Missing data imputation over academic records of electrical engineering students

ESTEBAN JOVE, *Department of Industrial Engineering, University of A Coruña, Avda. 19 de febrero s/n, 15405 Ferrol, A Coruña, Spain.*

PATRICIA BLANCO-RODRÍGUEZ, *Department of Construction and Manufacturing Engineering, University of Oviedo, Campus de Viesques, 33204 Gijón, Spain.*

JOSÉ-LUIS CASTELEIRO-ROCA* AND HÉCTOR QUINTIÁN, *Department of Industrial Engineering, University of A Coruña, Avda. 19 de febrero s/n, 15405 Ferrol, A Coruña, Spain.*

FRANCISCO JAVIER MORENO ARBOLEDA, *Universidad Nacional de Colombia, Facultad de Minas, Sede Medellín, Carrera 80 No 65-223, 050034 Medellín, Colombia.*

JOSÉ ANTONIO LÓPEZ-VÁZQUEZ, BENIGNO ANTONIO RODRÍGUEZ-GÓMEZ, MARÍA DEL CARMEN MEIZOSO-LÓPEZ AND ANDRÉS PIÑÓN-PAZOS, *Department of Industrial Engineering, University of A Coruña, Avda. 19 de febrero s/n, 15405 Ferrol, A Coruña, Spain.*

FRANCISCO JAVIER DE COS JUEZ, *Department of Mining Exploitation, University of Oviedo, Calle San Francisco, 1, 33004 Oviedo, Spain.*

SUNG-BAE CHO, *Department of Computer Science, Yonsei University, Seoul 03722, South Korea.*

JOSÉ LUIS CALVO-ROLLE, *Department of Industrial Engineering, University of A Coruña, Avda. 19 de febrero s/n, 15405 Ferrol, A Coruña, Spain.*

Abstract

Nowadays, the quality standards of higher education institutions pay special attention to the performance and evaluation of the students. Then, having a complete academic record of each student, such as number of attempts, average grade and so on, plays a key role. In this context, the existence of missing data, which can happen for different reasons, leads to affect adversely interesting future analysis. Therefore, the use of imputation techniques is presented as a helpful tool to estimate the

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

Vol. 00, No. 0, © The Author(s) 2019. Published by Oxford University Press. All rights reserved.
For permissions, please e-mail: journals.permission@oup.com.
doi:10.1093/jigpal/jzz056

Downloaded from https://academic.oup.com/jigpal/advance-article-abstract/doi/10.1093/jigpal/jzz056/5675544 by Yonsei University user on 16 March 2020

value of missing data. This work deals with the academic records of engineering students, in which imputation techniques are applied. More specifically, it is assessed and compared to the performance of the multivariate imputation by chained equations methodology, the adaptive assignation algorithm (AAA) based on multivariate adaptive regression splines and a hybridization based on self-organisation maps with Mahalanobis distances and AAA algorithm. The results show that proposed methods obtain successfully results regardless the number of missing values, in general terms.

Keywords: Student performance, data imputation, SOM, MARS, MICE, AAA

1 Introduction

According to guidance of quality assurance systems under the European Higher Education Area, studies tracking is not only legally regulated but it is also compulsory for official university degrees [53]. Consequently, internal quality systems of educational institutions make great efforts to enhance their quality ratios or indicators in terms of academic results and performance [20]. Faculties or higher education schools need thus tools to support or assist in this complicated task [27, 31, 37, 39].

In this context, missing data on academic records imply a significant problem that must be tackled. Then, the use of techniques to estimate missing values is presented as a good solution.

In earlier research works, common methodologies involved obtaining a model based on a historical data set, through either traditional techniques or more advanced procedures [2, 9–11, 14, 16, 19, 21, 34, 45]. This traditional method could be a problem in general terms, given the need to have previous cases with similar performance [1, 5–8, 13, 22, 25, 33, 44].

It must be emphasized that the case under study could change, so the model must be adaptive for novel cases with different casuistic and performance [4, 12, 18, 26, 32, 40, 57]. Therefore, imputation procedures based on evolutionary methods appear to be a good solution to accomplish the studies tracking task.

This paper evaluates different imputation techniques to estimate missing values on academic records of university students. In particular, the initial data set is composed by the grades and number of attempts of electrical engineering degree students. The students academic records are from the Polytechnic University College (in the University of A Coruña). The data set includes a total of 7 academic course, from 2001/2002 to 2008/2009.

The techniques employed are the adaptive assignation algorithm (AAA) [46, 48, 56], the multivariate imputation by chained equations (MICE) algorithm [52] and a hybridization based on self-organization maps (SOM) with Mahalanobis distances and AAA algorithm, which is a combination of pattern recognition and machine learning methodologies. The performance of each one on the filling in missing data task is assessed and compared.

The results obtained over a real data set show that, in general terms, for one and two missing values, the AAA algorithm offers better performance than MICE. On the other hand, with three missing values, the MICE algorithm achieves better results. Proper combination of these two algorithms appears to be a potential good solution that requires to clearly establish the boundary between application of both methodologies. Furthermore, the hybridization based on SOM, Malahanobis distances and AAA offers successfully results when it is applied to a data set with a number of missing data randomly selected.

This paper is structured as follows: after the present introduction, the case of study is detailed in the next section. Then, the data imputation techniques is used, and the approach is explained. The outcomes and the best model are exposed in the results section. Finally, the conclusions and future works are exposed.

Downloaded from https://academic.oup.com/jigpal/advance-article-abstract/doi/10.1093/jigpal/jzz056/5675544 by Yonsei University user on 16 March 2020

Downloaded from https://academic.oup.com/jigpal/advance-article-abstract/doi/10.1093/jigpal/jzz056/5675544 by Yonsei University user on 16 March 2020

2 Case of study

The present work deals the attempts prediction by missing data imputation of an engineering degree. To achieve this goal, a data set consisting on the academic records of electrical engineering degree students was used. This degree is taught at the Polytechnic University College (EUP) of the University of A Coruña (UDC) and it is structured in three courses with a total amount of 25 subjects: the first and second courses have 9 subjects and the last course has 7 subjects. The set of subjects for each degree corresponds to 236 academic credits.

To achieve the final data set used on this research, two different steps were followed. At first, the academic data of students enrolled in the EUP from 1996/1997 to 2008/2009 were considered. With these data, an initial statistical analysis was performed taking into account different factors such as access method, admission grade, average grade, gender or geographic origin. From these data, some conclusions are reached:

- The 80% of the students are male.
- The 99.4% of the students come from Galicia.
- The 91.65% of the enrolled people come from secondary education.
- The first enrolment is done at an average age of 20.86.
- The students obtain their graduate at an average age of 26.40.
- An average of 5.96 courses is needed to finish the studies.

After this step, two different groups of students were discarded: the students who enrolled the university before 2001/2002 due to lack of information about their academic records and the students that abandoned their studies in the EUP. The final data set considered on this work includes for each student the number of attempts to pass each subject and the grade achieved on each one.

The data set includes the next information:

- number of attempts to pass each subject;
- the grade achieved for each subject;
- academic background: secondary school or vocational education and training;
- admission grade.

With the aim of checking the performance of the used algorithms, different percentages of missing values were emulated. This step is vital to establish the proper boundaries of each method and hence develop a hybrid model useful in a wider range of cases.

3 Data imputation techniques

This section describes in a detailed way the data imputation methods used in the present research work.

3.1 MICE algorithm

The MICE algorithm was developed by van Buuren and Groothuis-Oudshoorn [55]. It is based on Markov chain Monte Carlo method, where the state space is the data set of all the values imputed.

Like other Markov chain methods, given the fact that the results must converge, this algorithm needs to meet the following characteristics [3, 35, 51, 54]:

- *Irreducible*: Chain has to able to comprise all elements of the space of state.

- *Aperiodic*: Chain must not fluctuate among different states.
- *Recurrence*: Any chain of this type could be considered recurrent if the likelihood of the Markov chain to start from i and return to i is equal to the unit.

From a practical point of view, the MICE algorithm convergence is obtained with a very low quantity of iterations, generally from 5 to 20 [30, 35, 38, 42, 49].

In accordance with the know-how of the method owner, five iterations are usually adequate. However, some different contingencies need more iterations.

Then, the iterations quantity for the present research work was fixed to five, a very low value compared with other Markov chain Monte Carlo methods applications, where it is necessary a large quantity of iterations, even thousands. It is very important to emphasize that each iteration of the MICE method could spend within some minutes up to hours. Also, the time spent on the iteration usually depends on the number of variables implied, and it does not depend so much on the amount of cases. Furthermore, it must be taken into account that the data imputed could have an appreciable random noise, which depends on the variables correlations.

If the variables are independent or if the correlations between them are low, then the algorithm convergence tends to be faster. It is also remarkable that high rates of missing data (more than 20 %) lead to slow convergence problems.

The necessary steps for the multivariate missing data imputation are the following [35]:

1. Specify an imputation model $P(Y_j^{mis}|Y_j^{obs}, Y_{-j}, R)$ for variable Y_j with $j = 1, \ldots, p$
2. MICE algorithm gives the later R distribution using interactive sampling from the represented conditional formula shown above. The R parameters are specific to the respective conditional densities and are not necessarily the product of a factorization of the true joint distribution.
3. For each j, fill in starting imputations Y_j^0 by random draws from Y_j^{obs}.
4. Repeat for $t = 1, \ldots, T$ (iterations).
5. Repeat for $j = 1, \ldots, p$ (variables).
6. Define $Y_{-j}^t = (Y_1^t, \ldots, Y_{j-1}^t, Y_{j+1}^{t-1}, \ldots, Y_p^{t-1})$ as the currently complete data except Y_j.
7. Draw $\phi_j^t \sim P(\phi_j^t|Y_j^{obs}, Y_{-j}^t, R)$.
8. Draw imputations $Y_j^t P(Y_j^{mis}|Y_j^{obs}, Y_{-j}^t, R, \phi_j^t)$.
9. End repeat j.
10. End repeat t.

Y represents an $n \times p$ matrix of partiallyobserved sample data, R is an $n \times p$ matrix, $0 - 1$ response indicators of Y, and ϕ represents the parameters space. Please note that with MICE imputation [41], the initial guesses for all missing elements are given for the $n \times p$ matrix of partially observed sample.

The data set is divided into two subsets for each variable with missing elements, one of them containing all the missing data.

If the subset with all the data available is in regression with all other variables, the subset missing is predicted from the regression, by the replacement of the missing values by those achieved from the regression. This methodology is repeated for all variables with missing elements.

To conclude the procedure, all missing elements are imputed according to the above explained algorithm and the regression and predictions are repeated till the stop criterion is reached or, in this case, till some quantity of successive iterations are within the defined tolerance for each of the values imputed.

Downloaded from https://academic.oup.com/jigpal/advance-article-abstract/doi/10.1093/jigpal/jzz056/5675544 by Yonsei University user on 16 March 2020

Downloaded from https://academic.oup.com/jigpal/advance-article-abstract/doi/10.1093/jigpal/jzz056/5675544 by Yonsei University user on 16 March 2020

3.2 The AAA algorithm

With the aim to explain the AAA method [17], it is necessary to assume that the data set includes a set of variables $v_1, v_2, ..., v_n$. To obtain the missing values of the $i - th$ column, all rows without missing values in that column are used and, then, some MARS models are estimated.

It is possible to find rows with a very variable quantity of missing data, from 0 (no missing values) to n (all values are missing).

Also, the columns that contain all missing values can be removed so that they will not be used for the model calculation and will not be imputed. Furthermore, a number of missing data between 0 and $n - 2$ are feasible (all variables except one with missing values).

Then, the number of MARS models that must be taken into account to estimate the missing values in the column v_i of a data set formed by n variables $v_1, v_2, ..., v_n$ is shown in (1):

$$\sum_{k=1}^{n-1} \left(\begin{array}{c} n-1 \\ k \end{array} \right). \tag{1}$$

The present research work has 10 variables, so the maximum number of MARS models is 5,110 (511 by variable). Once all the available models were obtained, the next step is to compute all the missing data of each row. To achieve this goal, the models without missing data at this row must be used. When the model was not obtained, missing data must be substituted by the column average value. Remark that the large data sets with not too high missing data percentage are infrequent.

When a value is estimated using more than one MARS model, then it must be checked the model with the highest quantity of input variables; the value could be obtained with some of those chosen models.

Exceptionally, if there is no available model for estimation, then the variable average value can be employed in the imputation process.

3.3 The hybrid proposed methodology

The algorithm described on this section is based on the hybridization of the SOM neural networks with the Mahalanobis distances. Also, the hybrid topology achieved is combined with the above shown algorithm AAA [7] that is based on MARS. The new proposed methodology performance is better than the developed on previous research when it was used with the same data set. It is considered a hybridization due to the fact that this proposal combines pattern recognition and machine learning methods, achieving a hybrid model that allows the missing data imputation [8].

If there is a data set composed by c different variables $v_1, v_2, ..., v_c$ that are the matrix columns with a total number of r rows, then the algorithm could be applied by following the steps described in the next subsections.

3.3.1 New matrix creation of missing values as from the original data set If the algorithm is applied to a data set in which the missing data are going to be imputed, then the present step of the algorithm is not necessary. However, in this research it is necessary to validate the proposal with a complete data set.

A is the original matrix ($r \times c$) of r rows and c columns. With the aim to obtain a matrix with missing data, a rate of p elements must be removed. B will be the new ($r \times c$) matrix, with a rate p of missing elements.

The removal is accomplished in a randomly manner, then the imputation method that is going to be checked is defined as missing completely at random.

3.3.2 The reduced matrix creation It is created a new matrix where all the rows that have missing data are removed. The new matrix is denominated as B^{red}. Despite the quantity of rows s ($s \leq r$) of this matrix will change depending on the matrix to be imputed, in cases where the data removal was made completely at random, and proportionally to p, the quantity of remaining rows u follows (2):

$$u = r \cdot (1 - p)^c, \tag{2}$$

where

- p: proportion of missing data taken into account;
- r: number of rows of the original matrix;
- c: number of columns of the data matrix.

Then, the B^{red} matrix is normalized.

3.3.3 The director vectors obtaining through SOM The SOM is a kind of unsupervised artificial neural network with a wide range of applications. One of these applications is the visualization and the subsequent interpretation of data sets, usually high dimensional [40].

This kind of artificial neural networks is employed to perform the available observations, with very high accuracy, thanks to a model quantity reduction. This is why this method has been selected to accomplish this work.

N is the dimension of the n director vectors, $X(t) \in \mathbb{R}_n, t = 1, 2, ..., n$. Each sample vector is labelled. As a result, a two-dimensional output is obtained, which has a rectangular mesh of $k = 1, ..., xdimydim$ nodes. All the nodes are used as a codebook vector Wk of dimension N.

The weight vector values are obtained following the next algorithm [29]. It is necessary to establish a number of iterations, following these next steps on each one:

1. Choose one sample vector $X(t)$ at random.
2. Search for the nearest weight vector: $W_c : ||X - W_c|| = min_j \left||X - W_j|\right|$.
3. Update the weights W_i by means of the rule in (3):

$$W_i(t + 1) = W_i(t) + h_{ci}(t) \cdot [X(t) - W_i(t)], \tag{3}$$

where $h_{ci}(t)$ is the neighbour function, which, in the case of this work and being very usual in the bibliography [40], is a Gaussian equation (4):

$$h_{ci}(t) = \alpha(t) \cdot exp\left(\frac{||W_c - W_i||}{2 \cdot \sigma^2(t)}\right). \tag{4}$$

The neuron weights that are at the neighbourhood $h_{ci}(t)$ of the winning neuron are displaced near to $X(t)$. When the number of iterations increases, the learning rate $\alpha(t) \in [0, 1]$ decreases monotonically. After some iterations and if the reduction of $\alpha(t)$ and $\sigma(t)$ is slow, only a single node is covered by a neighbourhood, and consequently the map is created. It is necessary to remark that if these neuron weights are close in the parameter space W, then they are also close at the mesh.

Then, the achieved director vectors are denormalized. The chosen quantity of director vectors to implement the SOM for the present algorithm is proportional to the rows quantity in the B^{red} matrix. u is the quantity of rows of the matrix B^{red}; the number of director vectors will be a range of values $d = e \cdot \frac{u}{e} \in [0.05, 0.8]$. In the result section, this range values are explained and justified.

Downloaded from https://academic.oup.com/jigpal/advance-article-abstract/doi/10.1093/jigpal/jzz056/5675544 by Yonsei University user on 16 March 2020

3.3.4 The closest director vectors obtaining with Mahalanobis distances This technique is a non-Euclidean distance measure based on correlations between variables [15]. Some patterns are identified and analysed thanks to these correlations. This indicator is a very helpful way to achieve similarities over an unknown sample set to a known set. It is used on this work to compare each one of the rows of the matrix with missing data with all the director vectors.

It is defined with (5):

$$d_A(x_1, x_2) = \sqrt{(x_1 - x_2)^T \cdot A \cdot (x_1 - x_2)}\,. \tag{5}$$

x_1 and x_2 are the sets of variables of two rows of the data matrix. $A \in \mathbb{R}^{nxn}$ is a positively semi-definite matrix, which is the inverse of the covariance matrix of class $\{I\}$. With the eigenvalue decomposition, A can be decomposed into $A = W \cdot W^T$.

For the present approach, the Mahalanobis distance of each vector row is calculated with at least two missing data points to all the director vectors. Remark that all those variables with missing data in the row that come from the data matrix are removed in the director vector. The director vector with the lowest Mahalanobis distance value is chosen and the missing variables are filled with the present values in the corresponding row of the director vector.

The original matrix is reconstructed and the missing data value of those rows with one or two missing data points are imputed with the AAA algorithm [7, 48, 57].

4 Results

To check the performance of the algorithms, different percentages of missing data were tested using the validation measurements explain in the following subsection.

4.1 Models validation

It has been employed cross-validation leave-one-out to analyse the interpolated data spatial error [28, 50]. The use of eight of the nine stations in the model is made to achieve the estimated value at the ninth, with the aim to obtain this station. The procedure is calculated once for each station. Then, nine times for all the process.

The three methods have been compared in terms of performance, thanks to useful common statistics indicators root mean square error (RMSE), mean absolute error (MAE) and median absolute deviation (MAD).

$$RMSE = \sum_{i=1}^{n} \sqrt{\frac{1}{n}(\hat{G}_i - G_i)^2} \tag{6}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{G}_i - G_i| \tag{7}$$

$$MAD = median(|G_i - median(G)|) \tag{8}$$

G_i and $\hat{G}_i$ are the measurements and the modelestimated, respectively, and n is the number of data points of the validation set. RMSE is considered a very important model validation metric and it weights large estimation errors more strongly than small estimation errors. Furthermore, MAE is a useful complement of the measured–modelled scatter plot near the 1-to-1 line [23, 41, 47].

Downloaded from https://academic.oup.com/jigpal/advance-article-abstract/doi/10.1093/jigpal/jzz056/5675544 by Yonsei University user on 16 March 2020

TABLE 1 Algorithm results with 1 missing value.

	MSE		MAD		RMSE	
	MICE	AAA	MICE	AAA	MICE	AAA
Column 1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Column 2	0.4333	0.0000	0.1482	0.0000	0.7992	0.0000
Column 3	0.1666	0.0000	0.0000	0.0000	0.3125	0.0000
Column 4	0.1333	0.0000	0.1482	0.0000	0.2230	0.0000
Column 5	0.2333	0.0000	0.0000	0.0000	0.4280	0.0000
Column 6	0.4000	0.0000	0.1482	0.0000	0.5908	0.0000
Column 7	0.1333	0.0014	0.0000	0.0018	0.2787	0.0018
Column 8	0.1666	0.0000	0.0000	0.0000	0.3604	0.0000
Column 9	0.0666	0.0000	0.0000	0.0000	0.1154	0.0000
Column 10	0.2000	0.0000	0.1482	0.0000	0.1087	0.0000

TABLE 2 Algorithm results with 2 missing values.

	MSE		MAD		RMSE	
	MICE	AAA	MICE	AAA	MICE	AAA
Column 1	0.0502	0.0000	0.0474	0.0001	0.0370	0.0000
Column 2	0.4766	0.0000	0.2179	0.0001	0.8325	0.0000
Column 3	0.2132	0.0001	0.0294	0.0001	0.3672	0.0001
Column 4	0.1848	0.0001	0.1769	0.0001	0.2562	0.0000
Column 5	0.2721	0.0000	0.0594	0.0001	0.4879	0.0001
Column 6	0.4707	0.0000	0.1877	0.0001	0.6390	0.0000
Column 7	0.1879	0.0015	0.0693	0.0019	0.3144	0.0019
Column 8	0.2083	0.0000	0.0435	0.0001	0.3973	0.0000
Column 9	0.1258	0.0000	0.0604	0.0000	0.1515	0.0001
Column 10	0.2630	0.0000	0.2144	0.0000	0.1614	0.0001

4.2 Validation results

It should be noted that the results shown in the next tables are obtained taking into account only 10 columns of the data set. The content of each column represents a different subject randomly selected. The percentage of missing data represents a 10% of the whole data set. However, the number of real missing data varied depending on the test from one to three.

The results obtained with only one missing value in each case are shown on Table 1. The AAA algorithm offers better performance than MICE in this case.

The performance of each technique for two missing values is represented in Table 2. For this test, the AAA algorithm shows again better performance than the MICE. However, both performances are decreased with respect to the previous case.

If the number of missing values is raised to 3, then MICE algorithm offers better results, as shown in Table 3.

To take advantage of the best algorithm for each situation, a 'hybrid combination' of both techniques can be taken into account. Table 4 shows the performance of the hybrid combination, where the percentage of missing values is fixed to 10% and the number of missing values is randomly

Downloaded from https://academic.oup.com/jigpal/advance-article-abstract/doi/10.1093/jigpal/jzz056/5675544 by Yonsei University user on 16 March 2020

TABLE 3 Algorithm results with 3 missing values.

	MSE		MAD		RMSE	
	MICE	AAA	MICE	AAA	MICE	AAA
Column 1	0.1629	0.8613	0.1374	0.7392	0.1462	1.0162
Column 2	0.5811	0.8746	0.3514	0.7610	0.9278	0.8031
Column 3	0.3389	0.7967	0.1134	0.9031	0.4530	1.0407
Column 4	0.2830	0.7968	0.2748	0.9244	0.3656	0.8192
Column 5	0.3680	0.6989	0.1785	1.0166	0.5548	0.7210
Column 6	0.5255	0.8510	0.3003	0.9616	0.7245	0.9048
Column 7	0.3074	0.8079	0.1905	0.7252	0.3805	0.7783
Column 8	0.3333	0.8661	0.1320	0.8535	0.4980	0.8040
Column 9	0.2277	0.8602	0.1448	1.0770	0.2629	0.8335
Column 10	0.3417	0.9510	0.3103	0.7686	0.2387	0.9942

TABLE 4 Algorithm results with random missing values and hybrid combination.

	MSE		MAD		RMSE	
	MICE	Hybrid	MICE	Hybrid	MICE	Hybrid
Column 1	0.0545	0.0059	0.0532	0.0065	0.0430	0.0049
Column 2	0.4819	0.0045	0.2229	0.0054	0.8389	0.0036
Column 3	0.2191	0.0059	0.0350	0.0028	0.3743	0.0060
Column 4	0.1905	0.0038	0.1824	0.0050	0.2630	0.0063
Column 5	0.2783	0.0051	0.0654	0.0039	0.4925	0.0079
Column 6	0.4759	0.0047	0.1929	0.0047	0.6459	0.0045
Column 7	0.1930	0.0050	0.0747	0.0082	0.3192	0.0073
Column 8	0.2120	0.0053	0.0481	0.0037	0.4018	0.0050
Column 9	0.1312	0.0033	0.0654	0.0063	0.1580	0.0051
Column 10	0.2665	0.0028	0.2199	0.0040	0.1649	0.0016

selected. If the number of missing values is less than 3, then the 'hybrid system' uses AAA algorithm while MICE is chosen otherwise. Using this configuration, the error obtained is smaller than the MICE algorithm in all cases.

Finally, the results obtained with the hybrid approach proposed on Section 3.3 are shown in Table 5. This configuration combines the unsupervised algorithm SOM with techniques that measure similarity between two random multidimensional variables and the imputation algorithm AAA.

Figure 1 represents the RMSE values of each column for the two best algorithms when the number of missing data varies randomly. It must be emphasized that MAAA algorithm outperforms the hybrid model, which is a combination of AAA and MICE algorithms.

5 Conclusions and future works

Very good results have been obtained in general terms with the data imputation techniques employed in this study. It is possible to predict the students' attempts to pass subjects in the cases discussed

Downloaded from https://academic.oup.com/jigpal/advance-article-abstract/doi/10.1093/jigpal/jzz056/5675544 by Yonsei University user on 16 March 2020

TABLE 5 Algorithm results with random missing values and the new proposed hybrid combination based in SOM, Mahalanobis distances and AAA.

	MSE		MAD		RMSE	
	MICE	MAAA	MICE	MAAA	MICE	MAAA
Column 1	0.0729	0.0042	0.0732	0.0053	0.0575	0.0030
Column 2	0.5017	0.0029	0.2379	0.0041	0.8493	0.0024
Column 3	0.2380	0.0046	0.0452	0.0009	0.3936	0.0048
Column 4	0.2045	0.0022	0.1956	0.0037	0.2827	0.0044
Column 5	0.2905	0.0037	0.0757	0.0028	0.5045	0.0060
Column 6	0.4936	0.0032	0.2064	0.0035	0.6653	0.0032
Column 7	0.2108	0.0039	0.0875	0.0072	0.3319	0.0056
Column 8	0.2288	0.0040	0.0675	0.0021	0.4166	0.0035
Column 9	0.1473	0.0015	0.0803	0.0046	0.1753	0.0037
Column 10	0.2841	0.0011	0.2312	0.0022	0.1830	0.0015

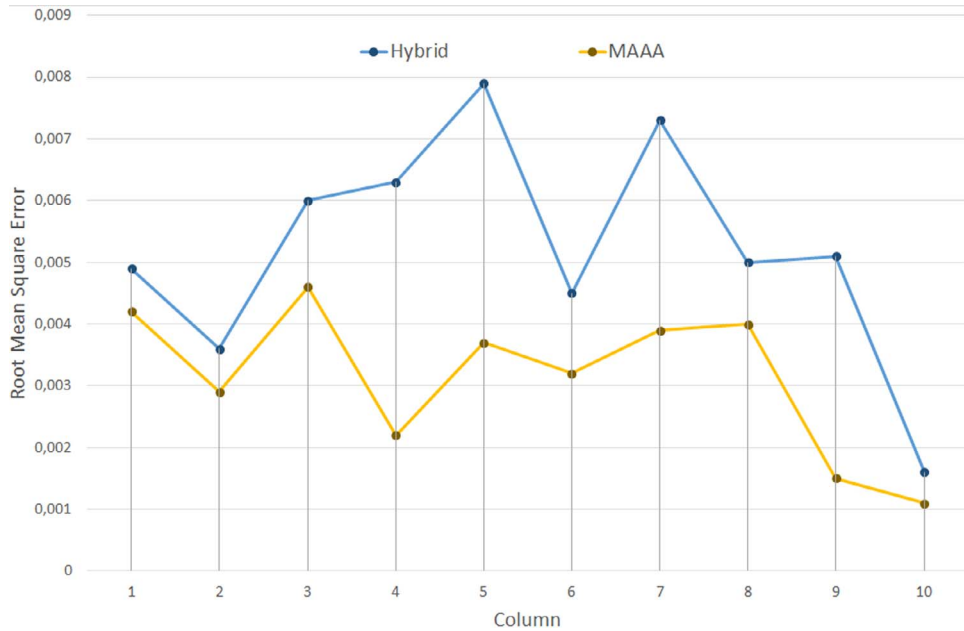

FIGURE 1 Distribution of the RMSE depending on the column and the algorithm.

in this paper, assuming that the data do not exist and comparing the estimate results with the real data set.

The AAA algorithm presents a very good performance for 1 and 2 missing values, with an RMSE value from 0.0000 to 0.0019. However, when the number of missing values increases to three, the RMSE varies from 0.7210 to 1.0407. On the other hand, the use of MICE algorithm offers better results with respect to AAA when three values are missing (RMSE from 0.1462 to 0.9278). This circumstance led to use a hybrid system that combines AAA and MICE algorithms depending on

Downloaded from https://academic.oup.com/jigpal/advance-article-abstract/doi/10.1093/jigpal/jzz056/5675544 by Yonsei University user on 16 March 2020

the number of missing values when this number is random. This hybrid configuration, outperforms MICE algorithm in all cases, with an RMSE from 0.0036 to 0.0079. Furthermore, the use of MAAA leads to even better results when the number of missing values is randomly assigned between one and three. In this case, the RMSE values vary from 0.0024 to 0.0060.

These techniques could be used to predict lack of data and to estimate students' performance and hence this would help to improve the academic records tracking and the quality standards of higher education institutions.

As future works, the use of support vector machines [36, 50] and other hybrid methods [24, 43, 58] will be considered by the authors in order to find a new algorithm with even higher performance. Also, as future works, the authors will check the procedure in other academic degrees, to ensure the generalization of the method and to study its application over some degrees at the University of A Coruña.

Acknowledgements

Authors greatly appreciate the support both from Spanish Ministry of Economy and Competitiveness through grant AYA2014-57648-P and from regional Ministry of Economy and Employment through grant FC-15-GRUPIN14-017.

This work was also supported by grant funded by 2019 IT promotion fund (Development of AI based Precision Medicine Emergency System) of the Korean government (Ministry of Science and ICT).

References

[1] 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.

[2] 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.

[3] B. Baruque, S. Porras, E. Jove and J. L. Calvo-Rolle. Geothermal heat exchanger energy prediction based on time series and monitoring sensors optimization. *Energy*, **171**, 49–60, 2019.

[4] A. G. Basden, D. Atkinson, N. A. Bharmal, U. Bitenc, M. Brangier, T. Buey, T. Butterley, D. Cano, F. Chemla, P. Clark, M. Cohen, J.-M. Conan, F. J. de Cos, C. Dickson, N. A. Dipper, C. N. Dunlop, P. Feautrier, T. Fusco, J. L. Gach, E. Gendron, D. Geng, S. J. Goodsell, D. Gratadour, A. H. Greenaway, A. Guesalaga, C. D. Guzman, D. Henry, D. Holck, Z. Hubert, J. M. Huet, A. Kellerer, C. Kulcsar, P. Laporte, B. Le Roux, N. Looker, A. J. Longmore, M. Marteaud, O. Martin, S. Meimon, C. Morel, T. J. Morris, R. M. Myers, J. Osborn, D. Perret, C. Petit, H. Raynaud, A. P. Reeves, G. Rousset, F. Sanchez Lasheras, M. Sanchez Rodriguez, J. D. Santos, A. Sevin, G. Sivo, E. Stadler, B. Stobie, G. Talbot, S. Todd, F. Vidal, E. J. Younger Experience with wavefront sensor and deformable mirror interfaces for wide-field adaptive optics systems. *Monthly Notices of the Royal Astronomical Society*, **459**, 1350–1359, 2016.

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

[6] J. L. Calvo-Rolle, H. Quintian-Pardo, E. Corchado, M. del Carmen Meizoso-López and R. Ferreiro García. Simplified method based on an intelligent model to obtain the extinction angle of the current for a single-phase half wave controlled rectifier with resistive and inductive load. *Journal of Applied Logic*, **13**, 37–47, 2015.

Downloaded from https://academic.oup.com/jigpal/advance-article-abstract/doi/10.1093/jigpal/jzz056/5675544 by Yonsei University user on 16 March 2020

[7] 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.

[8] J. L. Casteleiro-Roca, H. Quintián, J. L. Calvo-Rolle, E. Corchado, M. C. Meizoso-López and A. Pi nón Pazos. An intelligent fault detection system for a heat pump installation based on a geo-thermal heat exchanger. *Journal of Applied Logic*, **17**, 36–47, 2016.

[9] J.-L. Casteleiro-Roca, A. J. Barragán, F. Segura, J. L. Calvo-Rolle and J. M. Andújar. Fuel cell output current prediction with a hybrid intelligent system. *Complexity*, **2019**, 2019.

[10] J.-L. Casteleiro-Roca, J. L. Calvo-Rolle, J. A. M. Pérez, N. R. Gutiérrez and F. J. de Cos Juez. Hybrid intelligent system to perform fault detection on BIS sensor during surgeries. *Sensors*, **17**, 2017.

[11] J.-L. Casteleiro-Roca, E. Jove, J. M. Gonzalez-Cava, J. A. M. Pérez, J. L. Calvo-Rolle and F. B. Alvarez. Hybrid model for the ANI index prediction using remifentanil drug and EMG signal. *Neural Computing and Applications*, 1–10, 2018.

[12] 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.

[13] 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*, pp. 273–283. Springer, 2015.

[14] W. D. Cook, K. Tone and J. Zhu. Data envelopment analysis: prior to choosing a model. *Omega*, **44**, 1–4, 2014.

[15] 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, 2010.

[16] M. J. Crespo-Ramos, I. Machón-González, H. López-García and J. L. Calvo-Rolle. Detection of locally relevant variables using SOM–NG algorithm. *Engineering Applications of Artificial Intelligence*, **26**, 1992–2000, 2013.

[17] C. C. Turrado, F. S. Lasheras, J. L. Calvo-Rollé, 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.

[18] J. De Andrés, F. Sánchez-Lasheras, P. Lorca and F. J. De Cos Juez. A hybrid device of self organizing maps (SOM) and multivariate adaptive regression splines (MARS) for the forecasting of firms' bankruptcy. *Accounting and Management Information Systems*, **10**, 351–374, 2011.

[19] 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*, pp. 561–572. Springer International Publishing, 2014.

[20] F. H. G. Ferreira and J. Gignoux. The measurement of educational inequality: achievement and opportunity. *World Bank Economic Review*, **28**, 210–246, 2014.

[21] 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.

Downloaded from https://academic.oup.com/jigpal/advance-article-abstract/doi/10.1093/jigpal/jzz056/5675544 by Yonsei University user on 16 March 2020

[22] R. Ferreiro Garcia, J. L. C. Rolle, J. P. Castelo and M. R. Gomez. On the monitoring task of solar thermal fluid transfer systems using nn based models and rule based techniques. *Engineering Applications of Artificial Intelligence*, **27**, 129–136, 2014.

[23] R. F. Garcia, J. L. C. Rolle, M. R. Gomez and A. D. Catoira. Expert condition monitoring on hydrostatic self-levitating bearings. *Expert Systems with Applications*, **40**, 2975–2984, 2013.

[24] 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.

[25] 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.

[26] J. M. Gonzalez-Cava, J. Antonio Reboso, J. L. Casteleiro-Roca, J. Calvo-Rolle and J. A. M. Pérez. A novel fuzzy algorithm to introduce new variables in the drug supply decision-making process in medicine. *Complexity*, **2018**, 2018.

[27] J. A. Grissom, D. Kalogrides and S. Loeb. Using student test scores to measure principal performance. *Educational Evaluation and Policy Analysis*, **37**, 3–28, 2015.

[28] 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.

[29] D. Guzmán, F. J. de Cos Juez, R. Myers, A. Guesalaga and F. Sánchez-Lasheras. Modeling a MEMS deformable mirror using non-parametric estimation techniques. *Optics Express*, **20**, 21356–21369, 2010.

[30] E. Jove, H. Alaiz-Moretón, I. García-Rodríguez, C. Benavides-Cuellar, J. L. Casteleiro-Roca and J. L. Calvo-Rolle. PID-ITS: an intelligent tutoring system for PID tuning learning process. In *International Joint Conference SOCO'17-CISIS'17-ICEUTE'17 León, Spain, September 6–8, 2017*, pp. 726–735. Springer International Publishing, Cham, 2018.

[31] E. Jove, J. A. Lopez-Vazquez, M. I. Fernandez-Ibanez, J.-L. Casteleiro-Roca and J. L. Calvo-Rolle. Hybrid intelligent system to predict the individual academic performance of engineering students. *International Journal of Engineering Education*, **34**, 895–904, 2018.

[32] E. Jove, P. Blanco-Rodríguez, J. L. Casteleiro-Roca, J. Moreno-Arboleda, J. A. López-Vázquez, F. J. de Cos Juez and J. L. Calvo-Rolle. Attempts prediction by missing data imputation in engineering degree. In *International Joint Conference SOCO'17-CISIS'17-ICEUTE'17 León, Spain, September 6–8, 2017, Proceeding*, pp. 167–176. Springer International Publishing, Cham, 2018.

[33] E. Jove, J. M. Gonzalez-Cava, J.-L. Casteleiro-Roca, J.-A. Méndez-Pérez, J. Antonio Reboso-Morales, F. J. Pérez-Castelo, F. J. de Cos Juez and J. L. Calvo-Rolle. Modelling the hypnotic patient response in general anaesthesia using intelligent models. *Logic Journal of the IGPL*, **27**, 189–201, 2018.

[34] 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.

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

[36] Y. Liu and S. D. Brown. Comparison of five iterative imputation methods for multivariate classification. *Chemometrics and Intelligent Laboratory Systems*, **120**, 2013.

Downloaded from https://academic.oup.com/jigpal/advance-article-abstract/doi/10.1093/jigpal/jzz056/5675544 by Yonsei University user on 16 March 2020

[37] J. A. López-Vázquez, J. A. Orosa, J. L. Calvo-Rolle, F. J. Cos Juez, J. L. Castelerio-Roca and A. M. A. Costa. A new way to improve subject selection in engineering degree studies. In *International Joint Conference: CISIS 2015*, Springer, 2015.

[38] A. Marrero, J. A. Méndez, J. A. Reboso, I. Martín and J. L. Calvo. Adaptive fuzzy modeling of the hypnotic process in anesthesia. *Journal of Clinical Monitoring and Computing*, **31**, 319–330, 2017.

[39] F. Martinez-Álvarez, A. Troncoso, H. Quintián and E. Corchado. A data structure to speed-up machine learning algorithms on massive datasets francisco. In *International Conference on Hybrid Artificial Intelligence Systems*, vol. 9648, Springer, Cham pp. 365–376, 2016.

[40] J. Osborn, F. J. de Cos Juez, D. Guzmán, A. Basden, T. J. Morris, E. Gendron, T. Butterley, R. M. Myers, A. Guesalaga, F. S. Lasheras, M. G. Victoria, M. L. S. Rodriguez, D. Gratadour and G. Rousset. Open-loop tomography with artificial neural networks on CANARY: on-sky results. *Monthly Notices of the Royal Astronomical Society*, **441**, 2508–2514, 2014.

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

[42] E. Plaku and D. Le. Interactive search for action and motion planning with dynamics. *Journal of Experimental & Theoretical Artificial Intelligence*, **28**, 849–869, 2016.

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

[44] H. Quintián, J. L. Calvo-Rolle and E. Corchado. A hybrid regression system based on local models for solar energy prediction. *Informatica*, **25**, 265–282, 2014.

[45] H. Quintián, J.-L. Casteleiro-Roca, F. J. Perez-Castelo, J. L. Calvo-Rolle and E. Corchado. Hybrid intelligent model for fault detection of a lithium iron phosphate power cell used in electric vehicles. In *International Conference on Hybrid Artificial Intelligence Systems*, pp. 751–762. Springer International Publishing, 2016.

[46] H. Quintián and E. Corchado. Beta scale invariant map. *Engineering Applications of Artificial Intelligence*, **59**, 218–235, 2017.

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

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

[49] M. Thenmozhi and G. Sarath Chand. Forecasting stock returns based on information transmission across global markets using support vector machines. *Neural Computing and Applications*, **27**, 805–824, 2016.

[50] P. Tiengrod and W. Wongseree. A comparison of spatial interpolation methods for surface temperature in Thailand. In *2013 International Computer Science and Engineering Conference (ICSEC)*, pp. 174–178, IEEE, 2013.

[51] L. Tierny. Introduction to general state-space Markov chain theory. *Markov Chain Monte Carlo in Practice*, 59–71, 1996.

[52] C. Turrado, M. López, F. Lasheras, B. Gómez, J. L. Calvo-Rolle and F. J. Cos-Juez. Missing data imputation of solar radiation data under different atmospheric conditions. *Sensors*, **14**, 20382–20399, 2014.

[53] http:// www. ehea. info/, 2017. Online; accessed 19-March-2017.

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

Downloaded from https://academic.oup.com/jigpal/advance-article-abstract/doi/10.1093/jigpal/jzz056/5675544 by Yonsei University user on 16 March 2020

[55] S. Van-Buuren and K. Groothuis-Oudshoorn. Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, **45**, 1–67, 2011.

[56] R. Vega Vega, H. Quintián, J. L. Calvo-Rolle, Á. Herrero and E. Corchado. Gaining deep knowledge of android malware families through dimensionality reduction techniques. *Logic Journal of the IGPL*, **27**, 160–176, 2018.

[57] J. A. Vilán Vilán, J. R. Alonso Fernández, P. J. García Nieto, F. Sánchez Lasheras, F. J. de Cos Juez and C. Díaz Muñiz. Support vector machines and multilayer perceptron networks used to evaluate the cyanotoxins presence from experimental cyanobacteria concentrations in the trasona reservoir (northern Spain). *Water Resources Management*, **27**, 3457–3476, 2013.

[58] 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**, 112–121, 2014.

Received 00 Month 20xx

Downloaded from https://academic.oup.com/jigpal/advance-article-abstract/doi/10.1093/jigpal/jzz056/5675544 by Yonsei University user on 16 March 2020