Comparative Study of Imputation Algorithms Applied to the Prediction of Student Performance

CONCEPCIÓN CRESPO-TURRADO[*], *University of Oviedo, Maintenance Department, San Francisco 3, Oviedo 33007, Spain.*

JOSÉ LUIS CASTELEIRO-ROCA, *University of A Coruña, Departamento de Ingeniería Industrial, A Coruña 15405, Spain .*

FERNANDO SÁNCHEZ-LASHERAS, *University of Oviedo, Department of Mathematics, Facultad de Ciencias, Oviedo 33007, Spain.*

JOSÉ ANTONIO LÓPEZ-VÁZQUEZ, *University of A Coruña, Departamento de Ingeniería Industrial, A Coruña 15405, Spain.*

FRANCISCO JAVIER DE COS JUEZ, *University of Oviedo, Department of Mathematics, Facultad de Ciencias, Oviedo 33007, Spain.*

FRANCISCO JAVIER PÉREZ CASTELO, *University of A Coruña, Departamento de Ingeniería Industrial, A Coruña 15405, Spain.*

JOSÉ LUIS CALVO-ROLLE, *University of A Coruña, Departamento de Ingeniería Industrial, A Coruña 15405, Spain.*

EMILIO CORCHADO, *University of Salamanca, Departamento de Informática y Automática, Plaza de la Merced s/n, 37.008, Salamanca, Salamanca, Spain.*

Abstract

Student performance and its evaluation remain a serious challenge for education systems. Frequently, the recording and processing of students' scores in a specific curriculum have several flaws for various reasons. In this context, the absence of data from some of the student scores undermines the efficiency of any future analysis carried out in order to reach conclusions. When this is the case, missing data imputation algorithms are needed. These algorithms are capable of substituting, with a high level of accuracy, the missing data for predicted values. This research presents the hybridization of an algorithm previously

[*]E-mail: jose.luis.casteleiro@udc.es

Vol. 28, No. 1, © The Author(s) 2020. Published by Oxford University Press. All rights reserved.
For permissions, please e-mail: journals.permissions@oup.com.
Advance Access published 2 January 2020 doi:10.1093/jigpal/jzz071

proposed by the authors called adaptive assignation algorithm (AAA), with a well-known technique called multivariate imputation by chained equations (MICE). The results show how the suggested methodology outperforms both algorithms.

Keywords: Student performance, Data imputation, MARS, MICE, AAA.

1 Introduction

The academic performance of students in higher education is one of the main concerns at universities in Spain [1]. Education parameters such as exam absenteeism and failure rates, the grade point average (GPA) or the number of attempts needed to pass an exam show a low academic performance that has consistently been the case for years [2]. Please note that in this context, exam absenteeism is understood as all those students that are enrolled in certain subject but do not attend the exam.

Some research works on technical engineering degrees [3] show that only 11% of students graduate after the 3 years established by the educational curriculum. These studies also evidence that the average time in finishing the degree is 5.41 years, the dropout rate is around 70%, and the performance rate (credits passed by credits enrolled) is 56%.

In addition to the courses cited above, additional new studies on the Spanish University System (SUE; Sistema Universitario Español) show that engineering and architecture performance rates are still around 60% and are as such the lowest in the SUE. Compared to studies with better performance, such as degrees in health sciences, the rates for engineering and architecture are 20% lower.

In accordance with the guidelines of quality assurance systems under the European Higher Education Area (EHEA), tracking of studies is regulated from a legal point of view and is of course obligatory for official university degrees [1]. From this point of view, the internal quality systems of the educational institutions try to enhance their quality ratios or indicators in terms of academic results and performance, with the aim of ongoing improvement [2]. This fact means that the faculties or higher education centres need tools to support or assist with this task [4, 5].

In order to create a tool for making decisions, it is usually necessary to find a way to obtain the required knowledge. Traditionally, in past research works, the common method was to obtain a model based on a dataset of the historical, whether through traditional techniques or through other more advanced ones [6–13].

The above method could be problematic in general terms, given the need to have previous cases showing similar performance [14–22]. Also, it must be remarked that the case under study could change. If that were the case, the model must be adaptive for new cases with different casuistic and performance [23–28]. In this sense, the imputation methods based on evolutionary methods could be a good solution to the problem described here.

This paper evaluates two imputation methods which allow the system to fill in the missing data of any of the students' scores used in this research. One of the algorithms, the adaptive assignation algorithm (AAA) [29], is based on multivariate adaptive regression splines (MARS), and the other one is the multivariate imputation by chained equations (MICE) [30]. The first one performs well in general terms, when the percentage of missing data for a case is small; otherwise, the second method is more appropriate. A new algorithm that hybridizes the two aforementioned algorithms improves the results obtained. The right combination of both algorithms is a good solution that involves establishing the border application of both.

The document is structured in the following way. After the present section, the case of study is described. This consists of the dataset of students' scores in the Electrical Engineering Studies Degree of the University of A Coruña. Then, the techniques for missing data imputation are shown.

The results section shows the outcomes achieved with the imputation over the dataset for three different cases over the case of study. After that, the conclusions and future works are presented.

2 Case Study

The dataset used in this research is made up of students' scores in the Electrical Engineering Studies Degree of the University of A Coruña from the academic year 2001/2002 until 2008/2009. The dataset includes the scores for each subject in the degree: nine subjects in the first year, another nine in the second year, seven in the last year and the final project.

The data also includes the scores and the way to access the University studies. In Spain, there are two different ways: from secondary school and from vocational education and training. Moreover, the scores for the subjects in the degree include not only the mark; the time taken to pass each subject is also included.

The dataset under study has all the data. It is an important fact to test the performance of the algorithms used in this study. It will be possible to emulate several different percentages of missing values and compare both methods with the aim of establishing the right frontier of both methods of application. Then, with the combination, a hybrid model will be obtained to increase the applicability of the method in a wide range of possibilities.

2.1 The data imputation techniques used

The imputation process involves replacing missing values with plausible values. The quality of the imputation depends on the quality of the methodology employed and will greatly affect the final result of the problem under study. In this section the data imputation techniques employed in the present research are described. In a problem that presents complex incomplete data, multiple imputation methods are required [31], and many examples of these methodologies can be found in the bibliography.

2.2 The MARS algorithm

MARS is a non-parametric multivariate regression analysis technique in which the interaction of nonlinearities and variables can be modelled [32]. In a MARS model, any dependent variable can be represented by means of M basis functions by means of the following formula:

$$\hat{\vec{y}} = c_0 + \sum_{m=1}^{M} c_m B_m \left(\vec{x}\right) \tag{1}$$

The dependent variable is $\hat{\vec{y}}$, c_0 is a constant, $B_m \left(\vec{x}\right)$ is the *mth* basis function and c_m is its coefficient. Linear basis functions can be either constant or hinge functions, including their products. The MARS models are able to model nonlinearities and interactions as a weighted sum of basis functions [32, 33].

The generalized cross-validation method [34] is employed by the MARS methodology in order to determine which basis functions are to be included in the model. The model is built in two phases: first, a forward variable selection and then a backward deletion. In the forward stage, basis functions are added in order to reduce the training error, while in the backward stage, the model obtained is pruned in order to avoid overfitting [35]. At the end of the backward phase, from those best models of each size, a model with the lowest GCV value is selected and considered as the final one. The backward pass uses GCV to compare the performance of model subsets in order to choose the best subset, taking into account that lower values of GCV are better.

The GCV can be expressed as follows:

$$GCV(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}m\left(\overrightarrow{x}_i\right)\right)^2}{(1 - C(M)/n)^2} \tag{2}$$

In the formula represented above, the numerator is the mean squared error of the evaluated model in the training data, n is the number of data cases in the training database and $C(M)$ is a complexity penalty that increases with the number of basis functions in the model [36].

2.3 The MICE algorithm

The MICE algorithm developed by van Buuren and Groothuis-Oudshoorn [31] is a Markov chain Monte Carlo method where the state space is the collection of all imputed values. Like any other Markov chain, in order to converge, the MICE algorithm needs to satisfy the three following properties [37–39]:

- Irreducible: The chain must be able to reach all parts of the state space.
- Aperiodic: The chain should not oscillate between different states.
- Recurrence: Any Markov chain can be considered as recurrent if the probability that the Markov chain starting from i will return to i is equal to one.

In practice, the convergence of the MICE algorithm is achieved after a relatively low number of iterations, usually somewhere between 5 and 20 [39]. According to the experience of the algorithm creator, five iterations are generally enough, but some special circumstances would require a greater number of iterations. In the case of the present research, and due to the performance of the results obtained when compared with the other methods applied, five iterations were considered to be enough. This number of iterations is much lower than in other applications of the Markov chain Monte Carlo methods, which often require thousands of iterations. In spite of this, and from a researcher's point of view and experience, it must be also remarked that in the most common of the applications, each iteration of the MICE algorithm would take several minutes or even a few hours. Furthermore, the duration of each iteration is mainly linked with the number of variables involved in the calculus and not with the number of cases. It should be taken into consideration that imputed data can have a considerable amount of random noise, depending on the strength of the relations between the variables. So in those cases in which there are low correlations among variables or they are completely independent, the algorithm convergence will be faster. Finally, high rates of missing data (20% or more) would slow down the convergence process work. The MICE algorithm [39] for the imputation of multivariate missing data consists of the following steps:

1. Specify an imputation model $P\left(Y_j^{mis}|Y_j^{obs}, Y_{-j}, R\right)$ for variable Y_j with $j = 1, \ldots, p$
   The MICE algorithm obtains the posterior distribution of R by sampling interactive from the above-represented conditional formula. The parameters R are specific to the respective conditional densities and are not necessarily the product of a factorization of the true joint distribution.
2. For each j, fill in starting imputations Y_j^0 by random draws from Y_j^{obs}.
3. Repeat for $t = 1, \ldots, T$ (iterations).

```
4. Repeat for $j = 1,\ldots,p$ (variables).
5. Define $Y_{-j}^{t} = \left(Y_1^t,\ldots,Y_{j-1}^t, Y_{j+1}^{t-1},\ldots,Y_p^{t-1}\right)$ as the currently complete data
   except $Y_j$.
6. Draw $\varnothing_{\mathrm{j}}^{\mathrm{t}} \sim \mathrm{P}\left(\varnothing_{\mathrm{j}}^{\mathrm{t}} | \mathrm{Y}_{\mathrm{j}}^{\mathrm{obs}}, \mathrm{Y}_{-\mathrm{j}}^{\mathrm{t}}, \mathrm{R}\right)$.
7. Draw imputations $Y_j^t \sim P\left(Y_j^{mis} | Y_j^{obs}, Y_{-j}^t, R, \varnothing_j^t\right)$.
8. End repeat $j$.
9. End repeat $t$.
```

In the algorithm referred to, Y represents an $n \times p$ matrix of partially observed sample data, R is an $n \times p$ matrix, *0–1* represents response indicators of Y, and Ø represents the parameters space. Please note that in MICE imputation [40], initial guesses for all missing elements are provided for the $n \times p$ matrix of a partially observed sample. For each variable with missing elements, the data is divided into two subsets, one of which contains all the missing data. The subset with all the available data is regressed on all other variables. Then, the missing subset is predicted from the regression, and the missing values are replaced with those obtained from the regression. This procedure is repeated for all variables with missing elements. After this, all the missing elements are imputed according to the algorithm described above, while the regression and prediction are repeated until the stop criterion is reached: in this case, until a certain number of consecutive iterates fall within the specified tolerance for each of the imputed values. The MICE algorithm has performed well in previous studies where it was employed by the authors [30]. One of the main drawbacks to this algorithm is the high computational times required for the resolution of complex problems, and one of its main advantages is that no prior knowledge of the distribution is required. As far as it is known by the authors, this is the first time that the MICE algorithm has been used for the imputation of a kind of data such as that in the present article.

2.4 The AAA algorithm

With the purpose of explaining the AAA, let us assume that we have a dataset formed by n different variables $v_1, v_2, \ldots, v_n$. In order to calculate the missing values of the *ith* column, all the rows with no missing value in the said column are employed. Then, a certain number of MARS models are calculated. It is possible to find rows with very different amounts of missing data from *0* (no missing data) to n (all values are missing). Those columns with all values missing will be removed and will be neither used for the model calculation nor imputed. Therefore, any number of missing data from *0* to $n - 2$ is feasible (all variables but one with missing values).

In other words, if the dataset is formed by variables $v_1, v_2, \ldots, v_n$. and we want to estimate the missing values in column v_i, then the maximum number of different MARS models that would be computed for this variable (and in general for each column) is as follows: $\sum_{k=1}^{n-1}\binom{n-1}{k}$. In the case of the data under study in this research, with 10 different variables, a maximum of 5,110 distinct MARS models are trained (511 for each variable).

After the calculation of all the available models, the missing data of each row will be calculated using those models that employ all the available non-missing variables in the row. In those cases where no model was calculated, the missing data will be replaced by the median of the column. Please note that large datasets with a not-too-high percentage of missing data will be infrequent. As a general rule for the algorithm, it has been decided that when a certain value can be estimated using more than one MARS model, it must be estimated using the MARS model with the largest number of input variables; the value would be estimated by any of those models chosen at random.

Finally, in those exceptional cases in which no model is available for estimation, the median value of the variable will be used for the imputation. Please note that this lack of model only occurs in exceptional cases as they have a particularly low probability. For a more in-depth explanation, please see [29].

2.5 *The hybrid algorithm*

The hybrid algorithm combines MICE and AAA algorithms in search of a better performance, using one or the other depending on the most favourable conditions for their application. First, the MARS models of the AAA algorithm are trained using those rows without missing data as input values. In other words, the information from those rows without missing data is employed to train MARS models. Afterwards, the MARS models trained are employed for calculating missing values only in those rows with one or two missing values but not in those with three or more missing data. Finally, the resulting matrix, in which there are missing values only in those rows with three or more missing data, is imputed using the MICE algorithm.

This way of working takes advantage of the AAA algorithm for imputation in those cases of few (one or two) missing data per row where it has performed better and reduces the computational time required by the MICE algorithm, as fewer missing data are required to be imputed.

2.6 *Model validation*

Leave-one-out cross-validation has been used to analyse the spatial error of interpolated data [41, 42]. This procedure involves using eight of the nine stations in the model to obtain the estimated value in the ninth station (this one is left out) in order to calculate the performance of the different methods employed; mean absolute error (MAE), root-mean-square error (RMSE) and mean absolute deviation (MAD) have been employed.

The equations that represent these three metrics, frequently employed in literature [29, 30], are as follows:

$$\mathrm{MAE}\,(\%) = \frac{\mathrm{MAE}}{\frac{1}{n}\sum_{i=1}^{n} G_i} x100 \tag{3}$$

$$\mathrm{MAD} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{G}_i - G_i\right| \tag{4}$$

$$\mathrm{RMSE} = \sum_{i=1}^{n}\sqrt{\frac{1}{n}\left(\hat{G}_i - G_i\right)^2} \tag{5}$$

where G_i and $\hat{G}_i$ are the measurements and the model estimated and n is the number of data points of the validation set. The RMSE weighs large estimation errors more strongly than small errors, and it is considered a very important model validation metric. Also, MAE is a useful complement of the measured-modelled scatter plot near the 1-to-1 line [39].

3 Results

To calculate the performance of each algorithm, several tests were carried out, with differing quantities of missing data. First of all, it should be pointed out that, for the results shown in the tables, only 10 columns from the total dataset have been taken into account. Each column represents

TABLE 1. Results for algorithm with one missing value.

	MAE		MAD		RMSE	
	MICE	AAA	MICE	AAA	MICE	AAA
Column 1	0.15013	3.33E-15	0.07413	1.32E-15	0.25495	3.41E-15
Column 2	1.00E-03	1.11E-15	1.20E-03	0	0.00325	1.17E-15
Column 3	2.10E-03	8.44E-15	1.01E-03	1.25E-14	0.00247	8.45E-15
Column 4	0.07502	7.77E-15	0.07412	1.98E-15	0.11180	7.93E-15
Column 5	0.07503	5.11E-15	1.53E-03	0	0.15019	5.12E-15
Column 6	1.01734	5.11E-15	1.40805	1.32E-15	1.20623	5.24E-15
Column 7	0.19235	6.66E-16	2.36E-03	0	0.21015	1.33E-15
Column 8	0.02492	1.78E-15	0.02313	1.32E-15	0.04502	1.99E-15
Column 9	1.25013	1.55E-15	1.48260	0	1.36931	1.60E-15
Column 10	0.75024	2.89E-15	0.88956	0	0.90135	2.91E-15

TABLE 2. Results for algorithm with two missing values.

	MAE		MAD		RMSE	
	MICE	AAA	MICE	AAA	MICE	AAA
Column 1	0.22265	5.08E-05	0.11597	5.82E-05	0.35759	4.90E-06
Column 2	0.09623	8.26E-05	0.10724	4.36E-05	0.03246	4.08E-05
Column 3	0.01146	9.28E-05	0.03063	3.00E-06	0.02834	4.16E-05
Column 4	0.11066	6.47E-05	0.08836	5.62E-05	0.16341	6.84E-05
Column 5	0.12991	3.09E-05	0.04132	2.92E-05	0.22379	9.68E-05
Column 6	1.07473	7.28E-06	1.43598	3.69E-05	1.23973	5.79E-05
Column 7	0.26962	4.01E-05	0.09282	5.15E-05	0.22093	8.40E-05
Column 8	0.10302	6.47E-05	0.07025	3.14E-05	0.05451	7.37E-05
Column 9	1.34970	1.13E-05	1.49259	2.32E-05	1.46712	4.40E-05
Column 10	0.80736	6.71E-05	0.96982	2.71E-06	0.94504	9.29E-05

a different subject, and the selection was made at random. In all the tests, the percentage of missing data is the same, 10%, but the real missing data varied from one to more than three, depending on the test. From the authors' point of view, this way of proceeding allows us to say that the algorithm will perform robustly under many different real conditions.

Table 1 shows the performance of each algorithm with only one value missing in each case. According to the results obtained for MAE, MAD and RMSE, it may be appreciated that the AAA clearly performs better than MICE.

In Table 2, the performance was calculated for two missing values. In this case, as in the previous one, the AAA is clearly better than MICE, but the difference in performance between both algorithms is smaller.

The results present in Table 3 show that when the number of missing values increases to three, the MICE algorithm performs better than the AAA. Please note how for this case the results are clearly different to those presented in Tables 1 and 2.

With the aim of obtaining the best results, a hybrid of the two algorithms was created. The results of this hybrid system are shown in Tables 4 and 5. In these tables, the percentage of missing values is

TABLE 3. Results for algorithm with three of missing values.

	MAE		MAD		RMSE	
	MICE	AAA	MICE	AAA	MICE	AAA
Column 1	0.29825	1.29788	0.21251	0.97760	0.37785	0.33879
Column 2	0.10030	0.32404	0.23003	0.68342	0.08148	0.88620
Column 3	0.23256	0.76501	0.14919	1.28895	0.06039	1.19963
Column 4	0.22723	1.01881	0.21624	0.30309	0.32788	1.18293
Column 5	0.18359	0.66619	0.13611	0.46380	0.29628	0.84689
Column 6	1.14205	0.93627	1.61320	1.10473	1.36929	0.91176
Column 7	0.37556	0.78123	0.10937	0.68277	0.32463	0.72429
Column 8	0.14983	1.18887	0.14077	1.05084	0.06936	0.42409
Column 9	1.44081	1.01170	1.56163	1.21070	1.49719	0.94814
Column 10	0.89382	0.94070	1.01722	0.32494	1.09256	1.19549

TABLE 4. Results for algorithm with random missing values and hybrid combination (10% of missing values).

	MAE	MAD	RMSE
Column 1	0.12029	0.06570	0.16468
Column 2	0.02435	0.05345	0.06089
Column 3	0.03859	0.02057	0.05461
Column 4	0.09622	0.06510	0.07271
Column 5	0.06831	0.05347	0.11492
Column 6	0.05947	0.23464	0.23564
Column 7	0.11934	0.05872	0.12157
Column 8	0.05201	0.06290	0.13310
Column 9	0.09857	0.02255	0.23560
Column 10	0.42762	0.47831	0.46016

fixed at 10% and 15%, respectively, but the number of missing values is random. When the missing values are fewer than 3, the algorithm selected is the AAA, while the MICE is the chosen one in other cases.

Figure 1 shows the evolution of the RMSE for the two algorithms and the hybrid combination. The hybrid algorithm is not the best one for every case, but it does keep the values for the RMSE constant regardless of the number of missing values. The blue continued line represents the MICE algorithm, the red dotted line means the AAA algorithm, and the black dashed line is for the combined algorithm results.

4 Conclusions

With the proposal, it was possible to predict students' academic performance. The high level of accuracy attained could help the students to identify which subjects may represent more difficulties for them during the course. This could be a useful tool for improving different parameters such as

TABLE 5. Results for algorithm with random missing values and hybrid combination (15% of missing values).

	MAE	MAD	RMSE
Column 1	0.14507	0.11055	0.16490
Column 2	0.04801	0.09328	0.09792
Column 3	0.04459	0.03185	0.09607
Column 4	0.19098	0.08987	0.08459
Column 5	0.10614	0.08720	0.13741
Column 6	0.08390	0.45658	0.33184
Column 7	0.12288	0.10458	0.21315
Column 8	0.07592	0.10295	0.22643
Column 9	0.13944	0.03038	0.39976
Column 10	0.64091	0.65130	0.64978

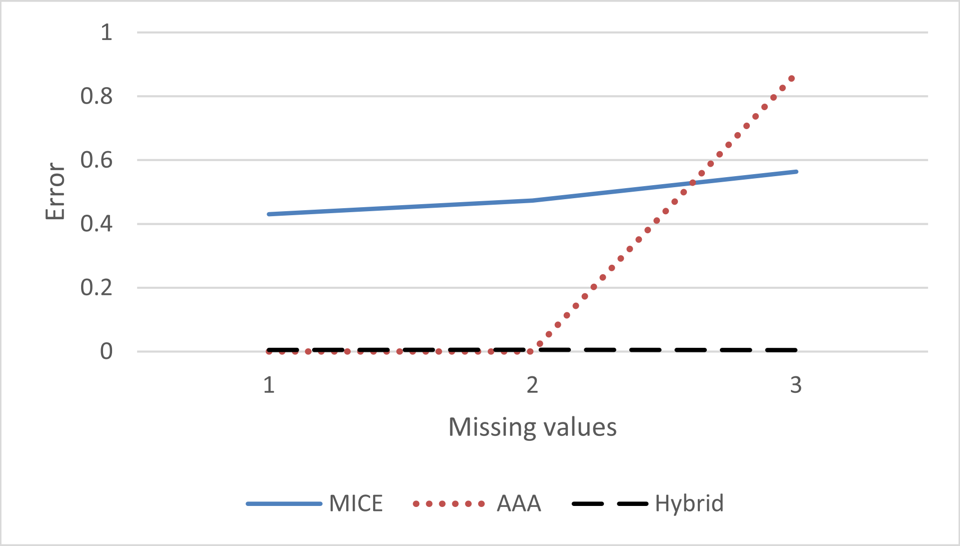

FIGURE 1. Plotting of the RMSE values for the algorithms.

academic performance, the GPA and dropout rates in all academic years. Overall, very good results have been obtained with the data imputation techniques employed in this study.

It is possible to predict the scores of the students for the three cases contemplated, assuming the data do not exist, by comparing the estimate results with the real dataset. The average of RMSE for MICE was 0.50759, varying from 2.47E-3 to 1.54849; for AAA, the average of RMSE was 0.29130, with a minimum of 3.11E-31 and a maximum of 1.29216. The hybrid combination of these two algorithms achieved a 4.92E-3 average RMSE, ranging from 4.26E-5 to 9.72E-3.

These techniques could be used to predict missing data and then undertake studies about students' performance, taking into account all the cases. As in many other cases, a hybrid algorithm performed better than the original set of algorithms employed for its construction. This is the reason why the authors consider that the use of a hybrid algorithm combining MICE and AAA is a most promising way of solving complex imputation problems.

In future research, the use of support vector machines [43, 44] and hybrid methods [44–47] will be explored by the authors in order to find a new algorithm with even higher performance than the one proposed in this research.

Acknowledgements

The authors appreciate support from the Spanish Economics and Competitiveness Ministry, through grant AYA2014-57648-P and the Government of the Principality of Asturias (Consejería de Economía y Empleo), through grant FC-15-GRUPIN14-017.

References

[1] M. Miguel Díaz. *Evaluación del Rendimiento en la enseñanza Superior: Resultados Entre Alumnos Procedentes de la LOGSE Y del COU*. Ministerio de Educación, Oviedo, 2002.

[2] F. H. G. Ferreira and J. Gignoux. The measurement of educational inequality: Achievement and opportunity. *The World Bank Economic Review*, **28**, 210–246, 2014. doi: 10.1093/wber/lht004.

[3] J. Bará, J. F. Córdoba, R. de Luis, J. Hernández and P. Martín. *Evaluación Transversal del Rendimiento Académico de Las Ingenierías Técnicas*. Ministerio de Educación, Cultura y Deporte, Madrid, 2002.

[4] J. A. Grissom, D. Kalogrides and S. Loeb. Using student test scores to measure principal performance. *Educational Evaluation and Policy Analysis*, **37**, 3–28. doi: 10.3102/0162373714523831.

[5] J. A. López-Vázquez, J.A. Orosa, J. L . Calvo-Rolle, F. J. Cos Juez, J. L. Castelerio-Roca and A. M. Costa. A new way to improve subject selection in engineering degree studies. In *International Joint Conference: CISIS'15 and ICEUTE'15*, Switzerland, 2015. doi: 10.1007/978-3-319-19713-5_47

[6] C. M. Kokkinos, A. Kargiotidis and A. Markos. The relationship between learning and study strategies and big five personality traits among junior university student teachers. *Learning and Individual Differences*, **43**, 39–47, 2015. doi: 10.1016/j.lindif.2015.08.031.

[7] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt and M. P. Wenderoth. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, **111**, 8410–8415, 2014.

[8] W. D. Cook, K. Tone and J. Zhu. Data envelopment analysis: Prior to choosing a model. *Omega*, **44**, 1–4, 2014. doi: 10.1016/j.omega.2013.09.004.

[9] E. M. Anderman, B. Gimbert, A. A. O'Connell and L. Riegel. Approaches to academic growth assessment. *British Journal of Educational Psychology*, **85**, 138–153, 2015.

[10] J. A. Romera Cabrerizo and M. Santos. ParaTrough: Modelica-based simulation library for solar thermal plants. *Revista Iberoamericana de Automática e Informática Industrial RIAI*, **14**, 412–423. doi: 10.1016/j.riai.2017.06.005.

[11] S. González, J. Sedano, J. R. Villar, E. Corchado, A. Herrero and B. Baruque. Features and models for human activity recognition. *Neurocomputing*, **167**, 52–60. doi: 10.1016/j.neucom.2015.01.082.

[12] J. L. Calvo-Rolle, O. Fontenla-Romero, B. Pérez-Sánchez and B. Guijarro-Berdiñas. Adaptive inverse control using an online learning algorithm for neural networks. *Informatica*, **25**, 401–414, 2014.

[13] J. L. Casteleiro-Roca, E. Jove, F. Sánchez-Lasheras, J. A. Méndez-Pérez, J. L. Calvo-Rolle and F. J. de Cos Juez. Power cell SOC modelling for intelligent virtual sensor implementation. *Journal of Sensors*, **2017**, 2017. doi: 10.1155/2017/9640546.

[14] J. L. Calvo-Rolle, I. Machón-Gonzalez and H. López-Garcia. Neuro-robust controller for non-linear systems. *Dyna*, **86**, 308–317, 2011. doi: 10.6036/3949.

[15] A. Ghanghermeh, G. Roshan, J. A. Orosa, J. L. Calvo-Rolle and Á. M. Costa. New climatic indicators for improving urban sprawl: A case study of Tehran City. *Entropy*, **15**, 999–1013, 2013.

[16] H. Alaiz Moretón, J. L. Calvo Rolle, I. García and A. Alonso Alvarez. Formalization and practical implementation of a conceptual model for PID controller tuning. *Asian Journal of Control*, **13**, 773–784, 2011.

[17] J. L. Casteleiro-Roca, J. L. Calvo-Rolle, M. C. Meizoso-López, A. J. Piñón-Pazos and B. A. Rodríguez-Gómez. Bio-inspired model of ground temperature behavior on the horizontal geothermal exchanger of an installation based on a heat pump. *Neurocomputting*, **150**, 90–98, 2015.

[18] J. L. Casteleiro-Roca, H. Quintián, J. L. Calvo-Rolle, E. Corchado, M. C. Meizoso-López and A. Piñón-Pazos. An intelligent fault detection system for a heat pump installation based on a geothermal heat exchanger. *Journal of Applied Logic*, **17**, 36–47, 2015.

[19] F. Alonso Zotes and M. Santos Peñas. Heuristic optimization of interplanetary trajectories in aerospace missions. *Revista Iberoamericana de Automática e Informática Industrial RIAI*, **14**, 1–15. doi: 10.1016/j.riai.2016.07.006.

[20] C. Pinzón, J. F. De Paz, J. Bajo, Á. Herrero and E. Corchado. AIIDA-SQL: An adaptive intelligent intrusion detector agent for detecting SQL injection attacks. *10th International Conference on Hybrid Intelligent Systems*, Atlanta, GA, 73–78, 2010. doi: 10.1109/HIS.2010.5600026

[21] H. Quintian Pardo, J. L. Calvo Rolle and O. Fontenla Romero. Application of a low cost commercial robot in tasks of tracking of objects. *Dyna.*, **79**, 24–33, 2012.

[22] J. L. Casteleiro-Roca, J. A. M. Pérez, A. J. Piñón-Pazos, J. L. Calvo-Rolle and E. Corchado. Modeling the electromyogram (EMG) of patients undergoing anesthesia during surgery. In *10th International Conference on Soft Computing Models in Industrial and Environmental Applications*, Switzerland, pp. 273–283, 2015. doi: 10.1007/978-3-319-19719-7_24

[23] J. Osborn, F. J. Cos Juez, D. Guzman, T. Butterley, R. Myers and A. Guesalaga. Using artificial neural networks for open-loop tomography. *Optics Express*, **20**, 2420–2434.

[24] D. Guzmán, F. J. Cos Juez, R. Myers, A. Guesalaga and F. Sánchez-Lasheras. Modeling a MEMS deformable mirror using non-parametric estimation techniques. *Optics Express*, **18**, 21356–21369.

[25] F. J. Cos Juez, F. Sánchez-Lasheras, P. J. García Nieto and M. A. Suárez Suárez. A new data mining methodology applied to the modelling of the influence of diet and lifestyle on the value of bone mineral density in post-menopausal women. *International Journal of Computer Mathematics*, **86**, 1878–1887.

[26] P. J. García Nieto, J. R. Alonso Fernández, F. Sánchez Lasheras, F. J. Cos Juez and C. Díaz Muñiz. A new improved study of cyanotoxins presence from experimental cyanobacteria concentrations in the Trasona reservoir (northern Spain) using the MARS technique. *Science of the Total Environment*, **430**, 88–92.

[27] J. L. Casteleiro-Roca, J. L. Calvo-Rolle, J. A. Méndez Pérez, N. Roqueñí Gutiérrez and F. J. de Cos Juez. Hybrid intelligent system to perform fault detection on BIS sensor during surgeries. *Sensors*, **17**, 2017. doi: 10.3390/s17010179.

[28] L. A. Fernández-Serantes, R. E. Vázquez, J. L. Casteleiro-Roca, J. L. Calvo-Rolle and E. Corchado. Hybrid intelligent model to predict the SOC of a LFP power cell type. *In International Conference on Hybrid Artificial Intelligence Systems*, Switzerland, pp. 561–572, 2014. doi: 10.1007/978-3-319-07617-1_49

[29] C. Crespo Turrado, F. Sánchez Lasheras, J. L. Calvo-Rolle, A. J. Piñón-Pazos and F. J. de Cos Juez. A new missing data imputation algorithm applied to electrical data loggers. *Sensors*, **15**, 31069–31082, 2015. doi: 10.3390/s151229842.

[30] C. Turrado, M. López, F. Lasheras, B. Gómez, J. Rollé and F. Juez. Missing data imputation of solar radiation data under different atmospheric conditions. *Sensors*, **14**, 20382–20399, 2014. doi: 10.3390/s141120382.

[31] C. O. Galán, F. S. Lasheras, F. J. de Cos Juez and A. B. Sánchez. Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions. *Journal of Computational and Applied Mathematics*, **311**, 704–717, 2017.

[32] J. H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, **19**, 1–141, 1991.

[33] J. C. Á. Antón, P. J. G. Nieto, F. J. de Cos Juez, F. S. Lasheras and C. B. Viejo. Battery state-of-charge estimator using the MARS technique. *IEEE Transactions on Power Electronics*, **28**, 3798–3805.

[34] J. De Andrés, P. Lorca, F. J. de Cos Juez and F. Sánchez-Lasheras. Bankruptcy forecasting: A hybrid approach using fuzzy c-means clustering and multivariate adaptive regression splines (MARS). *Expert Systems with Applications*, **38**, 1866–1875.

[35] A. S. Sánchez, P. R. Fernández, F. S. Lasheras, F. J. de Cos Juez and P. J. G. Nieto. Prediction of work-related accidents according to working conditions using support vector machines. *Applied Mathematics and Computation*, **218**, 3539–3552.

[36] J. R. A. Fernández, C. D. Muñiz, P. J. G. Nieto, F. J. de Cos Juez and F. S. Lasheras. Forecasting the cyanotoxins presence in fresh waters: A new model based on genetic algorithms combined with the MARS technique. *Ecological Engineering*, **53**, 68–78.

[37] L. Tierny. Introduction to general state-space Markov chain theory. In *Markov Chain Monte Carlo in Practice*, W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds, pp. 59–71. Chapman & Hall, London, UK, 1996.

[38] S. Van Buuren. *Flexible Imputation of Missing Data*. Chapman & Hall/CRC, London, UK, 2012.

[39] Y. Liu and S. D. Brown. Comparison of five iterative imputation methods for multivariate classification. *Chemom. Intell. Lab.*, **120**, 106–115, 2013.

[40] R. Perez, E. Lorenz, S. Pelland, M. Beauharnois, G. van Knowe, K. Hemker, D. Heinemannb, J. Remunde, S. C. Müllere and W. Traunmüllerf. et al. Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. *Solar Energy*, **94**, 305–326, 2013.

[41] F. V. Gutierrez-Corea, M. A. Manso-Callejo, M. P. Moreno-Regidor and J. Velasco-Gómez. Spatial estimation of sub-hour global horizontal irradiance based on official observations and remote sensors. *Sensors*, **14**, 6758–6787, 2014.

[42] P. Tiengrod and W. Wongseree. A comparison of spatial interpolation methods for surface temperature in Thailand. In *Proceedings of the International Computer Science and Engineering Conference (ICSEC)*, Nakhon Pathom, Thailand, 4-6 September, pp. 174–178. 2013.

[43] Y. Liu and S. D. Brown. Comparison of five iterative imputation methods for multivariate classification. *Chemometrics and Intelligent Laboratory Systems*, **120**, 2013. doi: 10.1016/j.chemolab.2012.11.010.

[44] P. J. García Nieto, J. R. Alonso Fernández, F. J. de Cos Juez, F. Sánchez Lasheras and C. Díaz Muñiz. Hybrid modelling based on support vector regression with genetic algorithms in forecasting the cyanotoxins presence in the Trasona reservoir (northern Spain). *Environmental Research*, **122**, 2013. doi: 10.1016/j.envres.2013.01.001.

[45] H. Quintian, J. L. Calvo-Rolle and E. Corchado. A hybrid regression system based on local models for solar energy prediction. *Informatica*, **25**, 2014. doi: 10.15388/Informatica.2014.14.

[46] X. M. Vilar-Martinez, J. A. Montero-Sousa, J. L. Calvo-Rolle and J. L. Casteleiro-Roca. Expert system development to assist on the verification of "TACAN" system performance. *Dyna.*, **89**, 2014. doi: 10.6036/5756.

[47] F. Alonso Zotes and M . Santos Penas. *Heuristic Optimization of Interplanetary Trajectories in Aerospace Missions*. Revista iberoamericana de automática e informática industrial, vol. **14**, 1–15, 2017. doi: 10.1016/j.riai.2016.07.006

Received 20 October 2017