# Column-wise Guided Data Imputation

Alessio Petrozziello[*] and Ivan Jordanov
*University of Portsmouth, Portsmouth, U.K.*
*Alessio.petrozziello@port.ac.uk, Ivan.jordanov@port.ac.uk*

**Abstract**

This paper investigates data imputation techniques for pre-processing of dataset with missing values. The current literature is mainly focused on the overall accuracy, evaluated estimating the missing values on the dataset at hand, however the predictions can be suboptimal when considering the model performance for each feature. To address this problem, a Column-wise Guided Data Imputation method (cGDI) is proposed. Its main novelty resides in the selection of the most suitable model from a multitude of imputation techniques for each individual feature, through a learning process on the known data. To assess the performance of the proposed technique, empirical experiments have been conducted on 13 publicly available datasets. The results show that cGDI outperforms two baselines and has always comparable or greater estimation accuracy over four state-of-the-art methods, widely applied to solve the problem at hand. Furthermore, cGDI has a straightforward implementation and any other known imputation technique can be easily added.

*Keywords:* Missing Data; Data Imputation; Multitude of imputation models.

# 1 Introduction

Most real world datasets contain missing data due to either sensors failures or human errors and dealing with it is an important step in the dataset pre-processing phase, since most statistical analysis techniques, data reduction tools, and machine learning methods require complete sets. The mechanisms of missingness are usually categorized into three groups (Enders, 2010): MCAR (Missing Completely at Random), MAR (Missing At Random) and MNAR (Missing Not At Random) and the approaches for dealing with missingness include (Enders, 2010): simple deletion (list-wise, attribute, and pairwise deletion); univariate imputation (*Random Guessing*, and *Mean/ Median Imputation*) (Sarro et al., 2016); and multivariate imputation. The last group includes methods that consider the correlation of the attributes. In this work, four different algorithms of this family are investigated: *Multiple Imputation Chained Equations* (MICE) (Lee & Mitra, 2016); *Bagged Tree Imputation* (BTI) (Frènay & Verleysen, 2014); *K-Nearest Neighbour Imputation* (KNNI) (Troyanskaya, et al., 2001) and *Bayesian Principal Component Analysis Imputation* (bPCA) (Schmitt et al., 2015). These methods have been widely applied and compared in the past few years showing discordant results (Gòmez-Carracedo et al., 2014), (Schmitt et al., 2015), (Jordanov et al., 2016). Most approaches of dealing with missingness would select a single method that outperforms the others based on a given performance measure. However, while a given approach might have the best performance across the whole dataset, it does not mean that it will be superior at the level of each individual feature. In the proposed approach, instead of selecting a single method which outperforms the others on the whole dataset, a column-wise selection is used to choose the best imputation method for each attribute of the dataset. To do that, we initially use the subset with complete data only. Then, we artificially introduce a percentage of missing data in it, which subsequently is imputed using the above mentioned methods. For each feature, the method that produced the lowest estimation error is then used to impute the missing values for the correspondent attribute. We propose a *Column-wise Guided Data Imputation* (cGDI) method which performance is compared with two baseline techniques (*Random Guessing* and *Median Imputation*) and four state-of-the-art methods (MICE, BTI, KNNI, and bPCA). The cGDI is extensively tested and validated on 13 publicly available datasets with a large degree of diversity (size and number of attributes) and its performance is assessed and compared with the other techniques using Wilcoxon Signed-rank test for statistical significance (Cohen et al., 2013). The rest of the paper is organized as follows. Section 2 proposes the cGDI method. Section 3 discusses the empirical study carried out. The results of this investigation are critically analyzed in Section 4 and finally, in Section 5 conclusion is given.

[*] Correspondent author

# 2 The Proposed Method

All methods described in (Schmitt et al., 2015), (Jordanov et al., 2016) have been widely used for solving missing data problems. However, while a given approach may produce low estimation error for the whole dataset at hand, this does not mean that the method produces the best results (min error) for every individual feature (usually, for some of the features other methods give better estimates). The investigated here *Column-Wise Data Imputation* (cGDI) is an approach which ensemble "weak" models choosing the best one for each feature (column) of the dataset (accepting that in the same time the chosen model may be 'weak' for the other features). In other words, when building an ensemble, the best imputation method for each feature of the dataset is selected among the "weak" techniques, and then included into the ensemble. During the learning phase, the algorithm is trained on artificially introduced missing data, and then, the combination of methods that performed best, is used to impute the missing values in the initial dataset. The complete subset (without missingness) is used for training the model introducing a percentage of MCAR (e.g., 25%) in each column. Once the data are imputed with each "weak" technique, an error function (e.g., RMSE, MAE) is used to select the best imputation method for each column of the dataset. To cope with the random nature of the algorithm and to ensure a more robust choice, this process is iterated for a given number of times, and the algorithm which produced the lowest median overall error for each feature is then chosen. For example, let's assume a set of *m* imputation methods ($M_1$, ..., $M_m$ $\in$ S) and dataset (*X*) composed of *v* variables and *n* samples, where *k* of them ($0 < k < n$) contain at least one missing value. Once the *n-k* complete samples (*X'* subset) are separated from those with missing values, a % of MCAR is added to each variable of *X'* (e.g., 25%). The missing data in *X'* are separately imputed using all methods of S, and the estimation error (e.g., RMSE) is calculated for each feature. This process is repeated *I* times (e.g., *I = 5*), and for each variable in *X'*, the imputation algorithm scoring the lowest median error is added to the ensemble (*E*). The ensemble of those techniques is then used to estimate the missing values of the whole set *X*. In particular, $\forall$ $M_i \in E$, $i = 1,..,m$, the dataset *X* is entirely imputed, and only the imputed values for the features where $M_i$ scored the lowest error are saved, discarding the others. Since *X* is imputed independently with each technique, the order of imputation is irrelevant, enabling the process to be parallelized.

# 3 Empirical Study

In this section, the design of the empirical study used to test and validate the proposed approach is presented. Firstly, the research questions that promoted this study are discussed, then the validation criteria and performance metrics are analysed, followed by the description of experimental settings and used datasets.

After extensive research of the literature to identify one imputation method able to win on every dataset, we found that there are discordant performance results regarding the four discussed data imputation techniques. Furthermore, a preliminary empirical analysis (see Section 4) highlighted that the performance of the considered techniques vary for different datasets as well as for each feature. These findings led to the investigation of our ensemble idea.

The proposed method (cGDI) is compared with the given univariate baseline and multivariate state-of-the-art (KNN, BTI, MICE and bPCA) imputation methods to assess its performance on the missing data estimation task. The results are reported in Section 4.

Variety of metrics employed for comparing and evaluating data imputation and predictive models can be found in the literature (Pan, et al., 2011). Among them, *Mean Squared Error* (MSE) and variants as *Root Mean Squared Error* (RMSE) and *Normalized Root Mean Squared Error* (NRMSE) are the most largely used. These metrics measure the difference between predicted and actual values while the two variants are used to mitigate the magnitude problem (taking the root of the error) and normalize the errors in the interval [0, 1]. The *Mean Absolute Error* (MAE) is argued to be more accurate and informative than the RMSE (Willmott, 2005), successively refuted by (Chai & Draxler, 2014), who states that the two measures picture different aspects of the error and therefore they should both be used to assess the results. The *Standard Accuracy* (SA) is argued to be good baseline estimation measures (Whigham et al., 2015). Some of these measures are used to evaluate the performance of the proposed method in this work. In particular, as suggested in (Willmott, 2005) and (Chai & Draxler, 2014), RMSE and MAE are implemented to compare the estimated missing values and the original ones, reflecting the average performance of the imputation method. Furthermore, the RMSE is employed as error function for the training phase of the cGDI. SA is used to compare the proposed model with the univariate baseline imputation techniques (discussed earlier). In particular, SA compares the prediction against the mean of a random sampling of the training response values (*SA = 1 – (RMSE(predicted,actual)/RMSE(randomGuess,actual))*).

| Model | Hyper-parameters |
|---|---|
| KNNI | K = 10, Distance = Euclidean |
| BTI | #Trees = 200 |
| MICE | #Iterations = 10 |
| bPCA | Method = Bayesian, Nboot = 5, Lstart = 1000, L = 100 |
| cGDI | #Iteration of training set = 5, Error function = RMSE |

**Table 1** Hyper-parameters setting

To validate the proposed method, a k-fold cross validation is applied, splitting the dataset into independent training and test sets. The test set is generated using a uniform sampling without repetitions, and the rest of the data is left as a training set. Since the *Shapiro Test* showed that many of our patterns came from non-normally distributed populations, the statistical *Wilcoxon Signed Rank Test* was used to prove which method is giving better performance (Cohen et al., 2013). Furthermore, the used test does not make any assumptions about the underlying distribution of the data. In this work, the following *NULL* hypothesis is tested: “The RMSEs (MAEs) provided by model $M_i$ are significantly smaller than the errors provided by model $M_j$", using a confidence level $\alpha=0.05$.

A 5-fold (80% training and 20% testing) cross validation is used to validate the proposed method. To calibrate the model during the training phase, 25% MCAR is added to each attribute of the training set, subsequently imputed using the five imputation techniques and the accuracy is evaluated using the RMSE. This process is run 5 times and for each attribute, the imputation model achieving the lowest median error (preferred to the mean due to robustness to outliers) is selected. Lastly, the ensemble of selected techniques is used to impute the data on the independent test set and the results are compared to all the other methods. Table 1 shows the hyper-parameters used for each algorithm.

| Dataset | #Instances | R | I | C | Dataset | #Instances | R | I | C |
|---|---|---|---|---|---|---|---|---|---|
| *Contraceptive* | 1474 | 0 | 9 | 0 | *PageBlock* | 5472 | 4 | 6 | 0 |
| *Yeast* | 1484 | 8 | 0 | 0 | *Ring* | 7400 | 20 | 0 | 0 |
| *RedWine* | 1599 | 11 | 0 | 0 | *TwoNorm* | 7400 | 20 | 0 | 0 |
| *Car* | 1728 | 0 | 0 | 6 | *PenBased* | 10992 | 0 | 16 | 0 |
| *Titanic* | 2201 | 3 | 0 | 0 | *Nursery* | 12960 | 0 | 0 | 8 |
| *Abalone* | 4174 | 7 | 0 | 1 | *Magic04* | 19020 | 10 | 0 | 0 |
| *WhiteWine* | 4898 | 11 | 0 | 0 | | | | | |

**Table 2** Datasets used in our empirical study. The last three columns show the number and type of attributes (R - Real, I - Integer, C – Categorical).

Thirteen publicly available datasets from KEEL repositories (Alcalá-Fdez, et al., 2011) are used in this work, namely *Contraceptive, Yeast, Red wine, Car, Titanic, Abalone, White Wine, Page Block, Ring, Two Norm, Pen Based, Nursery,* and *Magic04*. The selection of these datasets from the repositories’ “classification” area was driven by the intent to cover different application domains and data characteristics. In particular, the datasets differ in the number of instances (1484 to 19020), the number of features (3 to 20), and range and type of the features (real, integer and categorical). The selected datasets do not have missing values by default. The introduction of synthetic MCAR guarantees the reliability of the estimation through the experiments and the assessment and evaluation of the results. Table 2 provides descriptive statistics for each dataset, and more details about the features can be found in (Alcalá-Fdez, et al., 2011).

# 4 Results and Discussion

Preliminary results highlighted that the performance of the considered techniques vary for different datasets as well as for each feature. The imputation with the single techniques showed that for each attribute there is a different winner; hence, if one method is selected as the overall "best" - it will not be superior for every feature of the dataset. As a whole, the KNNI prevailed on one dataset (*Pen Based*), BTI on four datasets (*Car, Red Wine, White Wine,* and *Ring*), and bPCA on six datasets (*Contraceptive, Yeast, Titanic, Abalone, Two Norm,* and *Magic04*). In rare cases, the dataset at hand would have the same best imputation model for all the features (e.g., *Car* and *Two Norm*), and normally each feature will have different best imputation model.

| **Dataset** | **cGDI** | **KNNI** | **BTI** | **MICE** | **bPCA** | **Median** |
|---|---|---|---|---|---|---|
| *Abalone* | 0.68 | 0.62 | 0.57 | 0.66 | 0.72 | 0.28 |
| *Pen Based* | 0.54 | 0.56 | 0.49 | 0.47 | 0.45 | 0.27 |
| *Page Block* | 0.49 | 0.41 | 0.43 | 0.39 | 0.46 | 0.25 |
| *Magic04* | 0.47 | 0.42 | 0.41 | 0.32 | 0.45 | 0.28 |
| *Contraceptive* | 0.39 | 0.24 | 0.36 | 0.18 | 0.38 | 0.26 |
| *Red Wine* | 0.37 | 0.33 | 0.33 | 0.23 | 0.32 | 0.30 |
| *White Wine* | 0.36 | 0.34 | 0.34 | 0.16 | 0.34 | 0.28 |
| *Titanic* | 0.35 | 0.26 | 0.34 | 0.05 | 0.34 | 0.28 |
| *Two Norm* | 0.34 | 0.24 | 0.32 | 0.07 | 0.34 | 0.30 |
| *Yeast* | 0.33 | 0.24 | 0.32 | 0.06 | 0.33 | 0.28 |
| *Car* | 0.29 | 0.12 | 0.29 | -0.01 | 0.29 | 0.25 |
| *Ring* | 0.31 | 0.24 | 0.29 | -0.02 | 0.29 | 0.28 |
| *Nursery* | 0.30 | 0.09 | 0.25 | 0.00 | 0.29 | 0.23 |

**Table 3** Standard Accuracy (SA) values achieved by cGDI, the baseline (Median Imputation) and state-of-the-art (KNNI, BTI, MICE, SVD and bPCA) techniques over the 13 datasets for 5-fold cross validation with 25% MCAR. Higher values represent better estimation over the random guess

The *SA* values given in Table 3 show superior results for the imputation carried out with our model. It outperformed the baseline methods *Random Guessing* ($SA_{cGDI} > 0$) and the *Median Imputation* ($SA_{cGDI} > SA_{Median}$). The *Mean Imputation* was omitted in favor of the *Median Imputation*, since the latter is considered less biased to outliers. To finally assure that the proposed method is outperforming the baselines, a *Wilcoxon* test for statistical significance is run, testing the *NULL hypothesis* "The RMSEs provided by cGDI are significantly smaller than the errors produced by the models *Random Guessing* and *Median Imputation*". The results proved cGDI being better than both with *p-value* $< 0.05$ over all 13 datasets.

| **Model** | **win** | **tie** | **loss** |
|---|---|---|---|
| cGDI | 40 (37) | 9 (12) | 3 (3) |
| bPCA | 31 (24) | 9 (14) | 12 (14) |
| BTI | 26 (19) | 12 (15) | 14 (18) |
| KNNI | 15 (19) | 3 (11) | 34 (22) |
| MICE | 3 (4) | 5 (8) | 44 (40) |

**Table 4** RMSE (MAE) significance test for 5-fold cross validation with 25% MCAR in the test set. Each row shows how many times the model $M_i$ is better (win), comparable (tie), or worse (loss) than the other models in a Wilcoxon Signed Rank Test, with NULL Hypothesis "The RMSEs (MAEs) provided by $M_i$ are significantly smaller than the errors provided by the other models"

The *Standard Accuracy* analysis (Table 3) shows that the cGDI method not only outperforms the baselines, but it is also comparable, and even better than the state-of-the-art algorithms. As it can be seen from the table, the $SA_{cGDI}$ is higher than the *SA* of the other methods in 45 out of the 52 cases, comparable in 5 out of the 52 cases, and worse in only 2 case. To validate the significance of the difference, the *Wilcoxon* test is run justifying the *NULL hypothesis* "The RMSEs provided by cGDI are significantly smaller than the errors achieved by the state-of-the-art methods". Results in Table 4 show that the imputation improvement achieved by cGDI is significant (*p-value* $< 0.05$) in 40 out of the 52 cases, comparable in 9 out of the 52 cases and worse in 3 cases only. As suggested in (Willmott, 2005) the same *NULL* hypothesis was tested using the MAE metric. The cGDI resulted significantly better in 37 cases, comparable in 12 and worse in only 3 cases. The second-best imputation method (bPCA) for RMSE is significantly better in 31 out of the 52 cases, comparable in 9 and worse in 12 case, which shows an improvement for cGDI of 17% over the best single method. For the *MAE hypothesis*, bPCA results are significantly better in 24 out of the 52 cases, comparable in 14 and worse in 14 cases, showing 25% superiority for the cGDI over the best single method.

# 5 Conclusion

Missing data represents an important problem for datasets used in machine learning tasks, statistical analysis, and any other process requiring a complete set. Several models have been proposed in the literature, mainly focusing on the overall imputation accuracy. An initial analysis carried on 13 datasets showed that a model scoring the lowest overall error does not necessarily provide the best imputation for each feature of the dataset. For this reason, a *Column-wise Guided Data Imputation* method is introduced and proposed in this paper. Its novelty lies in its approach, which pairs each method with a feature in an attribute-wise fashion. The cGDI divides the complete records (without missing values) from those with missingness, and selects (using learning applied to the complete subset), the most suitable imputation method for each feature. The imputation performance is evaluated with four widely used imputation tasks metrics (SA, RE, RMSE, and MAE). The results are statistically assessed using the *Shapiro Test* to check the distribution normality, and the non-parametric *Wilcoxon Signed Rank Test*, validating the following *NULL hypothesis*: “The RMSEs (MAEs) provided by model $M_i$ are significantly smaller than the errors provided by model $M_j$" ($i, j = 1,\dots n$, where $n$ is the number of investigated models), using confidence level $\alpha=0.05$. The *Standard Accuracy* analysis shows cGDI to always have better accuracy than the two baselines and to produce superior estimation over the single state-of-the-art methods in 41 out of the 52 cases. The *Wilcoxon* on MAE and RMSE shows improvements of 25% and 17% respectively for the proposed method over the second best performing algorithm (bPCA). The results achieved in this work strongly suggest that the use of the proposed approach can be beneficial when considering multivariate imputation as a way of dealing with missingness. Another advantage of the cGDI approach is its straightforward implementation and easy incorporation of other known imputation methods.

# References

Alcalá-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., García, S., Sánchez, L., & Herrera, F. (2011). KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. *Journal of Multiple-Valued Logic and Soft Computing, 17*(2-3), 255-287.

Chai, T., & Draxler, R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?--Arguments against avoiding RMSE in the literature. *Geoscientific Model Development, 7*(3), 1247-1250.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences.* Routledge.

Enders, C. K. (2010). *Applied missing data analysis.* Guidford: Guildford Press.

Frènay, B., & Verleysen, M. (2014). Classification in the presence of label noise: a survey. *Neural Networks and Learning Systems, IEEE Trans. on, 25*(5), 845-869.

Gòmez-Carracedo, M., Andrade, J., Lòpez-Mahìa, P., Muniategui, S., & Prada, D. (2014). A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. *Chemometrics and Intelligent Laboratory Systems, 134*, 23-33.

Jordanov, I., Petrov, N., & Petrozziello, A. (2016). Supervised Radar Signal Classification. *Neural Networks (IJCNN), 2016 Int. Joint Conf. on* (p. 1464-1471). Vancouver: IEEE.

Lee, M. C., & Mitra, R. (2016). Multiply imputing missing values in data sets with mixed measurement scales using a sequence of generalised linear models. *Computational Statistics & Data Analysis, 95*, 24-38.

Pan, X.-Y., Tian, Y., Huang, Y., & Shen, H.-B. (2011). Towards better accuracy for missing value estimation of epistatic miniarray profiling data by a novel ensemble approach. *Genomics, 97*(5), 257-264.

Sarro, F., Petrozziello, A., & Harman, M. (2016). Multi-Objective Software Effort Estimation. *Software Engineering (ICSE), 2016 IEEE/ACM 38th IEEE Int. Conf. on*, (p. 619-630). Austin.

Schmitt, P., Mandel, J., & Guedj, M. (2015). A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics, 6*(1), 1-6.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., . . . Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics, 17*(6), 520-525.

Whigham, P. A., Owen, C. A., & Macdonell, S. G. (2015). A baseline model for software effort estimation. *ACM Trans. on Software Engineering and Methodology (TOSEM), 24*(3), 20.

Willmott, C. J. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research, 30*(1), 79-82.