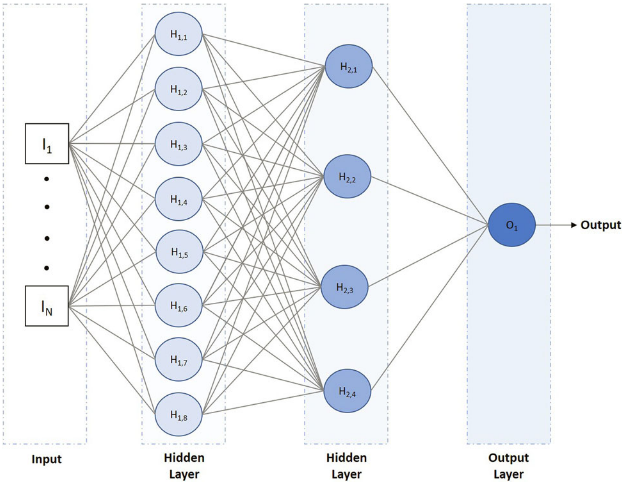

Fig. 2 The architecture of a traditional ANN for the analysis reported in the present work. The activation function of the neurons in the hidden layers is the sigmoid. The activation functions of the neurons in the input and output layers are linear functions

(2) For each of the following regressors, select the X_i that, if added to the previously selected inputs, maximises the reduction in the output error, if compared to the one obtained with the previous set of inputs

(3) Stop when adding any of the remaining candidate regressors to the input list does not result in a statistically significant reduction of the output error.

To increase the confidence in the results, it is advisable to implement the previous procedure also in reverse order. To this end, the first network is trained with all the available candidate regressors and then one is excluded at the time. The quantities, whose exclusion from the list of inputs has no statistically significant detrimental effect on the residuals, are considered not relevant and are eliminated from the list of regressors. The procedure is continued until removing any additional independent variable would result in an appreciable degradation in the network performance.

For the quantification of the error statistical significance, again traditional goodness-of-fit tests can be deployed [13].

In the studies reported in this paper, for clarity of explanation, only some simple indicators, widely known and used, are discussed. To this end, the training of the networks is typically repeated N times, to obtain a statistically relevant distribution of the results. This process is repeated for any new variable candidate to be added to (or eliminated from) the list of relevant regressors. The root main square error of the residuals and their standard deviations, obtained adding the new quantity are therefore indicated as:

$$\overline{\mathrm{RMSE}_{\mathrm{new}}} = \frac{1}{N}\sum\nolimits_i \mathrm{RMSE}_{\mathrm{new},i},\ \sigma_{\mathrm{new}} = \sqrt{\frac{\sum_i \left(\mathrm{RMSE}_{\mathrm{new},i} - \overline{\mathrm{RMSE}_{\mathrm{new}}}\right)}{N-1}}$$

where $\mathrm{RMSE}_{\mathrm{new},i}$ is the RMSE of the fit using the set including the additional candidate regressor. To evaluate the impact of the new quantity on the dependent variable, the quality of the residuals is to be compared with the one obtained with all the previous regressors without the new candidate. This impact can be quantified with the help of the indicator $Z_{\mathrm{score,new}}$:

$$Z_{\mathrm{score,new}} = \frac{\overline{\mathrm{RMSE}_{\mathrm{prev}}} - \overline{\mathrm{RMSE}_{\mathrm{new}}}}{\sqrt{\sigma^2_{\mathrm{prev}} + \sigma^2_{\mathrm{new}}}} \tag{9}$$

where the subscript “prev” indicates the results obtained in the previous step (i.e. the one without the new variable).

In the case the approach of subtracting, instead of adding, one quantity at a time from the list of regressors is adopted, and the indicator is calculated as:

$$Z_{\mathrm{score,remove}} = \frac{\overline{\mathrm{RMSE}_{\mathrm{removed}}} - \overline{\mathrm{RMSE}_{\mathrm{prev}}}}{\sqrt{\sigma^2_{\mathrm{prev}} + \sigma^2_{\mathrm{removed}}}} \tag{10}$$

where $\mathrm{RMSE}_{\mathrm{remove}}$ is the RMSE of the fit using the set, from which one candidate regressor has been removed (overbars again indicate averages).

4.2 Sufficiency of the regressors in each individual study of the DB

To determine whether the selected regressors are sufficient to model the independent quantity, or whether additional inputs would be necessary, again the level of output error can be analysed. If a sufficient set of regressors has been identified, the ANN should be able to reproduce perfectly the dependent variable and the errors in the output should be due only to the intrinsic uncertainties in Y (due to noise, fluctuations etc.). If the level of noise is negligible, the following R^2 indicator should tend to 1:

$$R^2 = 1 - \frac{\mathrm{RMSE}_{\mathrm{Allselvar}}}{\sigma(Y)} \tag{11}$$

where $\mathrm{RMSE}_{All\ sel\ var}$ indicates the root main square error of the residuals for the complete set of selected variables. Therefore R^2 quantifies how much uncertainties in the dependent variable are explained by the regressors.

In the eventuality, typically more relevant in practice, that the uncertainties are appreciable, again more sophisticated goodness-of-fit tests can be implemented to assess the quality of the ANN output. The null distribution of these statistical tests is calculated assuming that the outputs are drawn from the reference distribution, the one of the noise affecting the data in the present case. The goodness-of-fit tests previously mentioned, or others equivalent, can be deployed with this objective. Most goodness-of-fit tests typically provide a Z score as output, which can be expressed in such a way that the lower its value, the closer the residuals to the pdf of the null hypothesis (the one of the noise in our application). If the output, including all the selected inputs, does not approximate sufficiently well the noise statistics, additional quantities should be added to the DB to model the dependent variable.

4.3 Global consistency

Again, the first step, in the assessment of the coherence between the various contributions to a cumulative database, consists of verifying that the k subsets of entries, provided by the individual studies, satisfy the condition of sufficiency in terms of the same set of regressors. These two aspects, that the individual studies contain enough information to explain the dependent variable in terms of the same regressors, are a natural outcome of the tools described in the previous subsections. More delicate, as for the case of the information theoretic tools, is to ascertain consistency.

To test the consistency of the dependencies between the regressors and the dependent quantity, a set of k networks, each trained with the entries of an individual study, can be deployed. They can then be given as input one regressor scanned over its relevant range, while the other independent variables are kept constant. This process is repeated for each regressor and the values of the output stored. The obtained outputs can then be compared, and if they do not show the same trends, it is easy to determine which variables have a different effect on the output in the k individual studies. The individual studies, whose independent variables present a statistically significant different dependence on the regressors than the rest, should not be included in the global database without adequate justifications. This procedure is exemplified in Sect. 5.4.

4.4 Robust inference with neural computational tools

The neural computational techniques previously described are very powerful, but they are also completely data driven and therefore are affected by the limitations and imperfections of the data available in practice. Not requiring the identification of the data pdf, they are less demanding than the information theoretic tools, in terms of the amounts of data required to obtain reasonably robust conclusions. On the other hand, of course, great care must be taken anyway to ensure the quality of the results and of the inference process. The most important way to achieve reliability consists of adapting the procedures, used for ensembles of classifiers [14]. The tools are to be deployed several times, each one with slightly different data as inputs. The slightly different datasets can be derived from the original DB by the traditional tools of bagging or adding different realisations of the noise [15]. The final quantities, to base the inference process on, will then be appropriate summary statistics of the individual results. Basically, it is sufficient to verify that the estimates of the ANN present high values of goodness-of-fit indicators, such as the R^2. Again representative examples and a more detailed discussion about the inference process with neural computational tools, in presence of the typical practical limitations in the data, are provided in Sects. 5 and 6.

5 Examples and benchmarking of the proposed techniques with synthetic data

This section is meant to exemplify the potential of the previously described procedures with the help of synthetic data. The first subsection introduces the rationale behind the choice of the cases described in detail. The following three subsections report the results for the assessment of the three main aspects of regressor selection, sufficiency and global consistency.

5.1 Examples of some functional dependencies important in practice

The techniques and methodologies described in the previous sections have been successfully tested using databases of different compositions. All the main functional dependencies have been investigated, from linear and additive, to nonlinear, multiplicative and exponential. Different noise statistics have been analysed, from Gaussian and uniform to Poisson and gamma. The developed tools have proved to cope well with all these situations. Of course the requirements, in terms of both amounts of data and computational resources, become more stringent the more nonlinear the phenomena to investigate and the more exotic the noise statistics.

In the rest of this section, for the sake of brevity, the potential of the developed techniques is exemplified for the case of power laws. This is motivated by two main reasons. First, power laws are a very important class of nonlinear interactions, which are widely studied also with the help of cumulative databases. Secondly, the challenging real-life case, described in detail in Sect. 7, is typically modelled with this class of functional dependency.

For simplicity's sake, the cumulative database of synthetic data, analysed in this section, is supposed to be composed of the results coming from only two individual experiments or studies. The extension to higher numbers of individual contributions is immediate both conceptually and practically. The noise statistics considered is Gaussian, by far the most relevant in practice. Again, different noise statistics can be handled equally effectively.

The synthetic data have been generated as described in the following, for all the functional dependencies investigated. The independent variables X_i in each individual study are 6, sampled uniformly in the interval between 1 and 10 (10^3 points for each variable). Namely, the regressors are sampled in the intervals:

$$X_{1a} = \mathcal{U}(1,10); X_{2a} = \mathcal{U}(1,10);$$
$$X_{3a} = \mathcal{U}(1,10); X_{4a} = \mathcal{U}(1,10);$$
$$X_{5a} = \mathcal{U}(1,10); X_{6a} = \mathcal{U}(1,10)$$

$$X_{1b} = \mathcal{U}(1,10); X_{2b} = \mathcal{U}(1,10);$$
$$X_{3b} = \mathcal{U}(1,10); X = \mathcal{U}(1,10);$$
$$X_{5b} = \mathcal{U}(1,10); X_{6b} = \mathcal{U}(1,10)$$

where $\mathcal{U}$ indicates the uniform distribution. The quantities belonging to the first study are indicated with the subscript a and the ones belonging to the second with the subscript b. The individual and cumulative databases investigated are therefore:

$$X_a = [X_{1a}X_{2a}X_{3a}X_{4a}X_{5a}X_{6a}]$$

$$X_b = [X_{1b}X_{2b}X_{3b}X_{4b}X_{5b}X_{6b}]$$

$$X = [X_aX_b]$$

$$Y = [Y_aY_b]$$

The functional dependencies of Y_a and Y_b are specified in each case in the following subsections.

Noise of Gaussian distribution has been added to both the dependent and the independent variables. The results reported in this work have been derived for noise of zero mean and standard deviation equal to 10% of the entries amplitude.

Three important representative cases are discussed in the rest of this section. In the first database, the two individual studies are similar, but the dependent variable does not depend on the same regressors. Therefore interpreting the entire DB as a whole would not be correct. The second set of synthetic data is supposed to address a very important point related to sufficiency: the danger of missing important quantities with the related risk of using spurious ones in their place. In this case, not testing for sufficiency, at the global and individual level, would again result in completely wrong results. In the third collection of hypothetical studies, the same quantities are relevant for both studies, but their functional dependencies are different. Again only the type of analysis, proposed in this work for testing the global coherence of the studies, can detect the fact that interpreting the data as coming from a single individual experiment would be inappropriate.

5.2 Regressor selection

In this case, the functional dependencies investigated are:

$$Y_a = X_{1a} \cdot X_{2a} \cdot X_{3a} \tag{12}$$

$$Y_b = X_{1b} \cdot X_{2b} \cdot X_{4b} \tag{13}$$

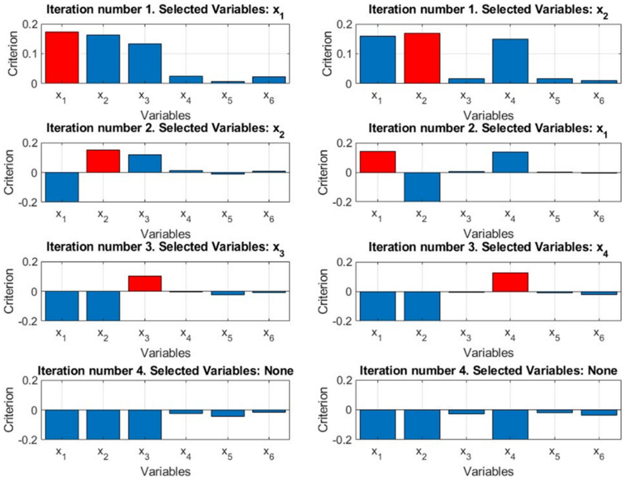

Fig. 3 Regressor selection for the data generated with Eqs. 12 and 13. Left-hand side column: sequence of selected variables with the relevance criterion for the first contribution a. Right-hand side column: sequence of selected variables with the relevance criterion for the second contribution b. The red bar indicates the variable selected at each step of the procedure

As in the other examples, the subscript number indicates the regressor and the subscript letter the single contribution to the global database.

The two individual studies are similar, but the dependent variable is not influenced by the same regressors in the two cases. Interpreting the entire DB as a whole would therefore not be correct.

On the other hand, both series of tools, information theoretic and neural networks, if applied to the cumulative database [Y_a Y_b], identify that the necessary variables are X_1, X_2, X_3, X_4. Given the presence of a low but significant level of noise, both the information theoretic indicators and the R^2 of the networks have problems identifying that something is amiss. Indeed the difference between these indicators and the ones expected in the case of a single database, depending only on X_1, X_2, X_3, X_4 are practically within the confidence intervals of the statistical indicators. On the other hand, the proposed tools, applied to the individual contributions of the database, manage to detect very well that the two systems are different, since the dependent variables depend on different regressors. In Fig. 3, the variable selection sequence, using the relevance criterion of Eq. 6, is reported. From inspection of the histograms shown, it is easy to see how the relevance criterion manages to correctly identify that the relevant regressors for contribution a are X_1, X_2, X_3, and that for contribution b the independent quantities of interest are X_1, X_2, X_4.

The indications of the conditional mutual information are perfectly coherent with the results obtained with the redundancy approach, producing the same selection of the variables as reported in Fig. 3.

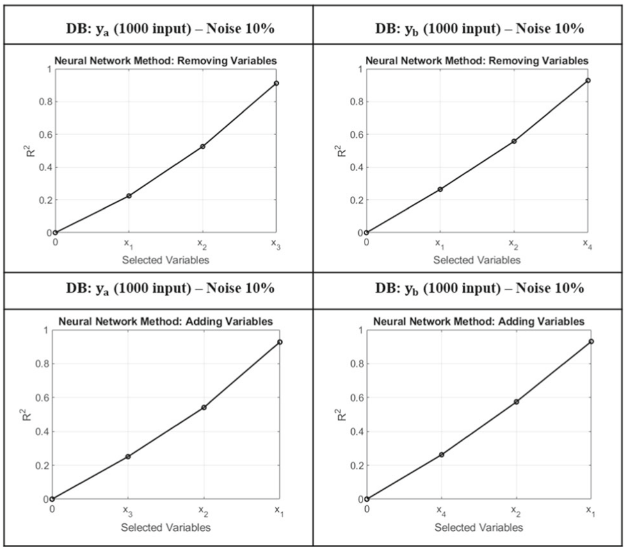

Fig. 4 Regressor selection for the data generated with Eqs. 12 and 13. Sequence of variable selection with the two approaches of removing and adding variables, explained in Sect. 4.1. The results are fully coherent with those obtained with the information theoretic criteria. The important variables are selected in the right order. On the ordinates, R^2 shows how much uncertainty in the dependent variable is explained by the regressors

Exactly the same conclusions can be derived also from the neural networks, whose results are shown in Fig. 4, using both approaches of adding and removing variables. Therefore in this case, the proposed detailed analysis of the individual primary studies, contributing to the cumulative database, would provide essential information for a consistent interpretation of the data available and for the subsequent meta-analysis and scientific exploitation.

5.3 Sufficiency

The issue of the global DB not containing all the relevant quantities, to perform the desired investigation and inference, is not to be underestimated. Among other things, as described in more detail in Sect. 6, it can render the analysis particularly vulnerable to overfitting. The nature of the problem can be easily appreciated by the following example. The two individual studies, contributing to the cumulative database, are assumed to generate data according to the equations:

$$Y_a = X_{1a} \cdot X_{2a} \cdot X_{3a} \tag{14}$$

$$Y_b = X_{1b} \cdot X_{2b} \cdot X_{3b} + 20X_{6b} \tag{15}$$

In this case, the two data generating equations depend exactly in the same way on X_1, X_2, X_3 but Y_b is also influenced by X_6. For the investigation of the issues posed by the lack of sufficiency, it is assumed that X_6 is not included in the variables contained in the database.

This problem is very difficult to interpret at the global level. Even in the present hypothetical case of only two studies with the same number of entries, inspection of both the residual entropy and the R^2 indicator does not reveal that some regressors are missing. For example, as reported in Fig. 5, for Eqs. 14 and 15 R^2 reaches almost 0.86, a value not incompatible with the noise level (the case shown for the approach of adding one regressor at the time is confirmed by the alternative of progressively deselecting quantities). On the contrary, the analysis particularised for the individual subsets of data reveals immediately that the sufficiency condition is satisfied by the individual study *a* and that it is the second set of entries, which is not complete and would need integrating with additional quantities. Indeed, the R^2 indicator assumes a value of 0.94, basically compatible with the level of noise, for the subset *a* of entries; the value of R^2 for the second set of entries *b*, on the contrary, is too low (0.82) and gives away the fact that one or more additional regressors would be required to interpret the dependent variable. The indications obtained with the help of the information theoretic indicators are in good agreement with the outputs of the ANNs.

It should be mentioned that this type of problem would be even more difficult to identify at the global level in case of multiple studies or in case the missing quantities belong to a contributor with a minority number of entries. However, the sufficiency indicators applied to the individual studies would show quite clearly, for which individual study the regressors do not constitute a complete list of independent quantities.

5.4 Global consistency

To exemplify the potential of the developed tools to address the issue of global consistency, a synthetic cumulative database has been generated, with the dependencies determined by the two following equations:

$$Y_a = X_{1a} \cdot X_{2a} \cdot X_{3a}^{2} \tag{16}$$

$$Y_b = X_{1b} \cdot X_{2b} \cdot X_{3a}^{-0.5} \tag{17}$$

In this case, the two systems, providing entries to the primary studies, present the same dependency on the quantities X_1 and X_2 but a completely different one on X_3. The tools developed in this work, when applied to the entire DB, identify X_1, X_2, X_3 as the relevant quantities to regress *Y*. However, both the residual entropy and the R^2 indicator basically assume values much lower than would be expected on the basis of the noise level affecting the data, as shown in Fig. 6. On the other hand, particularising the analysis for the individual studies *a* and *b* does not reveal any inconsistency in the individual studies. Indeed, the ANNs indicate even more clearly that X_1, X_2, X_3 are the

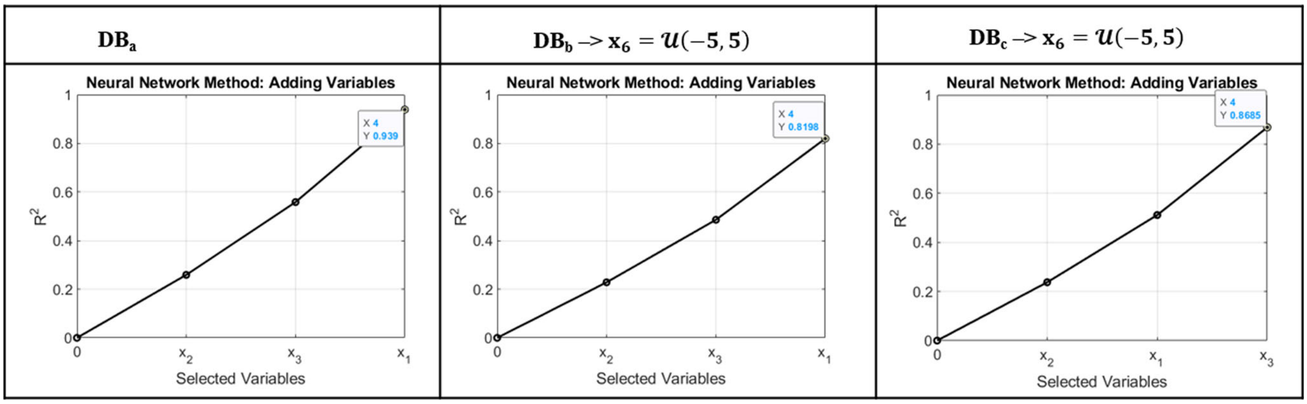

Fig. 5 Sufficiency identification for the data generated with Eqs. 14 and 15). The R^2 indicator assumes value compatible with the noise level only in the case of the database *a*, revealing that one or more regressors are missing in the study *b*. The database *c* is the union of *a* and *b*

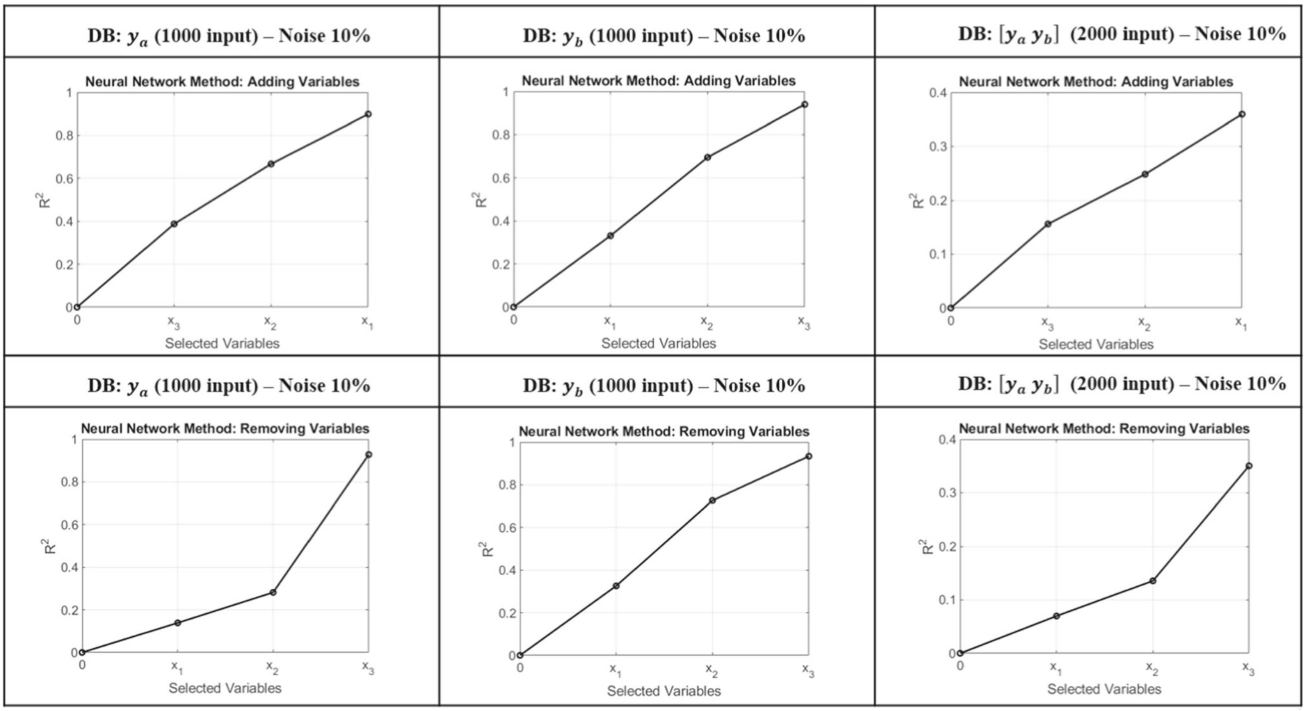

Fig. 6 Global consistency for the data generated with Eqs. 16 and 17. The values of the R^2 indicator are low only for the entire DB. This fact reveals that there is a global consistency issue, confirmed by the analysis particularised for the individual contributions, for which R^2 is sufficiently high to be compatible the noise level

right variables and R^2 values reach 0.9, compatible with 10% of random noise. It is worth mentioning once more that very coherent results are obtained with both approaches of adding to or removing from the network inputs one variable at the time. Therefore there is no indication that sufficiency is violated at the level of the individual studies. These conclusions are supported by the information theoretic indicators, both the conditional mutual information and the relevance.

For the systems described by Eqs. 16 and 17, only the final consistency checks reveal that something is amiss and that the two studies do not show the same dependencies and should therefore not be light heartedly included in the same regression. The values of the mutual information, between the regressors and the dependent quantity, indicate that the relations are different in the two studies contributing to the DB; these values are reported in Tables 1 and 2. For completeness sake, the Pearson correlation coefficients are also reported, to show how the linear correlations between X_3 and Y even change sign in the two subsets of entries. From inspection of these two tables, it appears very clearly that the information theoretic criteria

Table 1 Mutual information and Pearson correlation coefficient between the regressors and y for the subset a of data generated with Eqs. 16 and 17

Variables	Mutual information (y,x_i)	Pearson
X_1	0.1008	0.4008
X_2	0.0841	0.3887
X_3	0.2487	0.6203
X_4	0.0153	– 0.0002
X_5	0.0132	0.0302
X_6	0.0089	– 0.0110

Table 2 Mutual information and Pearson correlation coefficient between the regressors and y for the subset b of data generated with Eqs. 16 and 17

Variables	Mutual information (y,x_i)	Pearson
X_1	0.2190	0.5884
X_2	0.2202	0.5813
X_3	0.0796	– 0.3550
X_4	0.0105	0.0154
X_5	0.0050	– 0.0048
X_6	0.0107	0.0021

manage to identify the important regressors (the irrelevant variables have practically a negligible MI with the dependent quantity). The values of the mutual information for the relevant regressors X_1, X_2, X_3 are very different, indicating that the entries of the two individual studies cannot have been generated by the same system or process. These results are confirmed by the conditional mutual information, whose variable selection is reported in Table 3. The fact that the order of the quantities selected is different for the individual studies and for the entire DB reveals that something is amiss and that it is not legitimate to treat the individual contributions as part of a unique basis for the following analysis.

Table 3 Variable selection obtained with the CMI for the datasets generated with Eqs. 16 and 17

	DB–y_a	DB–y_b	DB–y_c
First selected variable	X_3	X_1	X_3
Second selected variable	X_2	X_2	X_2
Third selected variable	X_1	X_3	X_1

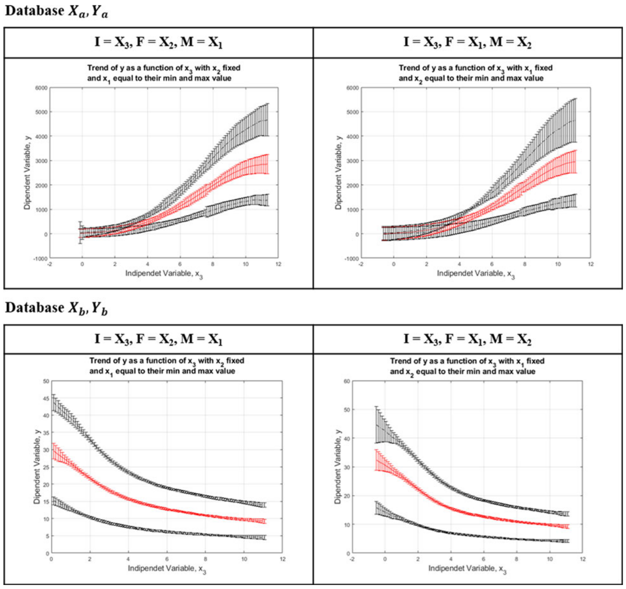

Fig. 7 Global consistency for the data generated with Eqs. 16 and 17. The y axis is the value of the dependent variable. The regressor scanned is X_3 and The other two independent variables X_1 and X_2 are alternatively kept fixed (F) or assume their median, maximum and minimum value (M), giving rise to the three curves

Also the networks show that the two individual studies are structurally different. The visualisation of the trends of variable X_3 for the two subsets of entries a and b, obtained with the trained network as explained in Sect. 4.3, shows quite clearly that the dependencies are different in the two individual studies, as reported in Fig. 7. The plots shown in this figure have been obtained by scanning the regressor X_3 versus alternatively one of the other two independent quantities, one at a time. For the quantity kept constant, three values have been chosen: mean, maximum and minimum in the confidence interval range. For each case therefore there are three different curves. Visual inspection of Fig. 7 reveals immediately that the trends of the effect of X_3 on the dependent variable are completely different in the two subsets of data a and b, indicating that the two studies cannot be naively aggregated in a single database. It is worth mentioning that whereas random-effects models prove not to be very useful in this case, psychometric meta-analysis of the correlation coefficients [10], particularised for the two subsets of data, would be able to detect the problem and would suggest not to combine the two studies.

6 Impact on model selection

To exemplify and illustrate the importance of the techniques proposed in this work, this section is devoted to the effects of overfitting on model selection. Indeed one of the main objectives of meta-analysis in the exact sciences consists of helping in the process of selecting the right model to interpret the available data. The techniques developed can help obtaining better results from the deployment of both frequentist and Bayesian or information theoretic model selection criteria. Next subsection is devoted to a brief overview of model selection criteria. The following addresses in detail how the tools described in this paper can contribute to avoid including spurious quantities in the models.

6.1 Brief overview of statistical criteria to select the best model among various alternatives

Choosing a model from a set of candidates, on the basis of the available evidence, is called model selection. It can be argued that the task of selecting the best mathematical model is the main and final objective of many scientific enquiries. To help performing it in sound way given the data, various statistical techniques have been developed.

Traditional frequentist techniques tend to emphasise the goodness of fit. In the context of selecting the important regressors for a specified family of models, the candidate quantities are given as inputs to the fitting routines and only the ones, providing a statistically significant improvement in the predictions, are retained. The statistical relevance of the results is assessed with aforementioned techniques such as Chi-squared, Anderson Darling and Kolmogorov–Smirnov [13].

Bayesian and information theoretic methods try to explicitly find a compromise between goodness of fit and complexity, by favouring simpler models. The Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC) are all consolidated and widely used indicators for this task [16]. The BIC is an unbiased estimator of the likelihood of a model. The form of the BIC indicator used in this paper is:

$$\mathrm{BIC} = n \cdot \ln\left(\sigma^2_{(\epsilon)}\right) + p \cdot \ln(n) \tag{18}$$

where $\epsilon = y_{\mathrm{data}} - y_{\mathrm{model}}$ are the residuals, $\sigma^2_{(\epsilon)}$ their variance, p is the number of parameters of the model, and n the number of y_{data} available, so the number of entries in the database (DB).

The AIC is a well-known model selection criterion, based on information theory. The AIC form most commonly used is:

$$\mathrm{AIC} = 2p + n \cdot \ln\left(\frac{\mathrm{RSS}}{n}\right) \tag{19}$$

where RSS is the Residual Sum of Squares between the experimental values and the estimates of the models, p is the number of parameters of the model and n the number of y_{data} provided, i.e. the number of entries in the database (DB).

All the aforementioned criteria are cost functions to be minimised, in the sense that better models have lower values of these metrics. This property can be appreciated by inspection of their mathematical structure. Indeed, BIC and AIC consist basically of two parts. The first one depends on the quality of the fit, represented by the residuals. Models closer to the data have lower values of this term. The second addend implements a penalisation for complexity, since it is proportional to the number of the parameters in the model equation. Therefore, parsimony is built in the cost function to avoid overfitting. The mathematical background to appreciate the relative merits of these model selection criteria, their strengths and weaknesses, can be found in [16].

6.2 The problem of overfitting due to spurious regressors

The issue of overfitting is a problem addressed carefully in machine learning and in statistics. To regress one quantity, it is important to pay attention not to include additional regressors, which have no impact on the dependent variable but have some level of correlation with other proper regressors. If additional spurious quantities are included in the set of regressors, the fit can be improved from a statistical point of view, reducing the χ^2 of the residuals. The model selecting criteria can therefore be induced to prefer models including spurious regressors, not really necessary to model Y. On the other hand, of course, these models lose physical meaning with the introduction of these spurious independent variables. This can therefore be considered a form of overfitting due to spurious regressors.

The nature of the problem can be easily appreciated by the following example. Let us assume that the cumulative database includes the contributions of three individual studies or experiments. The 8 available regressors considered are normalised between 1 and 10; they are sampled uniformly in this interval. The number of points in the synthetic database is 10^3. The three individual contributions have been generated with the following equations:

$$y_a = x_{1a} \cdot x_{2a} \cdot x_{3a} \cdot x_{4a} \tag{20}$$

$$y_b = x_{1b} \cdot x_{2b} \cdot x_{3b} \cdot x_{5b} \tag{21}$$

$$y_c = x_{1c} \cdot x_{2c} \cdot x_{3c} \cdot x_{6c} \tag{22}$$

The individual systems, providing entries to the database, present a power-law dependence that it is the same for the three quantities X_1, X_2 and $X_{3.}$ Then each depends also on a different quantity (again in power-law form).

Fitting the database with only three or with all the 8 regressors provides the results reported in Fig. 8. Inspection of the plots in this figure reveals immediately that all criteria would favour the choice of all the 8 regressors over the more limited set of three. Indeed, the R^2 is significantly higher for the model including all variables. Moreover, the

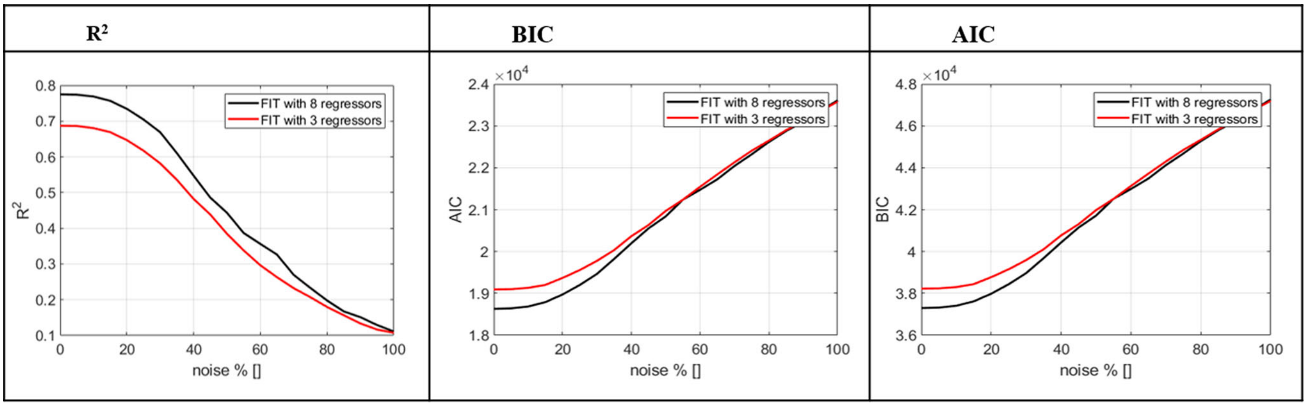

Fig. 8 Quality of the fits of the database generated by the three contributions, Eqs. 20, 21 and 22

value reached is not incompatible with the level of noise. AIC and BIC, being cost functions, are significantly lower for the model using all 8 regressors. Contrary to the case of global consistency problems investigated in Sect. 5.4, applying random-effects or psychometric meta-analysis does not help much in this situation [10]. Since the variables are not outliers but have concrete effects on parts of the database, also these techniques do not manage to clearly spot that something is wrong when considering the entire database. Only the analysis of the individual contributions, as suggested in this work, would therefore allow to detect that there is a problem with the cumulative database and that the individual studies should be considered separately. This is also the case of the real-life example reported in next section.

7 Application to a cumulative database of extreme relevance for thermonuclear fusion

In this section, the techniques previously described and benchmarked are applied to an experimental database, including measurements of the energy confinement time in the major Tokamak facilities in the world. Sect. 7.1 provides an overview of the physical background and a description of the problems that this cumulative database is meant to address. The technical details of the entries of the database, and the scaling laws originally derived from them, are discussed in Sect. 7.2. The results, obtained applying the tools developed in the context of the present work, are reported in Sect. 7.3.

7.1 Thermonuclear fusion and the scaling of the energy confinement time

The most exoenergetic reaction in the known universe is nuclear fusion, the coalescence of lighter nuclei to form heavier ones. Since the most relevant nuclei are hydrogen isotopes, they constitute an abundant fuel on earth. Consequently, nuclear fusion remains one of the most promising alternatives for the production of an almost unlimited quantity of energy without emission of greenhouse gasses, to satisfy the requirements of an increasingly power hungry world population. Magnetic Confinement Nuclear Fusion (MCNF) constitutes the most credible approach to a reactor, potentially capable of producing commercially viable power.

In MCNF, plasmas are confined by magnetic fields in a vacuum container and heated to temperatures higher than in the core of the sun. So far the best performance, in the perspective of the reactor, has been obtained with the configuration of the magnetic fields of the tokamak type. In any case, for every configuration, one of the most crucial quantities, to assess its reactor relevance, is the so called energy confinement time τ_E. This indicator quantifies how fast the internal energy of the plasma is lost [17]. Unfortunately, the transport mechanisms affecting the energy confinement in high temperature plasmas are very complex and nonlinear. Moreover, they include effects at many scales, ranging from microturbulence to macroscopic dimensions comparable to the devices' size. So, even if the understanding of the energy transport has progressed a lot in the last years, it remains very difficult, if not impossible, to properly estimate τ_E from theoretical or numerical calculations. As a consequence, since various decades, an empirical approach has been pursued, which consists of extracting robust scaling laws for τ_E from experimental data. Indeed, the energy confinement time is estimated in all the major devices on a routine basis. Multi-machine databases, containing information about this quantity, have been built. The most advanced remains the International Tokamak Programme Agreement (ITPA) database, which was expressively built to support advanced confinement time studies and includes validated measurements from all of the most relevant tokamak devices ever operated in the world [18, 19]. The DB3v13f version of the DB is also the one used to define the scaling chosen as a basis for the design of ITER, the IPB98 (y,2). This version, with the same selection rules reported in [19], is the DB analysed in the following as an application of the techniques proposed in the previous sections. The ITPA DB is a typical example of a cumulative database, because the entries are individual results but collected in different experiments performed in different devices.

7.2 The ITPA Database of the energy confinement time for the H mode

As mentioned, to maximise the generality of the results obtained with the tools described in the previous sections, an international database has been considered (DB3). To select a set of suitable predictors, the same variables, considered in the conventional scaling law for the plasma confinement time, have been taken into account in this study as well. These are:$I_p, B_T, P_{\mathrm{LTH}}, n_{\mathrm{el}}, M_{\mathrm{eff}}, R_{\mathrm{GEO}}, \epsilon, k_a$, where I_p is the plasma current, B_T is the toroidal magnetic field, P_{LTH} is the power loss across the last closed surface, n_{el} is the line average electron density, M_{eff} is the plasma isotopic composition, R_{GEO} is the plasma major radius, $\epsilon = \frac{a}{R_{\mathrm{GEO}}}$, where a is the plasma minor radius, and k_a is the volume measure of elongation [20].

The contributions of the various devices to the DB are reported in Table 4 for the case of the H mode of confinement. Inspection of the table suggests to divide the entries in three groups. The contributions of the largest devices JET, JT-60U and TFTR; the set of values provided by the medium size tokamaks ASDEX, AUG, DIII-D and PBXM; and finally the set made of the contributions of all the remaining smaller experiments.

Table 4 Entries per machine for the *H* mode case

TOK	#
ASDEX	431
AUG	526
CMOD	45
COMPASS	16
DIII-D	300
JET	1413
TDEV	3
TCV	11
TFTR	13
JFT-2 M	70
JT-60U	87
NSTX	5
PBXM	59
PDX	97
MAST	9
START	8

The most widely accepted scaling law, extracted from this database in terms of dimensional quantities, is the IPB98(y,2):

$$\tau_E = 5.62 \cdot 10^{-2} \cdot I_p^{0.93} B_t^{0.15} n_e^{0.41} P^{-0.69} R^{1.97} k^{0.78} \varepsilon^{0.58} M_{\mathrm{eff}}^{0.19} \tag{23}$$

The scaling reported as Eq. 23 has been obtained with log regression, assuming implicitly that the most appropriate mathematical form of the equation is a power-law monomial.

7.3 Results

Employing the feature selection algorithm proposed in this work, the main objective of the analysis consists of verifying whether all the predictor variables included in Eq. 23 are really necessary for a good description of τ_E, or whether a smaller subset of variables is sufficient. Unfortunately, the number of entries is quite low for the complexity and the dimensionality of the problem. The ANNs therefore perform much better to analyse the database and the information theoretic indicators can play only a confirmatory role, being affected by higher uncertainties.

Application of the ANN, proposed in this work, to the ITPA database provides the results reported in Table 5 for the entire dataset and the three subsets previously mentioned. To decide when to exclude additional variables, a backward (removing variable) method with ensembles of NN has been utilised. 36 NNs have been trained for 20 times, and the mean RMSE over the 36 NNs has been calculated. If, after removing a variable, the increase in the RMSE mean is statistically significant, that quantity is retained in the set of regressors. The significance level of 5% with a Z test is the threshold implemented to assess whether the improvement in the RMSE mean is statistically relevant.

To confirm the results obtained as just described, the following procedure has been deployed for each of the four groups. Log regression has been used to fit the database, firstly including the entire set of variables available and

Table 5 Variable selection for the ITPA database and three subsets

Database	Selected variables
JET/JT-60U/TFTR	$I_p, P_{\mathrm{LTH}}, n_{\mathrm{el}}, \epsilon$
ASDEX, AUG/DIII-D/PBXM	$I_p, P_{\mathrm{LTH}}, , n_{\mathrm{el}}, R_{\mathrm{GEO}}, B_T, M_{\mathrm{eff}}$
OTHERS	$I_p, P_{\mathrm{LTH}}, , n_{\mathrm{el}}, R_{\mathrm{GEO}}, M_{\mathrm{eff}}$
ENTIRE DATABASE	$I_p, P_{\mathrm{LTH}}, n_{\mathrm{el}}, R_{\mathrm{GEO}}, B_T, \epsilon$

secondly considering only the ones selected with the help of the ANNs. Then Gaussian noise has been added to the database entries, both to the regressors and to the predicted variables. The standard deviation of the noise has been chosen equal to the error bars in the entries of the database summarised in Table 6. Consequently, the two power laws previously obtained have been evaluated using the new data with added noise, and their goodness of fit has been calculated again with the R^2 indicator. Since the process of adding noise is random, the procedure above has been repeated again 500 times in order to provide sufficiently sound statistical basis for the conclusions. The results are two distribution of R^2, one for the power law with all the variables, and one for the power law with only the regressors selected by the ANNs proposed in this work. The two distributions are then compared to assess whether all the variables in the power law are really useful to describe τ_E or whether the set selected to obtain the IPB98 includes spurious quantities.

The R^2 achieved by fitting the datasets with all the eight variables of the IPB98(y,2) or with the ones selected by the ANNs are practically identical. When the noise is added, the power laws with all the variables in Eq. 23 present a goodness of fit lower than the ones using only the regressors selected by the ANNs and the difference is clearly statistically significant. A visualisation of the two R^2 distributions for each group is reported in Fig. 9, whose plots substantiate the previous statements. The proposed procedures therefore provide quite strong evidence that the variables removed by the algorithm do not bring any new information about τ_E and the IPB98 is affected by a problem of spurious quantities. The meta-analytic tools, introduced in this work, also indicate that the entries of the three groups of devices should not be included in the same dataset to apply a global fit, confirming previous preliminary investigations on the same subject [20].

To confirm the competitive advantage of the developed indicators, compared to traditional meta-analytic techniques, the data have been aggregated with psychometric techniques [10]. Unfortunately, all the correlation coefficients between the regressors and the dependent variable are statistically significant and therefore there is no indication that some of them should be discarded. Using these coefficients as a first guess for the log regression provides a power-law scaling, whose exponents are reported in Table 7. The scaling has a poor physical meaning and the extrapolation to ITER gives an estimate of significantly more than 4 s for the energy confinement time, clearly an unrealistically optimistic low value. This is because the traditional meta-analytic techniques are not designed to detect that the database contains sets of entries from experiments of different nature, which should not be grouped together. Consequently also in this real-life case study, the procedure proposed in the present work seems an indispensable preliminary step to any form of reasonable statistical inference.

Table 6 The uncertainties for the entries of the ITPA database

	I_p	B_T	P_{LTH}	n_{el}	M_{eff}	R_{GEO}	ϵ	k_a	τ
Rel. err	1%	1%	14%	5%	8%	1%	1%	10%	10%

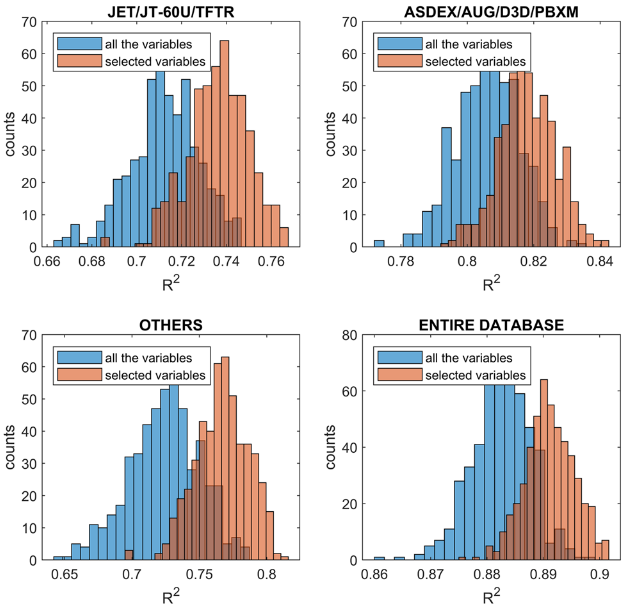

Fig. 9 Results of the tests to assess the level of overfitting in the ITPA database

8 Conclusions

Large databases, composed of the entries produced by different experiments and studies, are becoming increasingly important and popular in many fields of science, given the widespread use of sensors and storage technologies. In practice, these DBs pose significant interpretation issues, which cannot be addressed, neither with the traditional tools for primary studies nor with the meta-analytic techniques used to handle summary statistics. A series of data analysis techniques have therefore been developed to address the research synthesis issues inherent to the analysis of these databases. The combination of completely independent mathematical tools, belonging to information theory and neural computation, allows deriving quite robust results, as confirmed by a series of systematic tests with synthetic data. The deployment of the methodology, for the analysis of an international database

Table 7 Parameters of the power law scaling obtained with the help of meta-analytic techniques

	I_p	B_T	P_{LTH}	n_{el}	M_{eff}	R_{GEO}	ϵ	k_a
Exponent	$0.25_{-0.30}^{0.80}$	$0.35_{0.01}^{0.71}$	$-0.50_{-0.70}^{-0.30}$	$0.70_{0.30}^{1.10}$	$0.04_{-0.44}^{0.52}$	$2.58_{1.67}^{3.48}$	$0.05_{0.02}^{0.09}$	$1.13_{0.40}^{1.85}$

built by the thermonuclear fusion community, has revealed the potential of the approach to address practical and quite difficult examples. It is probably worth noting that the tools, developed to address regressors selection and consistency, can be deployed also to analyse large primary databases, if it is not clear which are the independent variables and whether they are enough to interpret the data.

With regard to future developments, from a methodological perspective, it would be important to extend the described tools to the case of classification. In this respect, the crucial point is the definition of an appropriate metric to minimise. In both supervised and unsupervised classification, it could be based on the distance to the boundary between the classes, but more work is required to properly develop this aspect. More generally, it should be remembered that the proposed techniques have been conceived to provide only a detailed analysis of the individual contributions to the global database. The potential of these tools, and additional ones, to contribute to other applications of research synthesis, remains to be explored. More sophisticated versions of the developed indicators could contribute significantly to multilevel synthesis, which is typically performed mainly with linear tools [21]. In particular, how to best combine the individual studies, to obtain more general conclusions, is a subject worth investigating in detail. The potential of other machine learning families, in particular evolutionary computation, should also be considered seriously, given the great results of these techniques in many fields [22–26]. Regarding thermonuclear fusion, the analysis performed in Sect. 7 should be applied also to the other major cumulative databases used in the community, such as the ones for the L mode of confinement and the Stellarator configuration at different levels of optimisation. Particularising the results for devices with a metallic wall, such as JET ITER Like Wall [27, 28], is also a priority, together with the extension of the approach to profiles and distributed quantities [29, 30].

Funding Open access funding provided by Università degli Studi di Roma Tor Vergata within the CRUI-CARE Agreement. The authors did not receive support from any organisation for the submitted work.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Data Availability The data used in this paper are publicly available.

Code Availability Not applicable.

References

1. Hall Judith A, Tickle-Degnen L, Rosenthal R, Mosteller F (1993) Hypotheses and problems in research synthesis. In: Cooper H, Hedges LV (eds) The handbook of research synthesis. Russell Sage Foundation
2. Nordmann AJ, Kasenda B, Briel M (2012) Meta-analyses: what they can and cannot do. Swiss Med Wkly. https://doi.org/10.4414/smw.2012.13518.PMID22407741
3. O'Rourke K (2007) An historical perspective on meta-analysis: dealing quantitatively with varying study results. J R Soc Med 100(12):579–582. https://doi.org/10.1258/jrsm.100.12.579.PMC2121629.PMID18065712
4. Pearson K (1904) Report on certain enteric fever inoculation statistics. BMJ 2(2288):1243–1246. https://doi.org/10.1136/bmj.2.2288.1243.PMC2355479.PMID20761760
5. Pratt JG, Rhine JB, Smith BM, Stuart CE, Greenwood JA (1940) Extra-sensory perception after sixty years: a critical appraisal of the research in extra-sensory perception. Henry Holt
6. Glass GV (1976) Primary, secondary, and meta-analysis of research. Educ Res 5(10):3–8. https://doi.org/10.3102/0013189X005010003
7. Wilson DB, Lipsey MW (2011) Practical meta analysis thousand oaks. Sage
8. M Borenstein, LV Hedges, JPT Higgins, HR Rothstein "*Introduction to meta-analysis*" 2011
9. https://training.cochrane.org/resource/grade-handbook
10. Deniz SO, Chockalingam V, Frank LS (2017) Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management. Human Resourc Manage Rev 27:1. https://doi.org/10.1016/j.hrmr.2016.09.011

11. Stone JV (2015) Information theory: a tutorial introduction. Septbel Press
12. Masters T (2017) Assessing and improving prediction and classification: theory and algorithms. Springer
13. Gregory WC, Dale IF (2014) Nonparametric statistics: a step-by-step approach. John Wiley and Sons Inc
14. Sollich P, Krogh A (1996) Learning with ensembles: how overfitting can be useful. Adv Neural Inf Process Syst 8:190–196
15. Cha Z, Yunqian M (2012) Ensemble machine learning: methods and applications. Springer Science & Business Media
16. Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. Springer
17. Wesson J (2011) Tokamak. Oxford University Press
18. McDonald D et al (2004) ELMy H-modes in JET helium-4 plasmas. Plasma Phys Control Fusion 46:519–534
19. http://efdasql.ipp.mpg.de/hmodepublic/DataDocumentation/ Datainfo /DB3v13/db3v13.html
20. Murari A et al (2020) Testing the consistency of multimachine databases for physical studies of regression. Nucl Fusion 60:015001. https://doi.org/10.1088/1741-4326/ab4285
21. Daniel C (2007) M*ultilevel Synthesis: From the Group to the Individual*" Springer, Cham
22. Schmid M, Lipson H (2009) Distilling free-form natural laws from experimental data. Science 324:81
23. Murari A et al (2013) Non-power law scaling for access to the H-mode in tokamaks via symbolic regression. Nucl Fus 53(4):43001. https://doi.org/10.1088/0029-5515/53/4/043001
24. Murari A et al (2015) Symbolic regression via genetic programming for data driven derivation of confinement scaling laws without any assumption on their mathematical form. Plasma Phys Controll Fus 57(1):14008. https://doi.org/10.1088/0741-3335/57/1/014008
25. Murari A et al (2015) A new approach to the formulation and validation of scaling expressions for plasma confinement in tokamaks. Nucl Fus 55:7. https://doi.org/10.1088/0029-5515/55/7/073009
26. Murari A et al (2016) Application of symbolic regression to the derivation of scaling laws for tokamak energy confinement time in terms of dimensionless quantities. Nucl Fus 56(2):26005. https://doi.org/10.1088/0029-5515/56/2/026005
27. Pamela J et al (2007) The JET programme in support of ITER. Fus Eng Des 82(5):590–602. https://doi.org/10.1016/j.fusengdes.2007.03.003
28. Romanelli F et al (2011) Overview of JET results. Nucl Fus 51:94008. https://doi.org/10.1088/0029-5515/51/9/094008
29. Mazon D et al (2003) Active control of the current density profile in JET. Plasma Phys Controll Fus 45:7. https://doi.org/10.1088/0741-3335/45/7/102
30. Craciunescu T et al (2009) A comparison of four reconstruction methods for JET neutron and gamma tomography. Nucl Instruments Methods Phys Res, Sect A: Accelerators, Spectrom, Detectors Assoc Equipment 605(3):374–3831. https://doi.org/10.1016/j.nima.2009.03.224

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