Evaluation of a Multiple Regression Model for Noisy and Missing Data

Chanintorn Jittawiriyanukoon
Assumption University, Thailand

Article Info

Article history:

Received Nov 13, 2017
Revised Jan 17, 2018
Accepted Apr 2, 2018

Keyword:

Big data
Classification
Noisy and missing data
Performance evaluation
Regression model
Root mean square error

ABSTRACT

The standard data collection problems may involve noiseless data while on the other hand large organizations commonly experience noisy and missing data, probably concerning data collected from individuals. As noisy and missing data will be significantly worrisome for occasions of the vast data collection then the investigation of different filtering techniques for big data environment would be remarkable. A multiple regression model where big data is employed for experimenting will be presented. Approximation for datasets with noisy and missing data is also proposed. The statistical root mean squared error (RMSE) associated with correlation coefficient (COEF) will be analyzed to prove the accuracy of estimators. Finally, results predicted by massive online analysis (MOA) will be compared to those real data collected from the following different time. These theoretical predictions with noisy and missing data estimation by simulation, revealing consistency with the real data are illustrated. Deletion mechanism (DEL) outperforms with the lowest average percentage of error.

Corresponding Author:

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

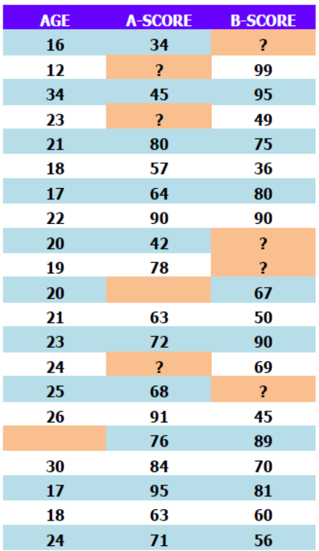
1. INTRODUCTION

During recent years, the concept of big data and the fluctuated applications of the Internet of Everything (IoE) per se have been chased with huge attention by the data scientists. Big data is a holistic, free formatted and time evolving of dynamic information but influence the quality of inhabitant life and to advance an environmental sustainable growth, digital economy and public in uninterrupted development. For instance, a smart and digitized city is crucial to accumulate vast data and to powerfully carry out decision-making on the upbringing. In order to accomplish so, it is to make use of data curation such as classification or clustering which will consent to straightforwardly re-organize an enormous data with the aspire of intelligently data analytics, evaluation, prediction and visualization; from unstructured data format through pre-processing, to the scrutiny of noisy data and missing data. For the improvement of the IoE and the smart city [1], wireless sensor [2] has been deployed in order to ensure real-time supervising of several devices that can enhance digital-age lifestyle, energy-saving and the quality time of community. This measureless amount of time series data generated by the sensor devices, simultaneously with the collected data from other digital devices, such as cell phones, monitoring software and social media, must be appropriately treasured and subsequently pre-processed in order to secure the insights. Scrutinizing gigantic amount of data is a big threat, thus big data curation [3] for gathering, saving and investigating datasets has to be associable to execute them properly and resourcefully.

In practice, missing data [4] arise as no value (blank or unexpected) has been found for any entries during the surveillance as shown in Figure 1. Missing data can commonly arise and have a considerable consequence on the final results which can be extracted from data entry. It is due to nonresponsive: blank information is filled for at least one or many places. Some private or sensitive accounts for example age, income, and etc are likely to produce a question than others. Missingness may arise as respondents give up before the questionnaire time terminates and acquired answers are missing. Data normally are incredibly missing in quantitative research in business, engineering and sciences since government officers opt not to, or feel reluctant to, fill critical/sensitive measurements. Occasionally those mistakes are motivated by researchers, for instance, improper data collection is conducted or human-errors are taken during data entry phase. These missingness thus simply create diverse categories, missing by chance or intentional missing. Consequently noisy data that detected amongthe navigation menu, advertising banner and other information content of the web document influences adversely theperformance of mobile applications that involves with thecontent as such [5]. In addition, Sharma and Bhatia [6] expanded a page replacement algorithmin order to separate noisy data from web document. Chae *et al*. [7] mention it is compulsory that big data analytics in supply chain management (SCM) be collective with SCM objectives to advance working performance and escalate the value of analytics. However, they have not been looking at practical issue of missingness.

Noisy data is described as worthless data. The term is called as an alternative expression for crooked data as depicted in Figure 2. Nonetheless, the meaning has included any incomprehensive data for instance unstructured format of data. Any unreadable data which has been detected by the machine will develop and can be defined as noise. Shabir and Padma [8] presented a denoise procedure to improve the quality of original image. Noisy data is worsening of data collection caused by external hazards. These noise include not only internal problems such as software or hardware incompatibility or viruses, system malfunction, failures, or flaws, but also environmental hazards such as dust, moist, extreme temperatures, black-outs, and water. Noisy data on the other hand redundantly requires the extraordinary amount of saving space and can unfavorably upset the outcomes of data analytics.

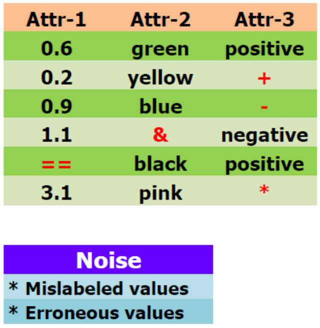
Thus analysis can overcome this problem by employing data collected previously (historical data) to filter out noisy data and ease data curation. Much of noisy data can affect failures in hardware processing and accuracy. Moreover, typos, slang, misspelling, careless and other abbreviations can obstruct machine learning. Corrupt data is a realistic trouble, induced either because of defective data sources or during data broadcast (traversing). Noise is able to seriously mess up machine learning process of collected data. It is a much more rigorous trouble in case of data streams as it connects with concept drift. If a greedy algorithm is concerned to concept drift, it may qualify noise by erroneously picturing it as data from a fresh concept. If it is to be too strict to noise, it may have to ignore drifts then fine-tune. Besides, the computational complexity of the K-means algorithm with datasets has been evaluated in [9].



AGE	A-SCORE	B-SCORE
16	34	?
12	?	99
34	45	95
23	?	49
21	80	75
18	57	36
17	64	80
22	90	90
20	42	?
19	78	?
20		67
21	63	50
23	72	90
24	?	69
25	68	?
26	91	45
	76	89
30	84	70
17	95	81
18	63	60
24	71	56

Figure 1. Example of missing data (bold orange)

The aim of this paper is to evaluate the accuracy of multiple regression analysis for noisy and missing data environment using MOA [10] simulation. Both noise and missingness will be experimentally taken into consideration for practical point of view. Firstly, the noisy data will be weeded out in order to avoid complication in processing. Secondly, an estimator will iron out missing data then multiple regression results from simulation are collected for validation of the estimation method. Lastly, prediction data based on multiple regression equation will be compared to following real data. The accuracy of noise filtering and non-filtering techniques then will be discussed after comparison.



Attr-1	Attr-2	Attr-3
0.6	green	positive
0.2	yellow	+
0.9	blue	-
1.1	&	negative
==	black	positive
3.1	pink	*

Noise
* Mislabeled values
* Erroneous values

Figure 2. Example of noisy data

2. RELATED WORK

In order to fresh the text document by removing the noisy data out has been discussed intensively using various techniques. Bar-Yossef *et al*. [11] applied an approach based upon document object model's tree of the web content. Another method which extracts noise in order to advance data mining outcomes has been proposed by [12]. They formerly employ a structural tree which is analogous to the tree structure of document object model. After that the noisy data will be discovered by an evaluation process of contents found in structural tree. However, construction of the preferable tree as such will be time-consumption. Debnath *et al*. [13] have recommended a technique comparable to method written in [12], on the other hand rather chosen data blocks, which are nontrivial but exceed a set threshold. Then individual data block will be projected. Not all collected data succeeds the same mistakenly defective pattern. A method which can pinpoint and remove pertinent content from web contexts has been introduced by [14]. The method specifies significance as texts which more independently interpretable than the image. But, the method facilitates only on web contexts. It should be an uncomplicated algorithm which can neutrally extract noisy data from the web. A noise reduction algorithm with three phases has been presented in [15].

The 1st phase as specified in this algorithm converts a web document to table format containing of *x* instances and y attributes, then the essential data can be added into the table if necessary. The 2nd and 3rd phase will solely weed out noisy data using filtering techniques. Other operational aggregated methods split data streams into fixed blocks and machine will learn an aggregation from individual blocks. As soon as a fundamental model is built, it will never amend new streams. In general there are two voting categories: uniform and weighted. These two types are not associated to our approach as our proposed method will construct an aggregation from these sequential divided blocks. But what is omitted from the above two approaches is an analytical tool handling noisy data. While there are some algorithms for noise recognition, so called anomaly detection, noise elimination crafts a considerable breach between data stream and the above mentioned approaches. Furthermore, the issue of acquiring a noisy data will be addressed.

Our study then is unlike two approaches as stated above, firstly, anomaly detection will be clenched up into the machine learning process for the reason that concept drift is a fraction of outlier detection. Secondly, the distance vector can be easily drawn by the classifier, rather from a formula specified by datasets per se. As a matter of fact, the anomaly detection and adaptive machine learning hence reciprocally support one another. In general, an accurate adaptive model nourishes to discover the anomalies. Alternatively, by properly finding and removing the anomalies at earlier stage, a further exact model can be executed. Adaptive learning model with reference to vigor and revision, correspondingly will be illustrated. Model mapping and calculation will be provided as well. Afterward investigational results will be listed.

3. NOISE AND MISSING DATASETS

In this section, characteristics of noise will be described, at the same time datasets which are inclusive of noise are outlined and the comprehensive discussion is given.

3.1. Self-generated noise

Our approach described in this paper will allow MOA simulation to read a noisy dataset. MOA tolerates comparison of single algorithm on datasets with dissimilar noise rates. The procedure is created by numerous ideas. A dataset reader initially will choose small fragments with specific value from dataset. Secondly, ample figures for the selected sample will be abandoned if it appears very doubtful. As each leaf in decision tree is designed, all figures from each attribute are measured as applicants for fragmenting. Through every round of calculation that does not decide to fragmentize, attributes are marked to be unfortunate if their values are less than the value from top attribute which is greater than the bound. Regarding to the bound, current attributes are improbable to be chosen in the decision tree, therefore the reference to this information is rejected from that calculation point onward. The two ideas are interconnected, repeating until the replaceable value is set. This approach can function as a foundation for calculating a range of predictable values for a noisy dataset.

MOA simulation can randomly augment noise to datasets. Perturbation amount is presented to statistical figures. A level of noise can be familiarized to the datasets after provoking. For distinct attributes, a perturbation probability governs the coincidence which any figures are changed to others but the original figure. In the case of statistical attributes, a level of random perturbation is reckoned to all figures, randomized from a Gaussian distribution function with the identical standard deviation as of the original figures timed by a probability of perturbation. For example, the algorithm may augment 10% perturbation to the decision tree dataset. It is aimed that experiment with noiseless and noisy data can contribute perception to how smart the algorithms can succeed perturbation.

To quantify noise in the dataset is now considered. Only the case in which a bound on the noise exists and the case where the noise is random. In the first one, optimization is guaranteed on the machine learned simulation. In the latter one, the machine learning will associate synthetic datasets and training set. To synthesize the targeted datasets with noise, a synthesizer for bit level will vigorously insert some false until the preferred datasets are achieved. In case of hefty datasets, a numerical synthesizer which can calculate probabilistic models for a dataset in repository then approximates the preferred dataset based upon the computation.

3.2. Experimental datasets

Eliminating substances which are noise is a vital objective of data curation (both filtering and outlier) since noise garbles and obstructs data analytics. Surviving filtering techniques emphasize on eliminating noise which is the creation of low-level data errors developed by a deficient data collection, but data substances which are inappropriate or dimly related can easily frustrate the analytics. For example, noise can direct to biased variables (negative consequence), resulting data scientist to consider that an association of any attributes occurs (fault conclusion) though in fact it may not be (a type one error). Hence, if it is to enrich data analytics as far as achievable, these substances must be detected as noise, regarding to the basic analytics. Therefore, it is essential for noise removal techniques in order to eliminate any types of noise. Due to big portions of noise during data collection process, any techniques have to abandon a fraction of data. This research investigates 3 types proposed for noise (N), missing data (MD) and the integration of both (NMD). Three proposed treatment techniques include Listwise Deltion (DEL), Single Assertion Mechanism (SAM), and Random Method (RAM) to improve data analytics in the occurrence of extraordinary noise levels. Three investigations and two techniques are based upon multiple regression model with five different datasets. These experiments are estimated in terms of their impact on the successive data analytics, explicitly, the following year data collection will be employed to compare to those estimations from multiple regression analysis.

4. SIMULATION RESULTS AND ANALYSIS

The open-source and acceptable tool MOA (Release 2017.06) [10] will be employed for the analysis. Two dissimilar datasets have been used and the performance evaluation of a multiple regression model for noisy and missing data has been figured out. The computation has been executed on an Acer Windows 8 with Intel® Core ™ i5 CPU, 1.60 GHz Processor and 8 GB RAM on board. The datasets have been selected in order that they are different in number of attributes, instances and size.

The public health parameters are composed of monthly salary (SAL), loss rate per thousand populations (LOS), number of medical doctors per hundred thousand population (DOC), number of hospitals per hundred thousand population (HOS) and population density per square kilometers (DEN). Public health dataset was provincially collected by information technologist who had been working closely in terms of public health policy and promotion (PHPP). Collected data are subjected to help manage and improve a Quality of Life (QoL) in provincial areas. Note that these data was partially aggregated by temporarily hired staffs who have lower skill in computer literacy. Due to this reason it led to noisy and missing data in the original dataset.

The first dataset contains 580 samples while the second dataset collects 198 samples. Its attributes correspond to the public health characteristics as depicted in Table 1, some parameters such as sex, age, and others which are neglected due the confidentiality of the sample's owner, however eventually it consists of investigative attributes which serve the essential label of the sample.

Table 1. Public Health Dataset Attributes

Attribute	Type	Role
Gender	Nominal	Regular
Sal	Integer	Regular
Los	Integer	Regular
Doc	Integer	Regular
Hos	Integer	Regular
Den	Integer	Regular
Comments	Nominal	Regular

The soccer parameters are constructed by player height (HEI), weight (WEI), number of successful goals after thousand attempts (GOL), number of passing the ball (PAS) and average goals out of ten national competitions (AVG). Soccer dataset was captured by a media company which had been involving with television broadcasting. Data is meant to help analyze player's performance in general. Note that these data was manually collected by company staffs who unintentionally developed noise in the original dataset as well. Its attributes exhibit the player characteristics as listed in Table 2, some parameters such as education background, income, and others which are omitted due the secrecy of the player's information, but it contains experimental attributes which remark the significant label of the information.

Table 2. Soccer player Dataset Attributes

Attribute	Type	Role
Name and Surname	Nominal	Regular
Team	Nominal	Regular
Hei	Integer	Regular
Wei	Integer	Regular
Gol	Integer	Regular
Pas	Integer	Regular
Avg	Integer	Regular
Comments	Nominal	Regular

4.1. Mean absolute error

The mean absolute error (MAE) is an amount used to quantity estimates of the ultimate results. The MAE is a mean of the absolute value of flaws and can be computed by:

$$\mathrm{MAE} = \frac{1}{n}\sum_{k=1}^{n}|x_k - \hat{x}_k| \tag{1}$$

where x_k is the definite observation time series and $\hat{x}_k$ is the predicted or estimated time series.

4.2. Root mean squared error

Root mean squared error (RMSE) is a quantity used to measure the differences between sample and population values forecasted by a model or estimated values of actual observations. The RMSE denotes the standard deviation of the difference between forecasts and observations. These differences are computed by the sample data performance over prediction errors as calculated out-of-sample.

The RMSE of forecasted values $\hat{x}_t$ for times *t* of a regression's dependent variable $x_t$is calculated for *n* different forecasts as shown in Equation (2).

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

Training a dataset in general will decrease the error rate for experiment set. Flaw rate for training dataset is relatively higher than that of the experiment set. If any two algorithms result the identical MAE then RMSE is taken into account for selecting the finest algorithm. Normally experimenting set has low flaw rate than the training dataset.

4.3. Noise and missing patterns

Fully at Random (FaR) means noise or missing patterns are not depending on any factors. For instance many questionnaires will ask for a random sample. Intentionally (INT) means noise or missing patterns reckon on confidentialities. For example, some respondents may awkwardly to report their annual income, age, personally sensitive data etc. Eventually they may fill up with blank intentionally or untrue figures. Most missing data patterns conclude either FaR or INT.

4.4. Madhu treatment method

Madhu and Nagachandrika [16] presented treatment method for data imputation based upon dual distance vectors which are employed to outline a representation between the nearest neighbor and the cluster centroid. They denote dataset elements as a representative *p* x *q* matrix. Dataset matrix *D* characterizes the elements of *p* rows and *q* columns and each row contains a sequence of q-tuples of data elements such as (*dk*1, *dk*2, *dk*3,..., *dk*(*q*-1), *dkq*) for each *k* = 1, 2, 3,..., *p.* All datasets are assumed to be a set of finite elements. An element *dkq* is considered to be a missing element whenever $\{dij$ = null, $1 \leq i \leq p;\ 1 \leq j \leq q\}$. Then a k-means algorithm to structure clusters and to define the centroids using the vector *Vn* is specified as listed in Equation (3) below:

$$V_n = \frac{1}{|T|}\sum_{x=1}^{T} d_{kx} \tag{3}$$

Hence, the nearest neighbor based upon a Euclidian distance vector will be computed for missing values of a given dataset. Assume that *D* is a set of finite elements and both *m* and *n* correspond to *D*. *N* is considered to be the nearest neighbor of *M* if and only if *N* is the nearest to *M* among other points located in *{D – M}.*

4.5. Proposed treatment methods with less bias

Three proposed methods for imputing data to handle the problem of noisy and missing value which are based upon listwise deletion, assertion, and randomization have been presented. Let *D* denote dataset matrix which illustrates a representation of *p* rows and *q* columns matrix (*dk*1, *dk*2, *dk*3,..., *dk*(*q*-1), *dkq*) for each *k* = 1, 2, 3,..., *p.* The dataset is assumed to be a finite set. An element *dkq* is a missing or noisy element (NMD) whenever $\{dij$ = null || noise, $1 \leq i \leq p;\ 1 \leq j \leq q\}$. The dataset with NMD elements is called unexecutable dataset. Then treatment methods to get over the unexecution and move the further analysis on using the estimated vector *En* are described in the following section.

4.5.1. Deletion mechanism (DEL)

Listwise deletion deals with the NMD values by removing them entirely in order that data scientist can analyze the estimated dataset. It is commonly used method and recommended when the missingness is Unplanned Missing (UM) case. DEL retains the humble and simple treatment technique whether or not the NMD of an input influences the future neglected values. Any *z* rows of matrix *D* possess an element *dij* with NMD where $\{dij$ = null || noise, $1 \leq i \leq p;\ 1 \leq j \leq q\}$ then the entire row is cancelled. The estimated *En* dataset is $\{dij \neq$ null || noise, $1 \leq i \leq (p\text{-}z);\ 1 \leq j \leq q\}$. The DEL treatment is known to cultivate a fair prediction and classical analytics if and only if dataset is large, where power is trivial then the listwise deletion is an interesting method. Note, the study in this paper handles a large sample and the assumption of UM is fulfilled then this deletion is deliberated to be an acceptable strategy.

4.5.2. Single assertion mechanism (SAM)

Employ dummy variable namely average value to impute data to substitute the missingness. Divide the given *D* dataset into 2 groups that is: a) 1st group is a dataset which contains elements with noisy data (N). b) 2nd group represents a dataset which contains missing data (MD). Now consider the first group as garbled dataset which is unexecutable. Any *z* rows of matrix *D* possess an element *dij* with noisy data (N) where $\{dij$ = noise, $1 \leq i \leq p;\ 1 \leq j \leq q\}$ then the entire row is cancelled. The second group dataset is $\{dij \neq$ noise, $1 \leq i \leq (p\text{-}z);\ 1 \leq j \leq q\}$. The substitution for estimated *En* dataset with data imputation for missing values is defined as follows:

$$d_{ij} = \frac{1}{|p-z|} \sum_{x=1}^{p-z} d_{xj} \tag{4}$$

The validation of the average value is that it is an acceptable prediction for a random parameter out of a normal distribution. In case of planned missing value, this treatment method will lead an unpredictable bias. Not only this method develops distinguish information but rather grows the size of population compared to DEL and encourages an underestimate values. Moreover, this technique is imperfect, but develops more parameters for a scale score rather.

4.5.3. Random method (RAM)

Use multiple assertions, maximum likelihood at random for replacement. Like SAM, the *D* dataset must be split into two groups. Now deal with the 1st group as corrupted dataset which cannot be unexecutable. Any z rows of matrix D with an element of *dij*, noisy data (N) where {*dij* = noise, 1 ≤ *i* ≤ *p;* 1 ≤ *j* ≤ *q*} are then removed. The second group dataset is {*dij* ≠ noise, 1 ≤ *i* ≤ *(p-z);* 1 ≤ *j* ≤ *q*}.

The minimum likelihood of attribute (column) *j* (where *j* = 1, 2, 3,..., *q*) is characterized by $d(\min)_j$ where $d(\min)_j$ = Min (*dkj*) for each *k* = 1, 2, 3,..., *(p-z)*. Similarly, the maximum likelihood of attribute *j* (where *j* = 1, 2, 3,..., *q*) is represented by $d(\max)_j$ where $d(\max)_j$ = Max (*dkj*) for each *k* = 1, 2, 3,..., *(p-z)*. The substitution for estimated *En* dataset with multiple imputations for missing values in each attribute *j* is randomly determined as follows:

$$d_{ij} = RND\left[d(min)_j, d(max)_j\right] \tag{5}$$

Clearly the proposed approach presents columnwise (attribute-orientation) operation by eliminating unexecutable noisy data then imputing a replacement data based upon less-bias-mechanisms as described above. However, Madhu's method is rowwise (instance oriented) operation which cannot be applied with realistic case of NMD for two reasons. One is ND will be an invalid figure in statistical calculation. The other is likelihood in each attribute is more significant than instance for predicting a future trend. A comparison with method explained in [16] is displayed in Table 3.

Table 3. Comparison with Existing Method

DATASET	Madhu Method	DEL	SAM	RAM
Health	N/A	✓	✓	✓
Soccer	N/A	✓	✓	✓

This research has been conducted to carry out an in-depth analysis of the error of estimation with less bias (DEL, SAM and RAM) comparable to two original datasets (public health and soccer players). The structures of these datasets are as listed in Table 1 and Table 2. Table 4 depicts the overall results for correlation coefficient (COEF), mean absolute error (MAE) and root mean squared error (RMSE). The RMSE values after applying deletion mechanism (DEL) for missing and noise data comparably differ and they are getting lower for public health dataset. When compare to other mechanisms, the MAE of DEL can also be found differently, this is the lowest value of 18.3. While, the results collected for when the soccer player dataset has been taken into account for the evaluation are fairly close.

Table 4. Estimation with Root Mean Square Error

DATASET	Prototype			DEL			SAM			RAM		
	COEF	MAE	RMSE	COEF	MAE	RMSE	COEF	MAE	RMSE	COEF	MAE	RMSE
Health	0.16	35.7	47.0	0.67	18.3	22.2	0.08	38.1	49.6	0.38	32.8	43.4
Soccer	0.17	4.67	6.16	0.23	4.58	5.98	0.22	4.56	6.01	0.24	4.51	5.85

Different patterns of unreally shaped missing observations had been functioned for two mentioned datasets. Once again, the humble deletion is intended to probe the sensitivity of the estimation to missing observations in each attribute and the effect of deletion on correlation coefficient. Finally, a systematic pattern of replacement has been exploited to compensate a deletion in case of missing data. This is selected to echo the case where, for instance, confidential income units are more hesitant to explode their income in the review. In most cases, two substitution patterns (SAM and RAM) are exercised in this research and MOA simulation runs per patterns are executed. The estimated multiple regression equations with their respective patterns for noise and missing data are simulated and results are summarized in Table 5.

Table 5. Summary of Multiple Regression Models

	Public Health
Prototype	X_5 = -8.56 X_1 +0.03 X_3 + 171.97
DEL	X_5 = -9.02X_1 +0.05X_3 + 140.97
SAM	X_5 = -7.82X_1-0.28X_2 +0.04X_3 + 193.84
RAM	X_5 = -8.98X_1 +0.42X_2 + 139.46
	X_1=Los, X_2=Doc, X_3=Hos, X_4=Sal and X_5=Den
	Soccer Player
Prototype	X_5 = -3.23X_1 +48.7X_3 +11.09X_4 + 2.97
DEL	X_5 = -4.60X_1 +52.48X_3 + 18.13
SAM	X_5 = -4.70X_1 +52.04X_3 + 19.34
RAM	X_5 = -4.04X_1 +49.83X_3 + 16.07
	X_1=Hei, X_2=Wei, X_3=Gol, X_4=Pas and X_5=Avg

The estimated regression equations can hence be paralleled with the real/authentic data in subsequent years in order to investigate the accuracy of each prediction. As explained in previous section, the multiple regression-based imputation forms a forecasting trend. Note that auto regressive model based imputation are presented in [17] to reckon missing values and the accuracy are evaluated by using RMSE metrics but it is regardless of multiple regression model. Not to mention they do not concern about the accuracy of the prediction trend at all. While an interesting data imputation for estimating the missing value is introduced in [16], however this new paradigm is not applicable for a prediction model as well. That is, it can be seen as the first reason that the imputed values reflect malfunction. This problem is represented in Figure 3 and Figure 4. The authentic data points diverge from the regression line by some extent but imputed values jam perfectly on the regression line. This problem is easily fixed simply by replacing random value to each imputed value (this relates to adding the improvement). The second reason there is slight variability shares to the fact that the regression equations employed in imputation are based upon a sample from total population. As illustrated in both Figure 3 and Figure 4, there should be additional compensation around the dotted linear line in order to lower the different values at maximum extent.

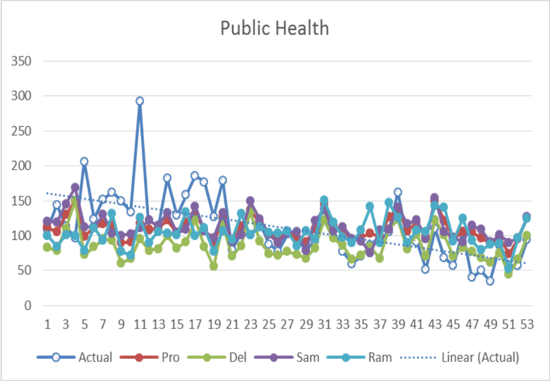


Figure 3. Comparison of public health data with authentic data

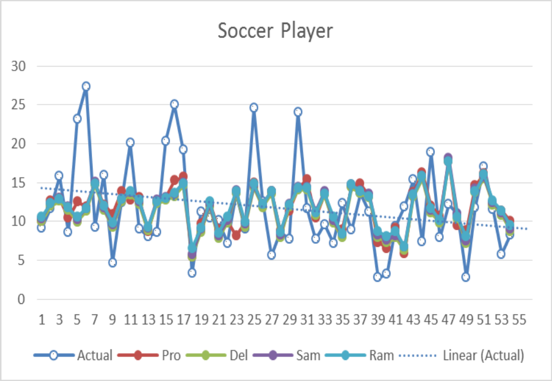


Figure 4. Comparison of soccer player data with authentic data

The estimation error (Err) of forecasted values y_t of a regression model is computed by comparing to real data x_t as listed in Equation (6).

$$\text{Err} = \frac{|x_t - y_t|}{x_t} \times 100 \qquad (6)$$

Figure 5 presents percentage error of prediction using multiple regression models tabularized in Table 4 after comparing to real data which has occurred subsequently. Figure 5 represents the investigation of public health dataset while Figure 6 denotes the study of soccer player dataset respectively

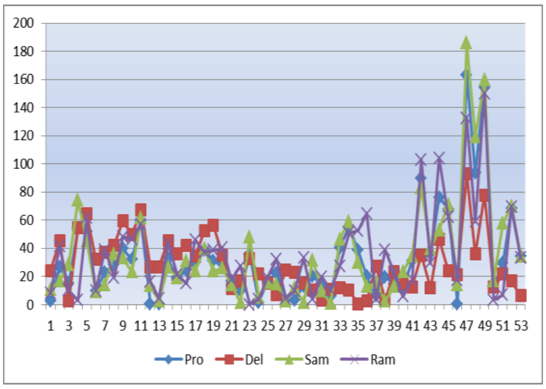


Figure 5. Estimation error of public health data comparing to authentic data

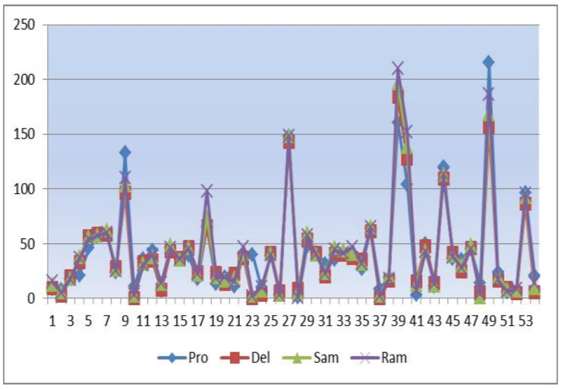


Figure 6. Estimation error of soccer player data comparing to authentic data

Table 6 summarizes the average error of each prediction by calculating from results shown in Figure 5 and Figure 6. It is apparent from the table DEL reflects the lowest average percentage of error and attains best accuracy among other paradigms.

Table 6. Average Error for Individual Estimations

DATASET	Average Error (%)			
	Prototype	DEL	SAM	RAM
Health	33.56	28.97	35.15	34.77
Soccer	42.74	40.7	42.73	44.01

5. CONCLUSION

The research exposes the analysis authors have adopted two general datasets of which both are incorporating missing data and noise. To get over the limitation of unexcutable dataset with NMD, N and MD are divided into two parts. After removing invalid figure, the treatment methods which customize estimation with less bias exercising deletion, single assertion and random mechanisms are proposed. These are of simplified tools and unsubstantial bias types. The key point of current investigation is that a prototype's file avails in order to cope with the missingness. Proposed mechanisms are constituted to reform the missing data and the verification is performed on this adjustment also. The performance of deletion mechanism is good on both datasets as compared to the rest. Remarkable point is that MAE and RMSE remain almost identically in all the two datasets. The unique aspects of the constancy of missing patterns are perceived in the current study. This is the leading time in the literature since no general dataset in practice is obtainable in the form of missing types rather than the re-construction of datasets in the MOA simulation model. From these results it is found that, the essential figures of average error are appropriate to judge which mechanisms can be opted for the sake of full reach. From table above, we can obviously state that DEL has achieved the best accuracy among others.The forthcoming work will consider the addition of the present research focusing on different variations of noise datasets approach in MOA. The next paper will grip a concept of depressing an error percentage of regressive estimation in MOA.

REFERENCES

[1] K. Su, *et al.*, "Smart City and the Applications," *International Conference on Electronics, Communications and Control (ICECC)*, pp. 1028-1031, 2011.
[2] A. Prabahar, "Development of High Performance Wireless Sensor Node for Acoustic Applications," *IEEE International Conference on Green High Performance Computing (ICGHPC)*, pp. 1-5, 2013.
[3] V. Marx, "The Big Challenge of Big Data," Nature 498.7453, pp. 255-260, 2013.
[4] C. Enders, "Applied Missing Data Analysis," Guilford Press, New York, 2010.

[5] H. Eldirdiery and A. H. Ahmed, “Detecting and Removing Noisy Data on Web Document using Text Density Approach,” *International Journal of Computer Applications*, vol. 112, no. 5, pp. 32-26, 2015.

[6] R. Sharma and M. Bhatia, “Eliminating the Noise from Web Pages using Page Replacement Algorithm,” *International Journal of Computer Science and Information Technologies (IJCSIT)*, vol. 5, no. 3, pp. 3066-3068, 2014.

[7] B. Chae, *et al.*, “The impact of advanced analytics and data accuracy on operational performance: A contingent resource based theory (RBT) perspective,” *Decision Support Systems*, vol. 59, pp. 119-126, 2014.

[8] M. A. Shabir and P. T. Deepali, “Satellite Image Denoising Using Discrete Cosine Transform,” *Indonesian Journal of Electrical Engineering and Informatics*, vol. 5, pp. 372-375, 2017.

[9] A. Dharmarajan and T. Velmurugan, “Lung Cancer Data Analysis by k-means and FarthestFirst Clustering Algorithms,” *Indian Journal of Science and Technology*, vol. 8, no. 15, 2015.

[10] A. Bifet, *et al.*, “MOA: Massive Online Analysis,” *Journal of Machine Learning Research*, vol. 11, pp.1601-1604, 2010.

[11] Z. B. Yossefand and S. Rajagopalan, “Template Detection via Data Mining and Its Applications,” *Proceedings of the International Conference on the World Wide Web*, ACM Press, pp. 580-591, 2002.

[12] L. Yi, *et al.*, “Eliminating Noisy Information in Web Pages for Data Mining,” *SIGKDD*, ACM Press, pp. 296-305, 2003.

[13] S. Debnath, *et al.*, “Automatic Extraction of Informative Blocks from Webpages,” *ACM Symposium on Applied Computing*, pp. 1722-1726, 2005.

[14] E. S. Laber, *et al.*, “Fast and Simple Method for Extracting Relevant Content from News Webpages,” *Proceedings of the ACM Conference on Information and Knowledge Management*, ACM Press, pp. 1685-1688, 2009.

[15] N. Raheja and V. K. Katiyar, “Noise Reduction Approach Based on n x 1 Table and XSL Display Method for Efficient Web Data Extraction,” *International Journal of Computer Applications*, vol. 64, 2013.

[16] G. Madhu and G. Nagachandrika, “A New Paradigm for Development of Data Imputation Approach for Missing Value Estimation,” *International Journal of Electrical and Computer Engineering*, vol. 6, pp. 3222-3228, 2016.

[17] R. Thirumahal and P. A. Deepali, “KNN and ARL Based Imputation to Estimate Missing Values,” *Indonesian Journal of Electrical Engineering and Informatics*, vol. 2, pp. 119-124, 2014.