Indonesian Journal of Electrical Engineering and Computer Science
Vol. 15, No. 2, August 2019, pp. 1009~1018
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v15.i2.pp1009-1018

Performance evaluation of listwise deletion for impaired datasets in multiple regression-based prediction

Chanintorn Jittawiriyanukoon
Assumption University, Thailand

Article Info

Article history:

Received Jan 16, 2019
Revised Mar 7, 2019
Accepted Mar 13, 2019

Keywords:

Missingness
Multiple regression-based prediction
Performance evaluation
Root mean squared error
Simulation

ABSTRACT

Multiple Regression-Based Prediction (MRBP) is an emerging calculation to or analysis technique cope with the future by compiling the history of data. The MRBP characteristic will include an approximation for the associations between physical observations and predictions. MRBP is a predictive model, which will be an important source of knowledge in terms of an interesting trend to be followed in the future. However, there is impairment in the MRBP dataset, wherein each form of missing and noisy data has caused an error and is unavailable further analysis. To overcome this unavailability, so that the data analytics can be moved on, two treatment approaches are introduced. First, the given dataset is denoised; next, listwise deletion (LD) is proposed to handle the missing data. The performance of the proposed technique will be investigated by dealing with datasets that cannot be executed. Employing the Massive Online Analysis (MOA) software, the proposed model is investigated, and the results are summarized. Performance metrics, such as mean squared error (MSE), correlation coefficient (COEF), mean absolute error (MAE), root mean squared error (RMSE), and the average error percentage, are used to validate the proposed mechanism. The proposed LD projection is confirmed through actual values. The proposed LD outperforms other treatments as it only requires less state space, which reflects low computation cost, and proves its capability to overcome the limitation of analysis.

Corresponding Author:

Chanintorn Jittawiriyanukoon,
Assumption University,
Samut Prakan, 10540, Thailand.
Email: pct2526@yahoo.com

1. INTRODUCTION

Currently, regression-based prediction (RBP) is astoundingly a booming tool; however, it can only be applicable to a limited range of experimental designs. The extendable regression-based ERP (Enterprise Resources Planning) introduced by [1] leverages any consolidations of nonlinear property, fractional confounding, and uninterrupted covariates. This study focuses on how nonlinear effects and correlations can be held within their proposed frameworks; filtering techniques and numerical analyses are examined as well. Prediction systems in the market survey [2] have turned out to be trendy in terms of the capability to classify what kinds of products/services clients are interested in. This study describes different strategies to estimate the users' search queries and self-generated content. As an overwhelming amount of information is escalating rapidly, analyzing dataset efficiently is more tedious for data scientists. In some situations, the system experiences missingness while cleansing irresistible datasets. Filtering the missing data is the most successful mechanism and is commercially opted nowadays. The crucial design behind this impressive mechanism is to educate users regarding numerous mistakes of similar users'. For example, a significant feature of filtering is that the user may not need to worry to manage their datasets much since the filtering will automatically suggest missing data based upon the "rules". In the course of suggestion, the system may agree to users'

conclusions (either implicitly or explicitly) about what they are fixing, which will help to update the rule of substitution. Steadily, the system will be more likely to extend recommendations, proving the robustness of the system. Usually, filtering will use some basic steps in prediction: (1) deletion of missingness is chosen as a result of the overall performance and; (2), an average value of neighbors is employed to predict the missing value. In the long run, practical filtering reveals that it is more helpful when compared with real content-based values. In numerical analysis, listwise deletion (LD) is a technique for leveraging missing data; hence, LD will eliminate a whole record from datasets if the value is missing. LD will induce statistics of the associated experiments [3]. The statistical function may reckon on the sample size when LD ignores records with missing values; this directly affects the sample, which is being experimentally analyzed. LD is ambiguous as the motivation for missingness may be systematic (i.e., questions in the survey leading to mine insightful information, such as current earnings, age, and an annual tax). Not to mention, subjective data will be removed from datasets, leading to unfairness in data analysis. For example, a survey may incorporate questions about respondents' criminal records, sexual seduction, and/or current incomes. Many of these questions in the questionnaire may not be responded to properly as they are considered to be disturbing questions. Multiple imputations are another technique for handling missing data to cope with this bias. Although LD seems to cause no problems, it is superior to compare LD with other mechanisms, such as weighted average substitution, single imputations, and maximum likelihood for manipulating missing data.

Other prediction approaches that employ the average substitution technique will be taken into account and these approaches will be discussed elaborately in section II. To overcome these issues, an RBP using LD is proposed. Based on these experiments, the proposed deletion gives higher accuracy without compromising in the performance, which obtains not only a lower average percentage of error, but also an experience with regards to the simple filtering approach. This paper introduces performance evaluation of the LD technique for treating the missingness in RBP. In our approach, different treatments to overcome missing data are discussed in section III. Furthermore, the statistical analysis using the Massive Online Analysis (MOA) simulation is executed, and the prediction results will be compared with the real values to know the estimation accuracy. The analysis will be demonstrated in section IV.

The contributions of our study are presented as follows:

a. Impaired datasets with different parameters, such as noisy data, missing data, attributes, instances, sizes, and value, are investigated.
b. Two steps of a treatment method for the given datasets are proposed, i.e., first, all the noisy data, which block the analysis, are removed; then, the missingness is substituted by an estimated value.
c. After the given dataset is deemed to be executable, the MRBP is applied and the model includes the historical events and forecasts associated with futuristic trends.
d. Statistical data, such as mean absolute error (MAE), root mean squared error (RMSE), correlation coefficient (COEF), mean squared error (MSE) and the average error percentage are calculated by the MOA software package, and the analytical results are discussed.
e. The real values are utilized to verify the outputs of each MRBP model after the treatment.

2. RELATED STUDIES

Predictive modeling is a step toward developing, measuring and verifying an outcome to fit an appropriate prediction. A modeling process ranging from artificial intelligence, statistics, and machine learning is always available in prediction analysis. An accurate RBP for app downloads which can help developers optimize some factors that influence apps can be found in [4]. The model is considered based on measurement, verification, and evaluation, employing the detection technique to estimate the effects from input data. The individual model exhibits its own pros and cons and is best fit for specific types of target complications. A discussion of problems in the Internet of Things (IoT) application [5-9] using regression analysis has been presented in [10]. In addition, each model can be recyclable and is built by a training set of algorithms using classical data. Simple regression is a statistical implication in which the decision is made as a potential prediction. For example, data that may consist of a disaster frequency is "predicted" by the length of the winter in each year. Many predictive models encompass several predictors and these transport to involve the correlation coefficient which is a correlation between paired predictors. In data collected on a quarterly basis, such as crime rate, and retail records, different weekly outcomes will be detected. Moreover, there may be swings in an association of the gathered data, such as, natural disasters and/or insurgency. To model the data, some "dummy" variables or substitutions can be taken into account to cover these consequences. Up to this limit, all datasets used in this paper will target to attributes and instances where conditional values of the prediction are assumed to be a regressive based model and exhibit a normal distribution. As a matter of fact, this is not a practical assumption, especially in the case of a bivariate example. The simple regressive model is rather of the form $Y= aX+ b + c$, where a and b represent the slope

of a straight line and the threshold value, respectively, while c is a compensated error for the sake of practical cases in which the points (X,Y) may not align with the straight line. It is assumed that the error c follows a normal distribution with zero mean and will be approximated by the figure of a and b. A simulation for the mobile application to rank algorithms in the app store has been investigated in [11]. The paper differs from [4], wherein a multiple regression-based prediction (MRBP) is opted to outline a dependent variable and three related attributes (independent variables). The multiple regression-based predictive model is a process of creating a continuous random dependent variable (also called the response variable), Y, and a number of independent variables, X1, X2, . . . , Xn. This model is also used to anticipate the value of the dependent variable using an estimated regression function of independent variables (also called the predictor or regressor variables). These independent variables are known as the trend of prediction, the model is of the form $Y = a1X1 + a2X2 + \cdots + anXn + b$, where b, the error variable, follows a normal distribution with zero mean, while a1, a2, a3, …., an are the coefficients. To find neutral predictions from sample parameters, i.e., if a multiple regression-based prediction is applicable, then the model must decide whether the variable Xn has a nontrivial, distinctive influence on the variable Y, after adjusting for the error variable b. The first step is to achieve an approximation of the coefficient an for an individual Xn estimating Y without the bias. The next step is to determine an implication of the prediction accuracy, i.e., to compute the MAE, RMSE [12], the average error percentage, and the interval of confidence about the prediction. In this investigation, methods for handling missing data in population parameters will be discussed. In performance evaluation, the degree of unfairness in parameter substitutions is presented with regards to whether or not there is a decent strategy for replacing missingness. This is also used to clarify that these methods for handling missing data are originally aimed to fix or efficiently handle the missingness [13]. However, these mechanisms described in the subsequent sections are employed to look over so that further analysis can follow.

3. PROPOSED TREATMENT METHOD AND PREDICTION

Missing values in datasets can be represented as the values of data, which are mistakenly collected for either attributes or instances in the population sample. The complication of missing values is typical in many data science research and can influence a quality of results after data manipulation [14]. For the same reason, several medical research studies have to concentrate on managing these missing values as presented in [15]. Lately, many studies have been determined based upon the assumption of comprehensive datasets. Missing values produce numerous complications, (1) the missingness reflects the analytical result, which may denote the probabilistic test and result to discard the null hypothesis; (2) the absence of data will ground unfairness in the prediction of population parameters; (3) the missing values will decrease the originality of the parameters; and (4) the missing values will be problematic for the conclusive analytical part of the research. These mentioned misrepresentations pressurize researchers in terms of the validity of the tests and result in misuse. In this research study, types of missing values will be classified based upon their preliminary causes why these values are missing. Thus, two types of missing values are categorized. Unplanned Missing (UM) is stated as whenever the probability of missing values is not associated with other any particular values. UM is a classical model, but an irrational notion for data science research studies. Conversely, if the value is missing by a technical failure or as a result of sampling techniques then these values become null in transit or technically substandard, such values are in the focus of UM as well. Thus, the approximated parameters are faired by the absence value. The other is Intentional Missing (IM) in which the missing characters or data are planned and purposely left as blank. The data of IM cases are ambiguous, and a fair approximation only will help to solve these problems. To approximate the missing values that are much more disoriented, they have to be modeled. In practice, the potential method of reckoning the missingness is to avoid the problem by capturing the data cautiously as well as empowering the questionnaire. To decrease the missing value in data science, these treatment methods will be proposed as follows:

Proposed treatment methods for asserting data have been developed to overcome the problem of noisy and missing values, such as LD, averaging variable (AV), and single imputation (SI). Let S be a given dataset matrix which comprises a rows and b columns (Sk1, Sk2, Sk3,…, Sk(b-1), Skb) for each k = 1, 2, 3,…, a. The S matrix is assumed to be a finite set. An element Skb is found to be a missing or noisy element whenever $\{Sij = \phi \parallel \infty, 1 \leq i \leq a; 1 \leq j \leq b\}$. The dataset with missing or noisy elements is called impaired dataset. Then, treatment methods to overcome the lack of executability as well as to advance the analysis by using the estimated vector Dn are presented in the subsequent section.

3.1. LD

LD deals with the missing values by removing them entirely in order, so that the data scientists can profit the manipulation as usual (only analyze the existing data). This is a commonly used method and is

recommended when the missingness is the UM case, and the analytical results are tolerant. Besides, the average percentage of errors from this deletion method is expressive as they will match the inconsistency about the approximation even if the approximation case was unfair. The point is that; the UM case is an uncommon incident in practice and that is why researchers may claim that this case will lead to an unfair prediction of the population parameters. This deletion has been recognized to develop a fair prediction and traditional analysis provided that it involves a large population, whrerin power is irrelevant, and therefore, LD is a realistic method. However, the research presented in this paper will involve a large sample, and if the assumption of UM is fulfilled, then this deletion is considered to be an acceptable strategy.

LD handles noisy and missing values by entire elimination so that the estimated dataset Dn can be computed and analyzed. This is a commonly used method and is endorsed as the missingness in terms of the unplanned case (UM). LD recalls the humble treatment mechanism whether the UM affects the removal. The zth row of matrix S contains an element Sij with noise and missingness, where $\{Sij = \phi \parallel \infty, 1 \leq i \leq a; 1 \leq j \leq b\}$, then the row is voided. The Dn dataset is now $\{Si \neq \phi \parallel \infty, 1 \leq i \leq (a\text{-}z); 1 \leq j \leq b\}$. The LD treatment is recognized as a fair estimation and traditional analysis if and only if the matrix S is large, and if the power is insignificant, then, LD is a practical method. Note that, the study employs a large dataset and the assumption of UM is satisfied. Thus, LD is a tolerable treatment method. Apparently, the state space of LD is reducible to [a-z, b] and the computation cost is O(ab).

3.2. AVs

In an AV, the average value [16] of all the parameters is calculated in place of the missing value. This method will allow the researchers to use the existing data in a missing dataset. The rationale of the AV is that the average value is an acceptable prediction for a random parameter out of a normal distribution. In case of missing values that are not randomly collected parameters, this substitution method will introduce an unpredictable bias. Moreover, this method does not develop distinguishing information but rather grows the size of a population and induces an underestimate value. Accordingly, this strategy is not generally perfect but has more statistical parameters for a scale score as such.

AV uses an average value to substitute the missingness. Dividing the given S dataset into two parts as follows: i) the first part is a dataset, which contains elements with noisy data (N). ii) the second part holds a dataset, which contains missing data (M). The first group is a corrupted element, which is not executable. The nth rows of matrix S contain an element Sij with noisy data (N), where $\{Sij = \infty, 1 \leq i \leq a; 1 \leq j \leq b\}$, then the entire row is discarded. The 2nd part is $\{Sij \neq \infty, 1 \leq i \leq (a\text{-}n); 1 \leq j \leq b\}$. The estimation for the Dn dataset with data imputation for missing values (M) is specified as follows:

$$S_{ij} = \frac{1}{|a-n|} \sum_{x=1}^{a-n} S_{xj} \quad (1)$$

The validation of the average value is that this value is an acceptable prediction for a parameter out of a normal distribution. In case of planned missing values, this treatment method induces a bias. This method does not only improve distinctive information, but also develops the size of population (state space) compared to LD (z = m + n; where z > n), and results an underestimated value. Furthermore, although this treatment is inadequate, it helps develop more parameters for a scale score. State space of AV can be reducible to [a-n, b] and computation cost is O(ab) + O(ab-bn).

3.3. SI

SI [17] is the step of substituting the missing data with any approximated values. Instead of removing values, similar to the LD case, this method rather keeps all the missing values by changing them with a value approximated by existing variables. Finally, the missing values will be substituted before the dataset is further manipulated. In SI, the existing information is employed to cope with estimation, and the estimated value is used to substitute as if the value was available. There are benefits to this mechanism as the SI can preserve much data compared to the LD and this mechanism can bypass considerably shifting either the variance or the distribution curve. Actually, this is a positive way to impute values particularly for the case of the greater correlation coefficient between the regressor variables and Y, since available information from data collection will be employed to approximate the missing values. However, SI may not be a good imputation all in all per se because when Y is depicted, there must be a difference between the regression curve and recognized values. Positively, with this method, the imputed values will mostly align the regression line.

SI uses single imputation, and maximum likelihood at random for substitution. Similar to AV, the S dataset must be divided into 2 parts. Now it labels the first part as garbled data. The nth rows of matrix S with

an element of Sij, noisy data (N) where {Sij = ∞, 1 ≤ i ≤ a; 1 ≤ j ≤ b} are removed. The 2nd part of the dataset is {Sij ≠ ∞, 1 ≤ i ≤ (a-n); 1 ≤ j ≤ b}.

The minimum likelihood of attribute (column) j (where j = 1, 2, 3,…, b) is characterized by S(min)j where S(min)j = Min (Skj) for each k = 1, 2, 3,…, (a-n). Similarly, the maximum likelihood of attribute j (where j = 1, 2, 3,…, b) is represented by S(max)j where S(max)j = Max (Skj) for each k = 1, 2, 3,…, (a-n).The substitution for the estimated Dn dataset with SI for missing values in each attribute j is randomly defined as follows:

$$S_{ij} = RAND\left[S(min)_j, S(max)_j\right] \quad (2)$$

State space of SI can be reducible to [a-n, b] but computation cost is O(ab) + O(2(ab-bn)).

In summary, regarding the performance, which is the computation cost metric, the existing estimations are problematic as the computation cost listed in Table 1 is very high.

Table 1. Disadvantages of Existing Estimations.

Estimation	Computation Cost
AV [11]	O(*ab*) + O(*ab-bn*)
SI [12]	O(*ab*) + O(2(*ab-bn*))
Proposed	O(*ab*)

3.4. Tentative Impaired Datasets

Taking out elements that are missing is an absolute idea in data curation for the reason that missingness will always impede and distort the analytics. Estimating for missing data will underline on missingness removal, which is the process of decreasing data errors generated by a poor population sample, but data elements that are missing can disturb the analysis. Missingness may induce unfair variables (negative results), directing data scientists to claim that an association of any attributes appears (which is fault analysis) while it does not (Type I error). In a case where this is used to conclude perfect analytics, these elements must be well handled. Due to the variation of missing data, this research will employ three ways to leverage this variation. Three methods comprising LD, AV and SI will be used to exploit and recover data analytics with regards to the amount of missingness. Experiments based upon three methods are reserved for MRBPs with ten datasets. These investigations are approximated in the sense of their outcomes on data analytics as the predictions from multiple regression-based analysis will be compared to actual data subsequently.

4. PERFORMANCE ANALYSIS

The commonly used and open-source software package called MOA (Release 2017.06) [18] will be manipulated for the analytics. Ten datasets have been opted and the performance evaluation of a multiple regression-based model for missing data has been resolved. The manipulation has been run on a Asus Windows 7 with Intel® Core ™ i5 CPU, 3.20 GHz Processor, and 2 GB RAM. The ten datasets have been explored in the sense that they all differ in the impairment, instances, sizes, and attributes. Note that some sample parameters, such as education, age, and sex will be intentionally omitted for the sake of maintaining the individual privacy of the sample's correspondents, but somehow it includes an adequate amount of attributes, which confirm the significance of the population sample.

4.1. MAE

MAE is used to quantify the estimated observations. MAE described in [19] is a favorable selection to calculate the estimating method; particularly, for cases in which metrics are based upon geometric distribution, for example, the geometric mean absolute error. The MAE is an average of the absolute value of failures and will be determined by the following (3), wherein xk represents the surveillance time series while $\hat{x}_k$ denotes the predicted time series.

$$MAE = \frac{1}{n}\sum_{k=1}^{n}|x_k - \hat{x}_k| \quad (3)$$

4.2. RMSE

RMSE is a figure used to quantify the variances between population values and sample sizes. RMSE can be estimated by a model of real observations. A method using RMSE as presented in [20] has been

proposed to improve the matching accuracy in case of an unstable scale. This high accuracy is a key function to analyze many image processing experiments. The RMSE designates the deviation of the difference between observations and predictions. These differences can be calculated by the sample population overestimation errors sometimes called an out-of-sample figure. The RMSE [21] of estimated values $\hat{x}_t$ for times t of a regression variable x_t can be computed for n different estimations as listed in (4). Training a dataset will minimize the error rate for the experiment set. The failure rate for the training dataset can be somewhat greater than if they are included in the experiment set. If any algorithms produce the equal MAE then, RMSE [22] will be considered for deciding the superior algorithm. In general, experimental set exhibits the minimum failure rate than the trained dataset.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{x}_t - x_t)^2} \quad (4)$$

4.3. Results and analysis

Approaches for managing the missingness can be classified into two common methods: a data-based approach, and a model-based approach. A data-based method will designate the missingness before carrying out the approximation, while the model-based method revises the algorithms to manage the missing data before the approximation of parameters is taken into account. The commonly used model-based approaches can be found in the SPSS software, which employs multiple imputations for manipulating missing data. Data-based approaches adopt Full Information Maximum Likelihood into consideration. Conversely, with the equation modeling algorithm, this feature is a bit complicated. As the algorithm has been designated to reflect parameter approximations specifically to the circumstances, this model can refer to a model-based method. If the algorithm leads to general results, such as a vector of mean values or covariance, then the model refers to a data-based approach instead. The error analysis method with ten different datasets using MOA will be considered. This is somehow a simple investigation of the targeted datasets, and the results are summarized in Table 2. The three errors in the table represent the COEF, MAE [21], and RMSE, respectively. Dataset 5 shows the smallest figures for both MAE [23-25] and RMSE while dataset 3 obtains the lowest COEF figures. The MRBP models after running an MOA simulation are outlined in Table 3.

Table 2. Prediction with RMSE for Ten Different Datasets

Dataset	COEF	MAE	RMSE
1	0.17	1.2	1.4
2	0.2	19.2	25.1
3	0.14	24.03	27.7
4	0.07	24.8	29.1
5	0.39	0.1	0.1
6	0.76	2.42	3.3
7	0.16	35.73	47.03
8	0.28	102.9	125.2
9	0.17	4.67	6.16
10	0.2	14.4	20.5

Table 3.MRBP Models for Ten Different Datasets

		Regression-based Prediction
DATASET	1	$X_4=0.09X_3+1.7$
	2	$X_8= 0.05X_5+0.05X_6+270.26$
	3	$X_4=0.31X_1+0.25X_2+ 108.76$
	4	$X_5=0.12X_1+ 40.5$
	5	$X_5=0.1X_4+0.48$
	6	$X_7= 0.06X_3+0.32X_4+0.14X_5+0.21X_6- 16.75$
	7	$X_5=-8.5X_1+0.03X_3+ 171.97$
	8	$X_7= 0.06X_5+0.04X_6+247.26$
	9	$X_5=-3.2X_1+48.7X_3 +11X_4 + 2.97$
	10	$X_7= -1.4X_1+1.3X_3-0.4X_4-18.52$

The diagnostic plots will be used to examine the precision of regressive-based predictions and their average error percentage. Figures 1-10 demonstrate graphs for an RBP prediction compared to the actual data. In this experiment, ten different datasets are used and the plots from these individual datasets are shown in Figures 1-10, wherein each figure corresponds to each dataset. The plot displays the prediction results

using the regression equation tabularized in Table 3 against the actual values. An LD, an AV and an SI curve have been plotted to highlight the prediction accuracy. The plots illustrate a roughly equal precision among three missing data estimations. In addition, for further details, the average error percentage is investigated to deal with best-fit estimation. Considering Table 3, the simple LD does seem to outclass other methods as well as is inexpensive in terms of the filtering process of the missingness. From the prediction results, all methods perform well with datasets 6, 7, 9, and 10. However, with datasets 1-5, and 8, the trends oppose to the actual data. The reason for this phenomenon is that the missingness and noise are highly correlated. For such cases, all estimations are not well aligned, but only display the direction of the curve. Thus, if considering these curves, those with opposite trends will reflect a small number of average errors since the average value of all curves listed in Table 4 is close to the identical direction.

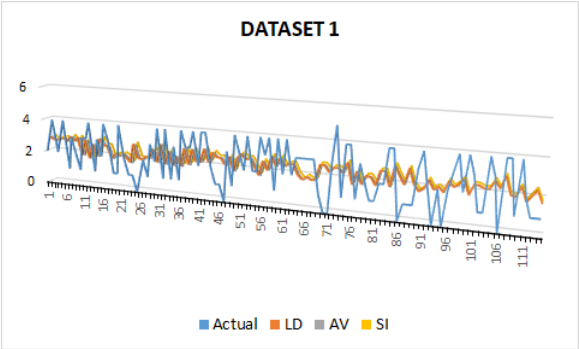

Figure 1. Comparison of dataset 1 with the actual data

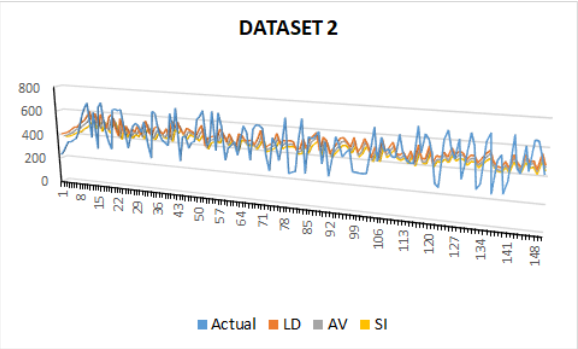

Figure 2. Comparison of dataset 2 with the actual data

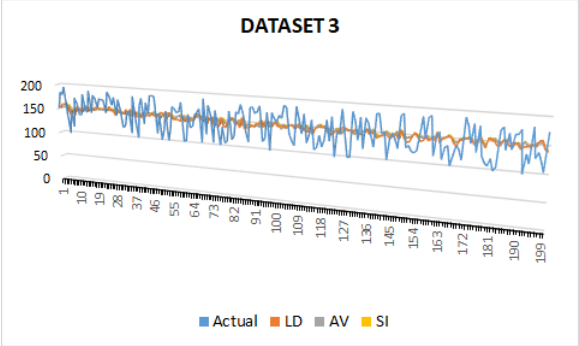

Figure 3. Comparison of dataset 3 with the actual data

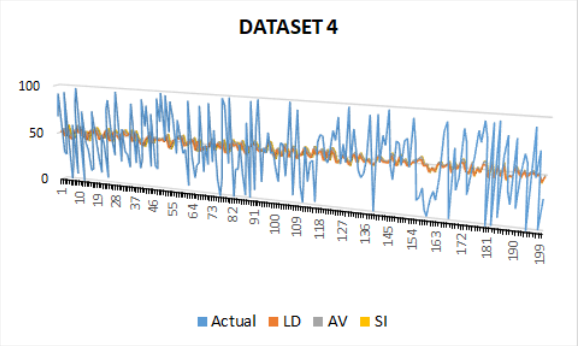

Figure 4. Comparison of dataset 4 with the actual data

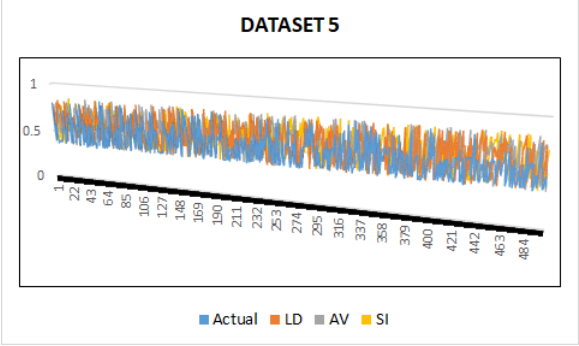

Figure 5. Comparison of dataset 5 with the actual data

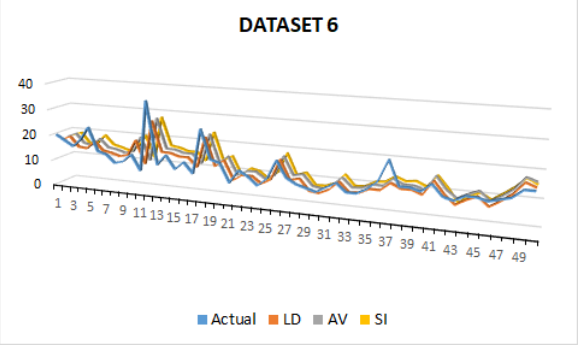

Figure 6. Comparison of dataset 6 with the actual data

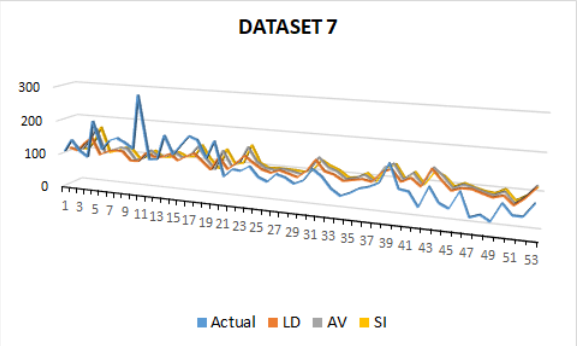

Figure 7. Comparison of dataset 7 with the actual data

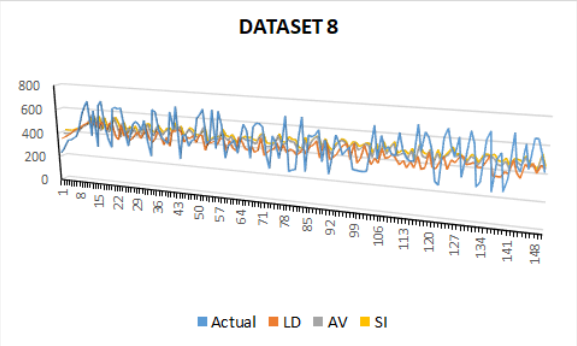

Figure 8. Comparison of dataset 8 with the actual data

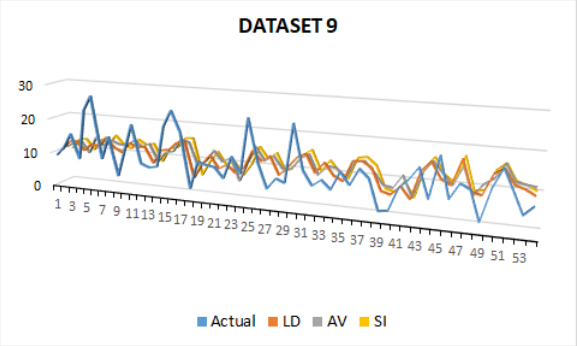

Figure 9. Comparison of dataset 9 with the actual data

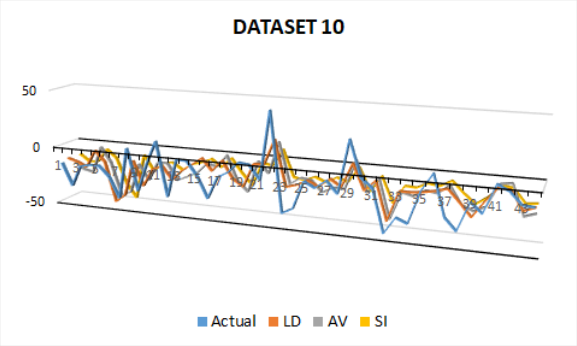

Figure 10. Comparison of dataset 10 with the actual data

Table 4. Average Error Percentage for Ten Different Datasets

Dataset	LD	AV	SI
1	58.4	57	58.47
2	24.5	23.1	23
3	16.7	17	16.9
4	23.8	23.7	23.4
5	2.52	14.6	58.6
6	14	12.75	13.8
7	33.9	34	32.5
8	23.5	23.1	23.9
9	43.6	47.6	40.9
10	20	26.2	24.3
Dataset	COEF	MAE	RMSE
1	0.17	1.2	1.4
2	0.2	19.2	25.1
3	0.14	24.03	27.7
4	0.07	24.8	29.1
5	0.39	0.1	0.1
6	0.76	2.42	3.3
7	0.16	35.73	47.03
8	0.28	102.9	125.2
9	0.17	4.67	6.16
10	0.2	14.4	20.5

5. CONCLUSIONS AND FUTURE WORK

In this paper, multiple regressive-based predictions are introduced as one of the critical diagnostic tools of advanced analytical systems. The classical methods of prediction have encountered practical missing data in time-series models, based on mistaken data collection for instances or attributes in the dataset. The problem of missing values is common in various research studies and can stimulate a quality of research

results after data curating. However, the effect of missingness has been widely taken into consideration in the case of recovery so that advanced analysis can be performed eventually. This research article determines to investigate the impact of missing data estimation as well as the precision of overcoming it. To attain this, ten datasets were chosen to emphasize the plots of their regressive- based models versus the actual values. Then, the percentage errors from the LD, AVs, and Sis were reviewed to obtain a best-fit filtering mechanism. The results conclude that the LD possesses a humble and simple filtering technique whether or not the missing values of an input influence the future neglected values. Additionally, LD can be utilized for predictions and is appropriate as well as better than the other AV or SI methods for prediction. Cross-correlation and covariance analysis of the population sample will be the focus of future research.

REFERENCES

[1] V. Kanchana, and S. Sri Ranjini, "Investigation and Study of Vital Factors in Selection, Implementation and Satisfaction of ERP in Small and Medium Scale Industries," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 2, pp. 1150-1155, 2018.

[2] S. Kashi, H. R. Karimi, K. Thoben, M. Lutjen, and M. Teucke, "A Survey on Retail Sales Forecasting and Prediction in Fashion Markets," *Journal of Systems Science and Control Engineering*, vol. 3, no. 1, pp. 154-161, 2015.

[3] A. Olinsky, S. Chen, and L. Harlow, "The Comparative Efficacy of Imputations Methods for Missing Data in Structural Equation Modeling," *European Journal of Operational Research*, vol. 151, no. 1, pp. 53-79, 2003.

[4] S. Wang, W. Wu, and X. Zhou, "App Store Analysis: Using Regression Model for App Downloads Prediction," *Social Computing, ICYCSEE 2016, Communications in Computer and Information Science*, Springer, Singapore, vol. 623, 2016.

[5] H. B. Chi, M. F. N. Tajuddin, N. H. Ghazali, A. Azmi, and M. U. Maaz, "Internet of Things (IoT) Based I-V Curve Tracer for Photovoltaic Monitoring Systems," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 13, no. 3, pp. 1022-1030, 2019.

[6] H. F. Hawari, A. A. Zainal, and M. R. Ahmad, "Development of Real Time Internet of Things (IoT) Based Air Quality Monitoring System," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 13, no. 3, pp. 1039-1047, 2019.

[7] M.A.F. Ismail, M. N. Md. Isa, S. N. Mohyar, M.I. Ahmad, M. N. M. Ismail, R. C. Ismail, A. Harun, and S.A.Z. Murad, "e-PADI: An IoT-Based Paddy Productivity Monitoring and Advisory System," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 14, no. 2, pp. 852-858, 2019.

[8] N. A. Khairi, A. B. Jambek and R. C. Ismail, "Performance Evaluation of Arithmetic Coding Data Compression for Internet of Things Applications," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 13, no. 3, pp. 591-597, 2019.

[9] I. Kang, H. Song, and H. Jung, "User Command Acquisition Based IoT Automatic Control System," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 13, no. 1, pp. 307-312, 2019.

[10] A. Rghioui, and A. Oumnad, "Internet of Things: Surveys for Measuring Human Activities from Everywhere," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 5, pp. 2472-2482, 2017.

[11] S. L. Lim, and P. J. Bentley, "Investigating App Store Ranking Algorithms Using A Simulation of Mobile App Ecosystems," *IEEE Congress on Evolutionary Computation (CEC)*, pp. 2672-2679, 2013.

[12] C. Jittawiriyanukoon, "Evaluation of a Multiple Regression Model for Noisy and Missing Data," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 4, pp. 2220-2229, 2018.

[13] H. Kang, "The Prevention and Handling of the Missing Data," *Korean Journal of Anesthesiology*, vol. 64, no. 5, pp. 402-406, 2013.

[14] R. J. Little, *et al*., "The Prevention and Treatment of Missing Data in Clinical Trials," *The New England Journal of Medicine*, vol. 356, no. 14, pp. 1355-1360, 2012.

[15] R. T. O'Neill, and R. Temple, "The Prevention and Treatment of Missing Data in Clinical Trials: An FDA Perspective on the Importance of Dealing With it," *American Society for Clinical Pharmacology and Therapeutics*, vol. 91, no. 3, pp. 550-554, 2012.

[16] C.J. Simon-Gabriel, A. Sciber, I. Tolstikhin, and B. Scholkopf, "Consistent Kernel Mean Estimation for Functions of Random Variables," *The 30th Conference on Neural Information Processing Systems (NIPS 2016)*, pp. 1-17, 2016.

[17] N. Masseran, A. M. Razali, K. Ibrahim, A. Zaharimand, and K. Sopian, "Application of the Single Imputation Method to Estimate Missing Wind Speed Data in Malaysia," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 6, no. 10, pp. 1780-1784, 2013.

[18] A. Bifet, R. Kirkby, G. Holmes, and B. Pfahringer, "MOA: Massive Online Analysis," *Journal of Machine Learning Research*, vol. 11, pp. 1601-1604, 2010.

[19] C. Chen, J. Twycross, and J. M. Garibaldi, "A New Accuracy Measure Based on Bounded Relative Error for Time Series Forecasting," *PLoS ONE*, vol. 12, no. 3, pp. 1-23, 2017.

[20] Z. Tang, P. Monasse, and J. M. Morel, "Improving the Matching Precision of SIFT," *IEEE International Conference on Image Processing (ICIP)*, pp. 5756-5760, 2014.

[21] S. Salisu, M. W. Mustafa, and M. Mustapha, "A Wavelet Based Solar Radiation Prediction in Nigeria Using Adaptive Neuro-Fuzzy Approach," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 12, no. 3, pp. 907-915, 2018.

[22] E. A. Abdullah, S. M. Zahari, S. S. R. Shariff, and M. A. A. Rahim, "Modelling Volatility of Kuala Lumpur Composite Index (KLCI) Using SV and Garch Models," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 13, no. 3, pp. 1087-1094, 2019.
[23] M. F. Talib, M. S. Anuar, and C. B. M. Rashidi, "Performance of Geometrical Effect in Wavelength Filtrate Detection Using 10GBPS Data Rate for Free Space Optical Communication System," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 13, no. 2, pp. 575-583, 2019.
[24] D. Kasiraja, and A. S. Vijendran, "Adaptive Data Structure Based Oversampling Algorithm for Ordinal Classification," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 12, no. 3, pp. 1063-1070, 2018.
[25] Y. C. Koo, and M. N. Mahyuddin, "An Enhanced Distributed Control-Theoretic Time Synchronization Protocol Using Sliding Mode Control for Wireless Sensor and Actuator Network," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 14, no. 2, pp. 688-696, 2019.