Data Imputation for Symbolic Regression with Missing Values: A Comparative Study

Baligh Al-Helali, Qi Chen, Bing Xue, and Mengjie Zhang
School of Engineering and Computer Science
Victoria University of Wellington, P.O. Box 600, Wellington 6140, New Zealand
Email: {baligh.al-helali; qi.chen; bing.xue; mengjie.zhang}@ecs.vuw.ac.nz

***Abstract*—Symbolic regression via genetic programming is considered as a crucial machine learning tool for empirical modelling. However, in reality, it is common for real-world data sets to have some data quality problems such as noise, outliers, and missing values. Although several approaches can be adopted to deal with data incompleteness in machine learning, most studies consider the classification tasks, and only a few have considered symbolic regression with missing values. In this work, the performance of symbolic regression using genetic programming on real-world data sets that have missing values is investigated. This is done by studying how different imputation methods affect symbolic regression performance. The experiments are conducted using thirteen real-world incomplete data sets with different ratios of missing values. The experimental results show that although the performance of the imputation methods differs with the data set, CART has a better effect than others. This might be due to its ability to deal with categorical and numerical variables. Moreover, the superiority of the use of imputation methods over the commonly used deletion strategy is observed.**

***Index Terms*—*symbolic regression*; *genetic programming*; *incomplete data*; *imputation*.**

I. Introduction

Symbolic regression (SR) aims to discover mathematical expressions that model a target variable in terms of input features from a given data set [1]. Compared with traditional regression methods, symbolic regression has the advantage of "white box" modelling without pre-assumptions [2]. Therefore, symbolic regression has a wide range of applications in many areas [3]. Although there are several methods to perform symbolic regression [4], the most widely used method is genetic programming. Genetic programming (GP) is one of a collection of biological-inspired techniques called evolutionary computation (EC). GP solves a given task by generating subsequent generations of computer programs using genetic operators such as crossover and mutation [1].

From a machine learning perspective, learning algorithms should perform the required tasks on given data sets. However, real-world data might be incomplete which hinders the ability of many methods to learn properly [5]. Incomplete data can contain three kinds of missingness: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [6].

The methods for dealing with data incompleteness can be classified into four main approaches. The first approach is to delete any instance (feature) that contain missing values then learn using only the remaining complete data portion [7]. Another approach is to utilise model-based procedures to model data distribution such as expectation—maximisation (EM) [8]. Thirdly, some learning methods can directly deal with incomplete data without explicitly estimating the missing values such as C4.5 [9], fuzzy approaches [10], and ensemble methods [11]. The fourth approach is called imputation, in which the missing values are firstly replaced by estimated values then the learning is carried out using the complete imputed data set [12], [13]. Data imputation is the process of estimating missing values and it is categorised into single imputation and multiple imputation [6].

Unlike classification, the investigation of symbolic regression with incomplete data has not received adequate efforts. In fact, the most common approach to dealing with the incomplete issue in the symbolic regression research community is the deletion approach. However, there are recent attempts to address this issue by utilising different learning techniques such as transfer learning [14], [15]. Although some studies propose imputation methods for symbolic regression with missing values [16], [17], [12], [18], to the best of our knowledge, no study has conducted a comparison between the existing imputation methods when performing symbolic regression with incomplete data. Such a comparison might help guide researchers on empirical practices they can follow in similar situations.

In this work, we present an empirical comparison between different widely used imputation methods on the performance of symbolic regression with real-world incomplete data sets. The specific objectives of this work include:

1) Investigating the impact of imputation approaches for symbolic regression on real-world data sets that contain missing values.
2) Comparing different state-of-the-art imputation methods regarding the symbolic regression performance.
3) Providing an insightful analysis for the experimental results, which includes the ability of different methods to select incomplete features.

The rest of this paper is organised as follows. Section II introduces the background and reviews the related work. In Section III, the experiment setup is presented. The experimental results are given and analysed in Section IV. Finally, Section V concludes this work and provides future directions.

978-1-7281-2547-3/20/$31.00 ©2020 IEEE

II. Background and Related Work

This section introduces a background of the related topics with a brief literature review.

A. Symbolic Regression via Genetic Programming

Genetic programming (GP) generates automatically computer programs for performing a user-defined task [19]. It starts from a high-level definition of the problem and creates a population of random programs then refines them progressively using variation and selection strategies until getting a satisfactory solution. There are several advantages of GP [20]. Firstly, there is no need for prior knowledge of the structure of the solution. Another advantage is representing the solutions using a formal language (symbolic expressions) which is suitable for human reasoning. Therefore, symbolic regression is typically performed using GP.

In GP-based symbolic regression, each individual in the population represents a potential solution to the underlying problem, which needs to specify [21]:

- The terminal set: inputs that can be constants or variables.
- The function set: functions that are domain-specific and combined with the terminal set for constructing potential solutions to the considered problem.
- The fitness function: it is a numeric value measuring how the solution is appropriate to the problem in question.
- The control parameters set: this set includes the probabilities of the crossover and the mutation and the size of the population.
- The termination criterion: it is a predefined parameter to state when the evaluation process should stop such as the number of generations and the fitness error tolerance.

It should be noted that the first three components above are responsible for determining the search space, while the other two affect the search quality and efficiency.

B. Missing data and imputation methods

There are three kinds of missingness mechanism: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [6]. In MCAR, the events that cause missingness are independent both of unobservable and observable knowledge, i.e. there is no relationship between the missingness of a data value and other data set values, existing or missing. MAR implies that the data missingness is not related to the missing data itself, rather, it is related to some other existing data. If the missingness is neither MAR nor MCAR, it is called MNAR and this means that the cause of missingness is related to the missing data.

To deal with incomplete data, several approaches have been used [6]. The main approaches are deletion, model-based, implicit learning, and imputation. Imputation is the process of estimating missing values in incomplete data sets. There are two imputation approaches: single imputation and multiple imputation [6]. In single imputation, each missing value is replaced with a single estimation directly. However, multiple imputation estimates the imputed values from multiple responses using statistical analysis.

There are several imputation methods from different approaches [5]. One common imputation method is called hot-deck where the imputed data is taken from a similar record selected randomly from the data set. By contrast, cold-deck obtains the value from another data set. Another imputation technique is the mean imputation where the missing value in a feature is replaced by the mean of this feature's values of complete instances. However, more advanced methods are commonly used and these are the methods considered in this work. The imputation using k-nearest neighbour (KNN) is a modification for hot-deck imputation as it imputes the missing value considering the k most similar instances. The linear regression model (LM) is also used to impute the missing values. A decision tree-based method called classification and regression trees (CART) is used for imputation by employing the sum of the squared errors as a criterion for splitting the regression tree. Another method for adopting the decision trees approach is random forest (RF). It is an ensemble method, that constructs a collection of decision trees providing the output as the mode of individual trees in the classification or the mean of individual trees in the regression. More details on the imputation methods can be found in [6].

To evaluate the performance of an imputation method, there are two main approaches: the modelling approach and the prediction approach [22]. The modelling approach evaluates the imputation method according to its impact on the main learning process, e.g. classification, regression, and clustering. The better learning performance, the better the imputation performance. On the other hand, the prediction approach evaluates the method based on its accuracy of predicting the missing values. This approach is usually used with synthetic incompleteness as it requires knowing the original ground truth values of the missing ones to be able to measure the accuracy. Therefore, it is not used in real-world incompleteness situations and, subsequently, it is not used in this work as only data sets with real missingness are considered.

C. Symbolic regression with missing data

The most commonly used strategy to deal with the incomplete data in symbolic regression research is to delete the instances having missing values. This approach is used in [23] to investigate the generalisation and bloat in symbolic regression. However, they used the Auto-MPG data set which has a few instances containing missing data. This approach is too risky when there is a high ratio of incomplete data. Another simple way to deal with missing values is to fill it with corresponding feature values from other instances. In [17], the prediction of time series and symbolic regression was improved using Affine Arithmetic. They conducted experiments on two-time series, corresponding to wind speed records in the southern Brazilian cities. Both series have missing values. To deal with such an issue, those missing entries are simply replaced by replicating the previous observations.

More advanced imputation-based methods are also presented to deal with the presence of missing values when performing symbolic regression. [12] study proposes an im-

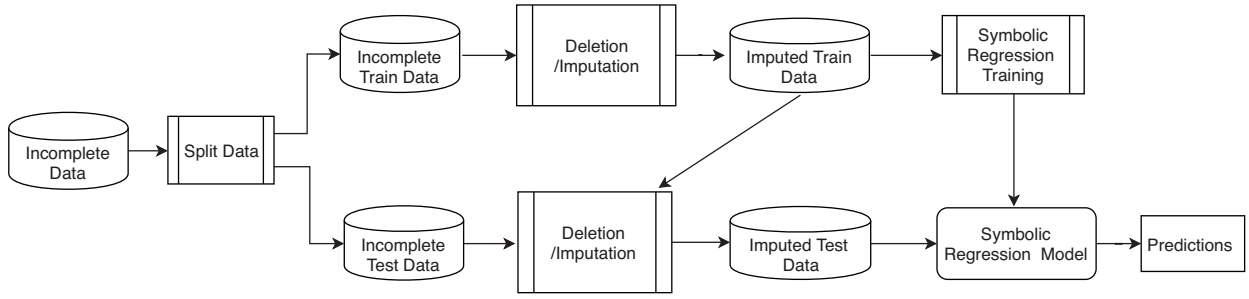


Fig. 1. The overall system of data imputation for symbolic regression with missing values.

putation method based on combining KNN and GP-based imputation. This method is evaluated using both synthetic and real incomplete data sets. In [24], [25], [26], GP-based feature selection methods are proposed for selecting imputation predictors in symbolic regression with missing values. These predictors are features that are used to construct GP imputation models. In [27], the predictor selection is based on arithmetic complexity measures. However, these methods focused on synthetic incompleteness as their main objective is to examine the effectiveness of the methods using controlled parameters.

From this brief review, no study has been conducted to compare well-known imputation methods when performing symbolic regression using real-world data sets that have real incompleteness. To achieve that, we present this study.

III. Experiment Setup

This section describes the settings considered when conducting the experiments in this work.

The purpose of the experiments is to evaluate the impact of using different imputation methods when conducting GP-based symbolic regression on real-world incomplete data sets. To achieve this goal, thirteen real-world regression data sets that have different percentages of missing values are used. Table I shows the statistics of the used data sets and the reader is referred to [28] for more details.

As most data sets are not originally split, the first step is to divide them into (70:30) training and test sets. The imputation/deletion processes are then performed on these sets independently. After that, the training data set is used to build the symbolic regression model and the obtained model is evaluated on the unseen test data set. The adopted methodology is shown in Fig. 1.

For each experiment, 100 independent runs are performed and the error of the best individual is obtained for each run (i.e. 100 best-of-run programs). The statistics of these errors are then aggregated to evaluate the performance. The Wilcoxon non-parametric statistical significance paired test is used to measure the significance of the differences between the results at a significance level of 0.05. The metric used to measure the prediction performance accuracy is the normalised root mean square error (NRMSE) which is calculated as:

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_i - \hat{y}_i)^2}}{y_{max} - y_{min}} \quad (1)$$

where n is number of instances, y_i is the predicted value of the i^{th} instance, $\hat{y}_i$ is the desired value of the i^{th} instance, and y_{max} and y_{min} are the maximum and minimum training target values, respectively.

In addition to the deletion strategy, the used imputation methods are [6]:

- Linear model (LM): this method uses a linear regression algorithm to fit a feature with missing values as the dependent variable and the other features as independent variables. LM is widely used for regression tasks in many areas. However, it requires presumptions such as linear relationship, no auto-correlation, multivariate normality, and homoscedasticity [29].
- K-Nearest Neighbour (KNN): KNN approximates a missing value by the values of its closest k points, based on other variables. It is useful for different kinds of missing data as it can be used with continuous, discrete, ordinal, and categorical data.

TABLE I
Statistics of the used data sets

Data set	#Features	#Instances	#Incomplete Instances	%Missingness ratio
Auto-mpg	7	398	6	1.58
SkillCraft1	19	3395	57	1.68
PRSA	9	43824	2067	4.72
Imports-85	15	205	54	26.34
ShanghaiPM	12	52584	29310	55.74
ChengduPM	12	52584	30684	58.84
ShenyangPM	12	52584	32404	61.62
GuangzhouPM	12	52584	32510	61.82
BeijingPM	12	52584	33227	63.19
CCN	122	1994	1676	84.05
CCUN	125	1994	1676	84.05
Wiki	53	913	737	89.72
AirQuality	13	8991	8164	90.80

- Classification and regression trees (CART): CART is used for estimating the missing values using classification and regression decision trees based on the complete features.
- Random forest (RF): random forest (RF) adopts the decision trees approach to regress the features having missing values considering the other features as predictive variables.

These imputation methods are implemented under the R package, simple imputation (Simputation) [30]. For the genetic programming setting, Table II shows the values for the GP parameters. The implementation of symbolic regression is carried out using the Python package DEAP [31].

TABLE II
THE USED VALUES FOR GP PARAMETERS

Parameter	Value
Generations	100
Population size	512
Crossover rate	0.9
Mutation rate	0.1
Elitism	Top 10 individuals
Selection Method	Tournament
Tournament size	7
Maximum depth	17
Initialisation	Ramped-half and half
Function set	+, -, *, protected div
Terminal set	features and constants $\in (-1, 1)$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Symbolic Regression Performance

For each used data set, the experimental results obtained by applying different methods to deal with missing data are shown. The mean, the standard deviation (Std), and the median of NRMSEs achieved by the best-of-run programs when using different methods on test data sets are shown in Table III. The best results are highlighted in bold and the worst are underlined.

For both the import-85 and Auto-mpg data sets, the best results are obtained when using the linear model (LM) and the worst are those of using the KNN method. Such findings may indicate that there is a linear relationship between the features having missing values and the fully observed features. In contrast, the features' values of different instances may not be highly related which is possibly the reason behind the poor results of the KNN method.

For the data sets CCUN, PRSA, BeijingPM, and ChengduPM, the best results are obtained when using KNN. This may indicate that the similarity between the instances is higher than the relationships between the features since the KNN estimates the missing values based on observed values from other similar instances. However, the other methods predict the missing values in a specific feature using regression models based on other features.

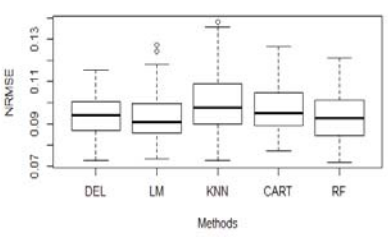
For the CNN data set, using the deletion strategy results in the worst test errors and the RF imputation method produces the best symbolic regression performance although there is no significant difference among the results of using different imputation methods. Such a result may be due to the high amount of missing values in this data set.

The most successful method is the CART method. It achieves the best results on more data sets than any other method. The CART method has the best results on the data sets SkillCraft, Wiki, AirQuality, ShenyangPM, and ShanghaiPM. However, surprisingly, the worst imputation method is RF. In fact, as an ensemble method, RF was expected to achieve the best results in many cases. The reason for these results can be related to the implementation of imputation methods. KNN and CART are designed to deal with mixed data whereas RF is used as a regression-based imputation method, so, CART and KNN provide better imputation.

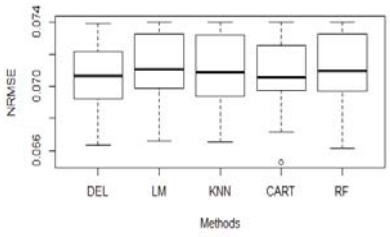
To have an overall analysis, whisker box plots are used to compare the methods by drawing their error results. Fig. 2 shows the box plot of the error distributions of the 100 best-of-run symbolic regression models with different imputation methods on the test data sets. For the SkillCraft1 data set (Fig. 2b), it can be seen that according to the median, which is represented by the line in the box referring to the centre of

TABLE III
THE SYMBOLIC REGRESSION TEST RESULTS AFTER USING DIFFERENT IMPUTATION METHODS.

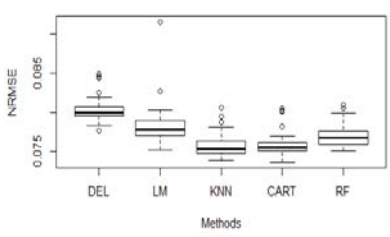
Data	Measure	DEL	LM	KNN	CART	RF
Auto-mpg	Mean	0.0939	**0.0928**	0.1001	0.0974	0.0936
	Std	0.0093	0.0104	0.0146	0.0112	0.0111
	Median	0.0942	**0.0908**	0.0975	0.0949	0.0926
SkillCraft	Mean	0.0709	0.0712	0.0710	**0.0709**	0.0712
	Std	0.0019	0.0020	0.0020	0.0019	0.0020
	Median	0.0706	0.0710	0.0709	**0.0706**	0.0709
PRSA	Mean	0.0804	0.0785	**0.0759**	0.0760	0.0770
	Std	0.0018	0.0029	0.0017	0.0018	0.0016
	Median	0.0800	0.0779	**0.0753**	0.0756	0.0768
Imports-85	Mean	0.0941	**0.0901**	0.0958	0.0914	0.0907
	Std	0.0108	0.0082	0.0114	0.0076	0.0078
	Median	0.0936	**0.0887**	0.0963	0.0903	0.0902
ShanghaiPM	Mean	0.1061	0.0688	0.0682	**0.0681**	0.0687
	Std	0.0022	0.0005	0.0008	0.0027	0.0005
	Median	0.1056	0.0685	0.0681	**0.0679**	0.0686
ChengduPM	Mean	0.0673	0.0525	**0.0521**	0.0528	0.0525
	Std	0.0016	0.0009	0.0009	0.0032	0.0008
	Median	0.0675	0.0525	**0.0519**	0.0520	0.0524
ShenyangPM	Mean	0.0794	0.0681	0.0668	**0.0666**	0.0685
	Std	0.0009	0.0006	0.0009	0.0008	0.0009
	Median	0.0794	0.0681	0.0669	**0.0663**	0.0681
GuangzhouPM	Mean	**0.0578**	0.0588	0.0590	0.0587	0.0592
	Std	0.0010	0.0007	0.0009	0.0006	0.0008
	Median	**0.0574**	0.0584	0.0589	0.0584	0.0594
BeijingPM	Mean	0.0824	0.0817	**0.0790**	0.0794	0.0813
	Std	0.0032	0.0020	0.0020	0.0018	0.0019
	Median	0.0822	0.0813	**0.0789**	0.0795	0.0812
CCN	Mean	0.2083	0.1626	0.1623	0.1629	**0.1620**
	Std	0.0182	0.0071	0.0074	0.0075	0.0077
	Median	0.2039	0.1619	0.1607	0.1611	**0.1607**
CCUN	Mean	0.0101	0.0059	**0.0027**	0.0040	0.0063
	Std	0.0149	0.0022	0.0021	0.0057	0.0022
	Median	0.0064	0.0051	**0.0015**	0.0019	0.0055
Wiki	Mean	0.1704	0.1644	0.1599	**0.1508**	0.1817
	Std	0.0111	0.0107	0.0089	0.0048	0.0274
	Median	0.1712	0.1649	0.1583	**0.1495**	0.1720
AirQuality	Mean	0.1447	0.1573	0.1509	**0.1291**	0.1418
	Std	0.0225	0.0313	0.0343	0.0297	0.0261
	Median	0.1419	0.1656	0.1511	**0.1286**	0.1384



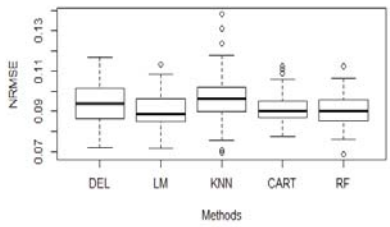
(a) Auto-mpg data set



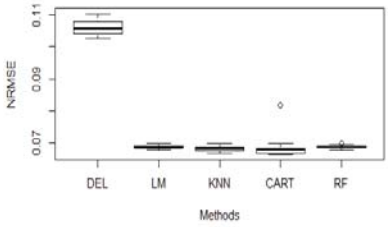
(b) SkillCraft1 data set



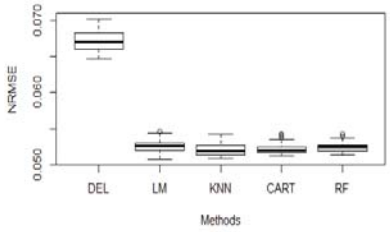
(c) PRSA data set



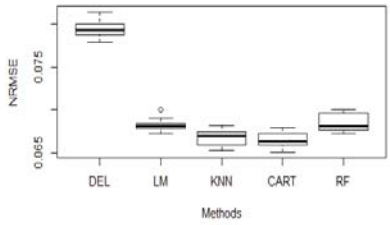
(d) Imports-85 data set



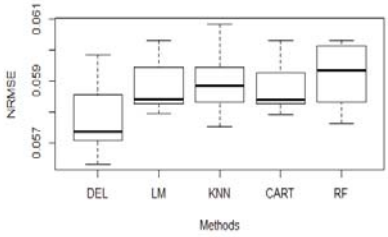
(e) ShanghaiPM data set



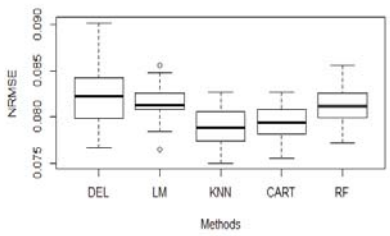
(f) ChengduPM data set



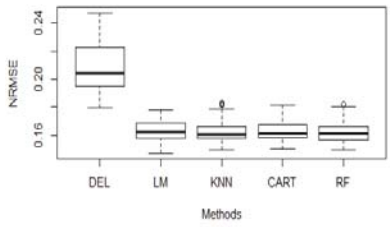
(g) ShenyangPM data set



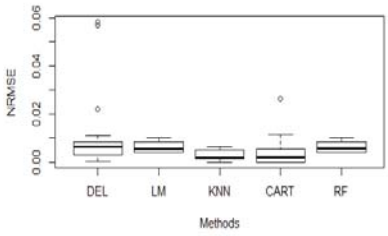
(h) GuangzhouPM data set



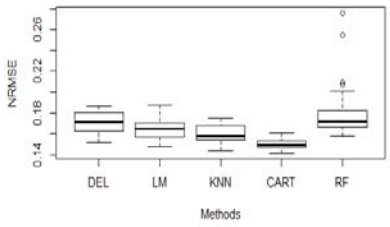
(i) BeijingPM data set



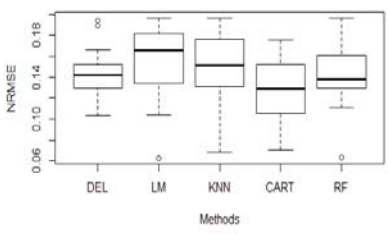
(j) CCN data set



(k) CCUN data set



(l) Wiki data set



(m) AirQuality data set

Fig. 2. Boxplots of best-of-run symbolic regression errors for different imputation methods on the test data sets

the results, the CART method gives slightly better results than the others. The median obtained using this method is smaller and the distribution of the results is represented by a shorter box. However, strangely, the deletion seems to produce good results as its median is very close to that of CART and the distribution is more stable around the median.

Fig. 2d shows the distribution of the results obtained using the Imports-85 data set and Fig. 2a for the Auto-mpg data set. According to these figures, the LM method has lower centres. However, for the Imports-85 data, the CART and RF methods result in shorter boxes referring to more consistent generalisation errors for the 100 runs. On these data sets, the KNN method is outperformed by the other imputation methods as it has the longest box plots with more outliers. For the CCN data set (Fig. 2j), there is a big difference when comparing the deletion strategy with any imputation method. Such significantly poor performance may be due to the high ratio of the incomplete instances. Deleting these instances causes the loss of more than 1670 samples representing about 84% of the given examples.

When comparing the imputation methods, the RF method produces the worst performance and we can see its unstable behaviour as the maximum values are extremely far from the box areas. On the other hand, even when it does not provide the smallest mean and median errors, the CART method is the most stable and consistent one among the imputation methods. The obtained skewed results refer to the non-normal distribution and outliers that affect the statistical results (e.g. the average). The method LM provides good results on small data sets such as Imports-85 and Auto-mpg.

B. Significance Comparisons

In this section, the Wilcoxon test is used to measure the significance of the differences between the results obtained using different methods. The comparison outcomes are shown in Table IV. The symbol "+" ("-") is used to refer that the method in the column is significantly better (worse) than the row-header method which means that the column method wins (losses) the one in the row, while "=" means there is no significant difference.

A summary of the comparisons is shown in Table V. The total number of wins (losses) are 13 (39), 20 (27), 29 (19), 39 (9), and 18 (27) for the methods DEL, LM, KNN, CART, and RF, respectively. The most wining method is CART as it outperforms the other methods 39 times. Moreover, it has the least number of losses. In contrast, RF is the worst symbolic regression performance among the imputation methods. RF has the least win cases and the most loss cases comparing to other imputation methods. The performance of using a different method is not related only to the used method, but also it is affected by the nature of the data set.

Interestingly, although RF and CART are DT-based methods and RF is supposed to be superior as an ensemble of DTs, CART achieved better performance. This seems to be due to the ability to deal with different data types. The RF-based imputation works as a random forest regressor, whereas CART

TABLE IV
THE SIGNIFICANCE COMPARISON BASED ON THE SYMBOLIC REGRESSION PERFORMANCE AFTER USING DIFFERENT IMPUTATION METHODS ON EACH REAL-WORLD INCOMPLETE DATA SET.

DS	Method	DEL	LM	KNN	CART	RF
Auto-mpg	DEL	=	+	-	-	+
	LM	-	=	-	-	-
	KNN	+	+	=	+	+
	CART	+	+	-	=	+
	RF	-	+	-	-	=
SkillCraft	DEL	=	-	-	+	-
	LM	+	=	+	+	+
	KNN	+	-	=	+	=
	CART	-	-	-	=	-
	RF	+	-	=	+	=
PRSA	DEL	=	+	+	+	+
	LM	-	=	+	+	+
	KNN	-	-	=	=	-
	CART	-	-	=	=	-
	RF	-	-	+	+	=
Imports-85	DEL	=	+	-	+	+
	LM	-	=	-	-	-
	KNN	+	+	=	+	+
	CART	-	+	-	=	=
	RF	-	+	-	=	=
ShanghaiPM	DEL	=	+	+	+	+
	LM	-	=	+	+	=
	KNN	-	-	=	+	-
	CART	-	-	-	=	-
	RF	-	=	+	+	=
ChengduuPM	DEL	=	+	+	+	+
	LM	-	=	+	+	=
	KNN	-	-	=	=	-
	CART	-	-	=	=	-
	RF	-	=	+	+	=
ShenyangPM	DEL	=	+	+	+	+
	LM	-	=	+	+	=
	KNN	-	-	=	+	-
	CART	-	-	-	=	-
	RF	-	=	+	+	=
GuangzhouP	DEL	=	-	-	-	-
	LM	+	=	-	=	-
	KNN	+	+	=	+	-
	CART	+	=	-	=	-
	RF	+	+	+	+	=
BeijingPM	DEL	=	+	+	+	+
	LM	-	=	+	+	=
	KNN	-	-	=	-	-
	CART	-	-	+	=	-
	RF	-	=	+	+	=
CCN	DEL	=	+	+	+	+
	LM	-	=	+	+	+
	KNN	-	-	=	-	=
	CART	-	-	+	=	+
	RF	-	-	=	-	=
CCUN	DEL	=	+	+	+	-
	LM	-	=	+	+	-
	KNN	-	-	=	-	-
	CART	-	-	+	=	-
	RF	-	+	+	+	=
Wiki	DEL	=	+	+	+	-
	LM	-	=	+	+	-
	KNN	-	-	=	+	-
	CART	-	-	-	=	-
	RF	+	+	+	+	=
AirQuality	DEL	=	-	-	+	+
	LM	+	=	+	+	+
	KNN	+	-	=	+	+
	CART	-	-	-	=	-
	RF	-	-	-	+	=

TABLE V
SUMMARY OF THE SIGNIFICANCE COMPARISONS.

Aggregation of the results					
	DEL	LM	KNN	CART	RF
+	13	20	29	39	18
-	39	27	19	9	27

works as both a classifier and a regressor based on the targeted variable.

It can be noticed that the symbolic regression performance after deleting the instances with missing values has the lowest number of wins against the other methods. Moreover, having better results when using the deletion approach in some cases does not mean that it is better to perform symbolic regression after getting rid of the instances having missing values. That is, the incomplete data may be faced frequently in some real applications and this situation is unavoidable. Hence, the incompleteness situation can occur in small data sets, where it is not affordable to lose more data.

According to [32], in real-world scenarios, bad generalisation can be a result of data incompleteness. The assumption that the training set is representative for all possible instances might not hold in the case of data missingness. In fact, while it is possible to learn models from complete instances ignoring the incomplete ones, these models cannot be applied to predict a new test instance that has missing values. If the training set does not contain examples for all system states that are likely to be observed, it is not possible to create a model that will generalise well to all new observations.

C. Furthermore Analysis

To have a deeper analysis, the Imports-85 data set is considered as it has comparatively a moderate ratio of incomplete instances ($\approx 25\%$). As shown earlier(see Table III), the deletion strategy results in the worst learning ability however the least test error occurred when performing symbolic regression after using the CART imputation method.

The Imports-85 data set has 15 numeric features and five of them have missing values. These features are F1, F9, F10, F12, and F13 with missingness ratios of 20%, 19.5%, 19.5%, 9.8%, and 9.8%, respectively. When considering the selection of the incomplete features in the symbolic regression models learned after using different imputation methods, Fig. 3 shows the frequency of each feature in the best GP individual over the 100 runs. This figure shows that the selection possibility of these features is less when adopting the deletion strategy than of using most imputation methods.

For more clarification, Table VI shows the frequency average of the incomplete features when using different imputation methods. This table shows that using the deletion method resulted in the lowest frequency for the features F1, F10, and F13, and the second-lowest frequency for the feature F12. When considering all the features including the complete ones, the most commonly selected features are F8 and F3. In fact, F8 is the most frequently selected feature regardless of the used imputation method. This means that the method used to

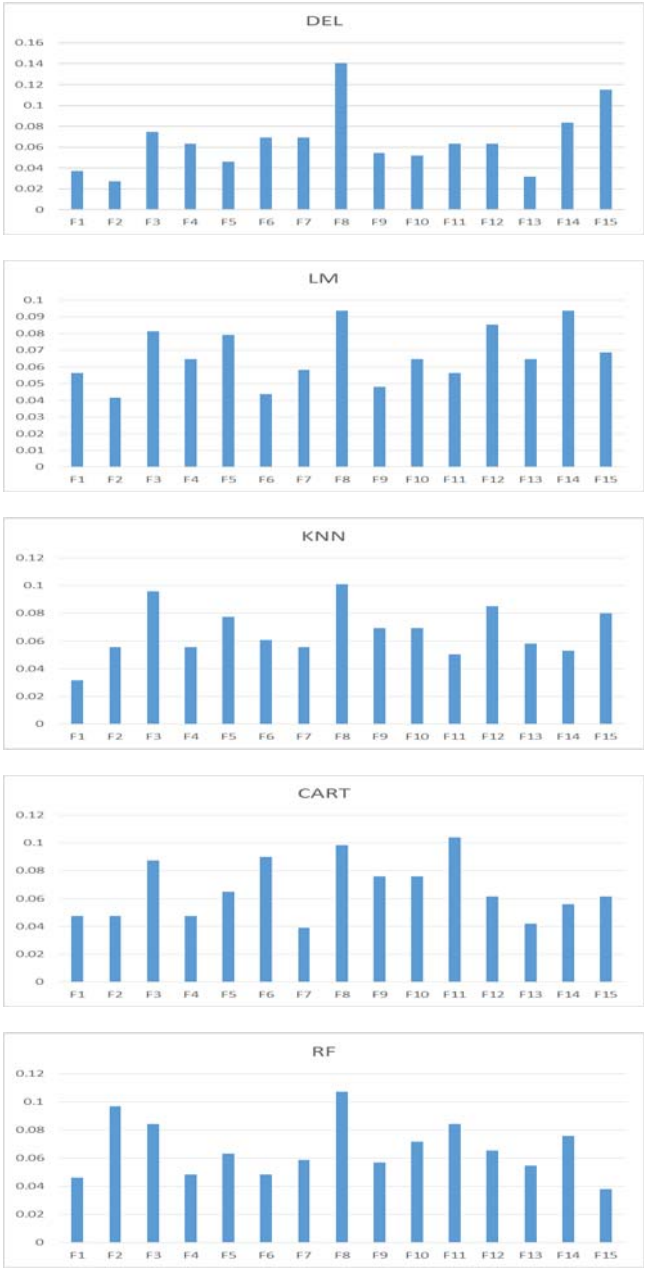


Fig. 3. The probability of selecting the Imports-85 data set features in the best symbolic regression models when using different methods.

TABLE VI
THE SELECTION FREQUENCY FOR THE INCOMPLETE FEATURES IN THE BEST SYMBOLIC REGRESSION MODELS WHEN USING DIFFERENT METHODS ON THE IMPORTS-85 DATA SET,

	DEL	**LM**	**KNN**	**CART**	**RF**
F2	0.0275	0.0416	0.0559	0.0478	0.0968
F9	0.0546	0.0479	0.0691	0.0758	0.0568
F10	0.0517	0.0646	0.0692	0.0759	0.0716
F12	0.0632	0.0854	0.0851	0.0618	0.0652
F13	0.0316	0.0646	0.0585	0.0421	0.0547

deal with missing values might not have a large impact on the selectability of complete features.

V. Conclusions and Future Work

In this work, the impact of popular imputation methods on GP-based symbolic regression with missing values is investigated. Moreover, these methods are also compared to the deletion strategy, where learning is after getting rid of instances that have missing values. The investigations show that the deletion strategy which is commonly used when dealing with missing values in the symbolic regression community is most likely to result in worse performance than imputing the missing values. Moreover, getting good results with the deletion strategy in some situations might not reflect the true pattern of the whole data set.

One main conclusion of this work is that, as the incompleteness issue is common in real-world applications, adopting imputation approaches has a positive effect when conducting symbolic regression on incomplete data. When comparing the imputation methods, although there does not exist a single method that is good for all data sets, the CART method achieved relatively good results compared with KNN, LM, and RF. This can be due to its ability to deal with mixed data more than other methods. Mixed data sets have features from different types (e.g. categorical and numerical features) which suppresses the effectiveness of regression-based imputation methods such as RF and LM.

For future work, more experimental work should be done to investigate the impact of different missingness mechanisms, i.e. MAR, MNAR, and MCAR on the performance of symbolic regression with incomplete data. For this purpose, synthetic incomplete data sets can be generated with different ratios and missingness kinds. The use of imputation methods can be then studied and analysed with more statistical evidence. Moreover, other factors such as data types could be more deeply analysed.

References

[1] J. R. Koza, *Genetic Programming II, Automatic Discovery of Reusable Subprograms*. MIT Press, Cambridge, MA, 1992.

[2] Q. Chen, B. Xue, and M. Zhang, "Genetic programming for instance transfer learning in symbolic regression," *IEEE Transactions on Cybernetics*, 2020.

[3] R. Rueda, L. G. B. Ruíz, M. P. Cuéllar, and M. Pegalajar, "An ant colony optimization approach for symbolic regression using straight line programs. application to energy consumption modelling," *International Journal of Approximate Reasoning*, 2020.

[4] J. Žegklitz and P. Pošík, "Benchmarking state-of-the-art symbolic regression algorithms," *Genetic Programming and Evolvable Machines*, pp. 1–29, 2020.

[5] K. Kleinke, J. Reinecke, D. Salfrán, and M. Spiess, "Missing data methods," in *Applied Multiple Imputation*. Springer, 2020, pp. 53–83.

[6] A. R. T. Donders, G. J. Van Der Heijden, T. Stijnen, and K. G. Moons, "A gentle introduction to imputation of missing values," *Journal of clinical epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006.

[7] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," *Neural Computing and Applications*, vol. 19, no. 2, pp. 263–282, 2010.

[8] Z. Ghahramani and M. I. Jordan, "Supervised learning from incomplete data via an em approach," in *Advances in neural information processing systems*, 1994, pp. 120–127.

[9] J. R. Quinlan, *C4.5: programs for machine learning*. Elsevier, 2014.

[10] M. R. Berthold and K.-P. Huber, "Missing values and learning of fuzzy rules," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 171–178, 1998.

[11] S. Krause and R. Polikar, "An ensemble of classifiers approach for the missing feature problem," in *Int. Joint Conf on Neural Networks*, vol. 1, 2003, pp. 553–556.

[12] B. Al-Helali, Q. Chen, B. Xue, and M. Zhang, "A hybrid GP-KNN imputation for symbolic regression with missing values," in *Australasian Joint Conference on Artificial Intelligence*. Springer, 2018, pp. 345–357.

[13] P. D. Allison, "Multiple imputation for missing data: A cautionary tale," *Sociological methods & research*, vol. 28, no. 3, pp. 301–309, 2000.

[14] B. Al-Helali, Q. Chen, B. Xue, and M. Zhang, "Multi-tree genetic programming for feature construction-based domain adaptation in symbolic regression with incomplete data," in *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, 2020, pp. 913–921.

[15] ——, "Multi-tree genetic programming-based transformation for transfer learning in symbolic regression with highly incomplete data," in *2020 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2020, pp. 1–8.

[16] T. Brandejsky, "Model identification from incomplete data set describing state variable subset only–the problem of optimizing and predicting heuristic incorporation into evolutionary system," in *Nostradamus 2013: Prediction, Modeling and Analysis of Complex Systems*. Springer, 2013, pp. 181–189.

[17] C. Pennachin, M. Looks, and J. de Vasconcelos, "Improved time series prediction and symbolic regression with affine arithmetic," in *Genetic Programming Theory and Practice IX*. Springer, 2011, pp. 97–112.

[18] B. Al-Helali, Q. Chen, B. Xue, and M. Zhang, "A genetic programming-based wrapper imputation method for symbolic regression with incomplete data," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2019, pp. 2395–2402.

[19] N. F. McPhee, R. Poli, and W. B. Langdon, "Field guide to genetic programming," 2008.

[20] Q. Chen, B. Xue, and M. Zhang, "Rademacher complexity for enhancing the generalisation of genetic programming for symbolic regression," *IEEE Transactions on Cybernetics*, 2020.

[21] J. R. Koza, "Genetic programming as a means for programming computers by natural selection," *Statistics and computing*, vol. 4, no. 2, pp. 87–112, 1994.

[22] C. T. Tran, M. Zhang, and P. Andreae, "Multiple imputation for missing data using genetic programming," in *Proceedings on genetic and evolutionary computation*. ACM, 2015, pp. 583–590.

[23] G. Dick, "Bloat and generalisation in symbolic regression," in *Asia-Pacific Conference on Simulated Evolution and Learning*. Springer, 2014, pp. 491–502.

[24] B. Al-Helali, Q. Chen, B. Xue, and M. Zhang, "Genetic programming for imputation predictor selection and ranking in symbolic regression with high-dimensional incomplete data," in *Australasian Joint Conference on Artificial Intelligence*. Springer, 2019, pp. 523–535.

[25] ——, "Genetic programming-based simultaneous feature selection and imputation for symbolic regression with incomplete data," in *Asian Conference on Pattern Recognition*. Springer, 2019, pp. 566–579.

[26] ——, "Genetic programming with noise sensitivity for imputation predictor selection in symbolic regression with incomplete data," in *2020 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2020, pp. 1–8.

[27] ——, "Hessian complexity measure for genetic programming-based imputation predictor selection in symbolic regression with incomplete data," in *European Conference on Genetic Programming (Part of EvoStar)*. Springer, 2020, pp. 1–17.

[28] D. Dheeru and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[29] M. A. Poole and P. N. O'Farrell, "The assumptions of the linear regression model," *Transactions of the Institute of British Geographers*, pp. 145–158, 1971.

[30] M. van der Loo, "simputation: Simple imputation," *R package version 0.2*, vol. 2, 2017.

[31] F.-A. Fortin, F.-M. D. Rainville, M.-A. Gardner, M. Parizeau, and C. Gagné, "Deap: Evolutionary algorithms made easy," *Journal of Machine Learning Research*, vol. 13, no. Jul, pp. 2171–2175, 2012.

[32] G. Kronberger, *Symbolic regression for knowledge discovery: bloat, overfitting, and variable interaction networks*. Trauner, 2011.