ISSN 2278-3091
Volume 8, No.3, May - June 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse05832019.pdf
https://doi.org/10.30534/ijatcse/2019/05832019

A Study on Sequential K-Nearest Neighbor (SKNN) Imputation for Treating Missing Rainfall Data

Wai Yan Lai[1], Kuok King Kuok[2], Shirley Gato-Trinidad[3], Derrick, Kuo Xiong Ling[4]
[1]Swinburne University of Technology Sarawak Campus, Malaysia, wlai@swinburne.edu.my
[2]Swinburne University of Technology Sarawak Campus, Malaysia, kkuok@swinburne.edu.my
[3]Swinburne University of Technology, Australia, sgatotrinidad@swin.edu.au
[4]Swinburne University of Technology Sarawak Campus, Malaysia, dling@swinburne.edu.my

ABSTRACT

This paper demonstrates a novel application of a gene imputation model, Sequential K-Nearest Neighbor (SKNN) imputation model to address the issues of missing rainfall data in Kuching City. To determine the reliability and robustness of SKNN imputation model in treating the missing rainfall data, an experiment was done to compare the imputation performance of SKNN against a conventional imputation model, K-Nearest Neighbor (KNN). The experiment was conducted using datasets with different missing entries (1%, 5%, 10%, 15%, and 20% of missing data entries). The datasets were created by artificially introducing the missing entries into a complete rainfall dataset. The imputation performance of the imputation models was evaluated with respect to Bias (B_S), Root Mean Square Error (RMSE), Coefficient of Correlation (r), and Index of Agreement (d). The SKNN was found to be superior to KNN in terms of accuracy and imputation performance. It was also reported that RMSE and B_S can express the relationship of missing data entries and imputation performance significantly.

Key words : K-Nearest Neighbor (KNN), missing rainfall data, rainfall imputation, Sequential K-Nearest Neighbor (SKNN)

1. INTRODUCTION

Missing data entries is one of the biggest issues suffered in many fields of research especially in the field of future data prediction and simulation studies. Due to the absence of complete datasets, the accuracy and performance of prediction studies are greatly affected hence resulting in unreliable and unrealistic studies. Missing data can be caused by human errors in collecting and managing the datasets, natural disaster, and machinery defects on site. Conventional methods such as hot-deck imputation, mean imputation, data deletion, and zero imputation are widely implemented to treat the missing data entries. However, these methods are not applicable in every scenario and they have several drawbacks that may lead to biased and inaccurate studies. Listwise deletion (LD) is a common imputation method used in treating missing data entries. Under LD, the missing data entries will be deleted, and useful information may be ignored and eliminated [1]. This will ultimately affect the accuracy of prediction model as bias may be introduced within the simulation studies. Missing rainfall data in Malaysia is currently treated using hot deck imputation method [2]. The missing rainfall data are substituted by other rainfall data collected from nearby gauging instruments or rainfall stations. However, this method is not feasible when the missing rainfall records occur simultaneously at the other gauging instruments or stations.

Statistical methods, data mining methods, and machine learning methods are utilised in the recent decade to treat missing data. These methods had been successfully deployed to address missing data in various fields. Tawarish and Satyanarayana [3] utilized data mining algorithm for predicting the stock market. Gold and Bentler [4] had successfully applied expectation maximization (EM) algorithm into two different imputation models, structured model expectation maximization and saturated model expectation maximization for imputing the incomplete data. The obtained results showed that the expectation maximization algorithm is capable in reconstructing the random missing data. Artificial neural network is also another viable option to predict the missing rainfall data where Kueh and Kuok [5] used bat optimisation neural network to predict future rainfall data. Tfwala, et al. [6] had successfully applied multilayer perceptron neural network to treat missing flow records. The study condition had been set to make sure that the flow records from adjacent stations were always available. The performance of multilayer perceptron neural network outperformed the coactive neurofuzzy inference system in imputing the missing flow records. Bennett, et al. [7] had used nearest neighbor by distance (ND), nearest neighbor by correlation (NC), inverse distance weighted (IDW), average of gauges selected by correlation (A) and weighted average of gauges selected by correlation (WA) to impute the missing rainfall data. The results showed that WA method outperformed the rest. Similar approaches were also adopted by Kamaruzaman, et al. [8] where inverse distance weighted (IDW), modified correlation weighted (MCW), combination correlation with inverse distance (CCID), and averaging correlation and inverse distance method (ACCID) were used to patch the missing rainfall data.

This research thus, aims to introduce a data mining algorithm, Sequential K-Nearest Neighbor (SKNN) imputation model in

Sarawak, Malaysia, for treating the missing rainfall records. SKNN is a variation of K-Nearest Neighbor (KNN) imputation model introduced by Kim, et al. [9] to impute missing DNA microarray data. KNN method is proven to be reliable and practical to treat missing hydrological data through the application of KNN imputation conducted by Lee and Kang [10]. Through the research, it is found that SKNN is yet to be introduced to treat any missing hydrological data. Thus, it is motivated to investigate the performance of SKNN to treat missing rainfall data. However, it is expected that it will be difficult to evaluate the imputation performance on missing rainfall data as rainfall data usually have random patterns that cause the prediction to be difficult [11]. Hence, using conventional evaluation method such as Coefficient of Correlation (r) may not be suitable or significant to analyse and differentiate the actual performance of imputation models. By referring to the aforementioned motivations and challenges, the objective of this research is outlined as below:

- To reconstruct the missing rainfall data via SKNN model
- To study the parameters that will affect the performance of SKNN model
- To compare the performance of SKNN model against conventional KNN model
- To identify the methods which are effective and reliable in evaluating the precipitation prediction model
- To evaluate the performance of SKNN model under different amount of missing data entries (1%, 5%, 10%, 15%, and 20% of missing rainfall data entries)

2. STUDY AREA AND DATASET

The state of Sarawak had been chosen to be the study area. The Sarawak River Basin of Kuching City as illustrated in Figure 1 was focused in this research study. The Kuching International Airport's rainfall station (Site 14030001) was selected to carry out the imputation study. The rainfall data of Kuching International Airport rainfall station in the year of 1951 was collected from Department of Irrigation and Drainage (DID) Sarawak. The collected rainfall data was used for the creation of input dataset for both SKNN and KNN imputation models.

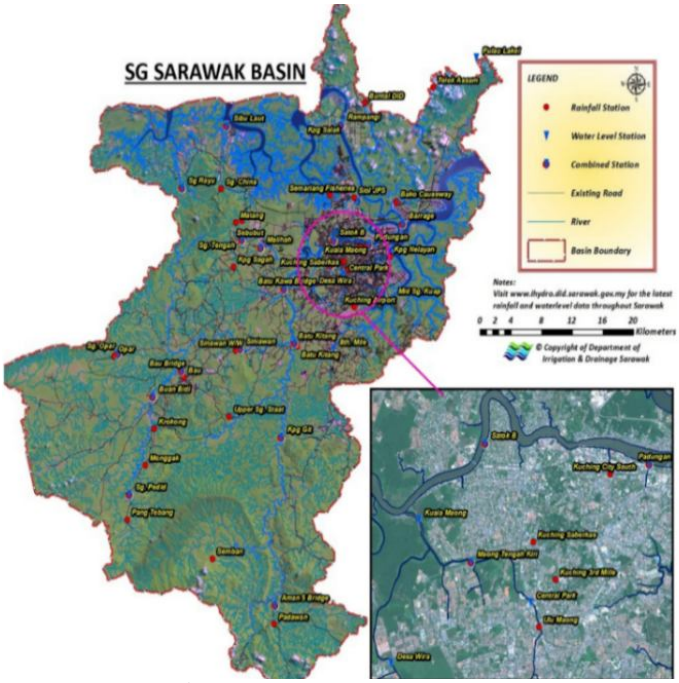

Figure 1: Rainfall Stations Within Sarawak River Basin [12]

3. METHODOLOGY

3.1 K-Nearest neighbor (KNN)

A built-in KNN imputation model of MATLAB was adopted in this study to compare the difference between the conventional KNN imputation model and SKNN imputation model. The built-in function for KNN, "knnimpute" was adopted in this study. The algorithm is set in such a way that it will identify the nearest neighbor column. The values from nearest neighbor column without any missing values will be referenced to carry out the imputation. The algorithm utilises distance metrics such as Euclidean distance and Minkowski distance to determine the nearest neighbor column. The default distance metric of KNN imputation in MATLAB, Euclidean distance as shown in (1), was selected to be the distance metric for KNN imputation model in this study. The missing values will be imputed by taking the weighted mean of the k^{th} nearest-neighbor columns. The parameter, K refers to the number of nearest neighbor that needs to be referenced by the imputation model to carry out the imputation. The weight of the i^{th} entry is calculated using (2).

$$Euclidean\ distance = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (1)$$

where p and q are the vectors of two different datasets

$$W = \frac{\frac{1}{D_i}}{\sum_{i=1}^{k}\frac{1}{D_i}} \quad (2)$$

D_i = the distance between i^{th} entry and an entry to be imputed

3.2 Sequential K-Nearest Neighbors (SKNN)

Kim, et al. [9] stated that the proposed SKNN method is slightly different with the conventional KNN imputation method. The gene imputation process of SKNN is carried out sequentially in two main stages. The term, "gene" in this case is noted as the gene data within the dataset used in their experiment. The proposed SKNN method will arrange and sort the gene accordingly with respect to their missing rate and followed by the imputation process. This means that the data set will be separated into incomplete and complete set, which only the incomplete set will have missing values. The genes in the incomplete set will be filled by taking the mean of the nearest neighbor genes in the complete set where Euclidean Distance is used as the distance metric to determine the nearest neighbor genes. The imputation is set to take place in sequence by referring to their respective missing rate. The imputed data set will move into the complete data set and it will be referenced for executing the rest of the imputation. It is reported that the SKNN method can execute faster than the KNN method. This is because SKNN method allows simultaneous imputation of all the missing values by referring to the selected neighbor gene in the complete set. To the knowledge of authors, the SKNN method is found yet to be used to treat any missing hydrological data. The results from the experiment performed by Kim, et al. [9] also show that SKNN method can be implemented for treating time series data. Hence, SKNN is proposed in this research for treating the missing rainfall data. From the literature above, it shows that KNN is reliable in treating missing hydrological data. Thus, KNN is also included in this study as a comparison against SKNN in treating missing rainfall data.

3.3 Evaluation Methods

Willmott [13] commented the Correlation Coefficient (r) and Coefficient of Determination (r^2) do not provide much information and they are not consistently related to the accuracy of the prediction. The author recommended other approaches such as Root Mean Square Error, Mean Square Error, Mean Bias Error and Index of Agreement to evaluate the performance of prediction model in the climatological field. Thus, the evaluation methods are selected by referring to the methods as proposed by Willmott [13] and other relevant hydrological papers that had used the similar approach. Wang, et al. [14] stated that it is difficult to estimate the convective precipitation forecasts using uniform verification method due to the fact that the convective precipitation fields change drastically and rapidly. The authors suggested to use Bias (B_S), Index of Agreement (d), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for evaluating the performance of the prediction model in hydrology application. Thus, r, d, B_S and RMSE were selected to evaluate the performance of the proposed imputation model and their formulae are expressed as in (3) to (6). A perfect estimation of values will result in r = 1, B_S = 1, d = 1 and RMSE = 0.

$$r = \frac{\sum(F - \bar{F})(O - \bar{O})}{\sqrt{\sum(F - \bar{F})^2 \sum(O - \bar{O})^2}} \quad (3)$$

$$d = 1 - \frac{\sum_{i=1}^{N} (O_i - F_i)^2}{\sum_{i=1}^{N} (|O_i - \bar{O}| + |F_i - \bar{O}|)^2} \quad (4)$$

$$B_S = \frac{\sum_{i=1}^{N} F_i}{\sum_{i=1}^{N} O_i} \quad (5)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (O_i - F_i)^2}{N}} \quad (6)$$

where,
F = imputed value or predicted value
O = original value or observed value
$\bar{O}$ = mean of original value or observed value
$\bar{F}$ = mean of imputed value or observed value
N = number of data

4. CASE STUDY

The missingness mechanism in this study was assumed to be Missing Completely at Random (MCAR) as suggested by Malek [15]. This is because the missing data in Malaysia is mainly contributed by errors and mistakes in data management, human resources, operation and maintenance. Hence, it is reasonable to induce that the occurrences of missing rainfall data do not depend on any occurrences of random events. The general outline of the case study is described under the list below:

Step 1: Creation of input datasets without any missing values using the daily rainfall data collected from DID Sarawak
Step 2: Introduction of different artificial missing entries for all the datasets (1%, 5%, 10%, 15%, and 20% of missing rainfall data entries)
Step 3: Import the rainfall data and source code into MATLAB
Step 4: Execution of imputation under different parameters settings (different K values and percentage of missing data entries)
Step 5: Evaluation on the performance of SKNN and KNN imputation model using different evaluation methods

The daily rainfall data of Kuching International Airport Station was arranged into a matrix of (X × Y) where X and Y represent daily rainfall amount and month, respectively. The missing data entries were introduced artificially into the complete dataset as created in Step 1 at different missing entries. The datasets at different missing entries were used as the input dataset for both imputation models. The performance of both the imputation models were evaluated against r, d, B_S and RMSE by referring to the original and imputed datasets as obtained from the experiment.

5. RESULTS AND DISCUSSION

In order to ease the analysis process, only a fraction of the results is included in this work. The imputation performance of both the imputation models are presented as in Table 1, Figure 2, and Figure 3. Table 1 only shows the best imputation

performance achieved by both imputation models. Generally, the imputation performance of both KNN and SKNN is about the same as all the tabulated results do not show a large difference in terms of all the evaluation methods at their respective percentage of missing entries. The obtained results are logical as the imputation performance become worse when the percentage of missing entries increase. This is because there will be more values needed to be predicted when the missing entries increase hence causing lower imputation performance at higher missing entries.

It is evident that SKNN is superior to KNN at the missing entries of 10% and above. This shows that SKNN is more capable in predicting more accurate results at larger missing entries when compare to conventional KNN imputation method. In terms of evaluation methods, B_S and RMSE are better than d and r. RMSE is more capable in showing the relationship between the percentage of missing entries and the imputation performance. In Table 1, it can be observed that the fluctuation of RMSE values is more significant as the missing percentage increases. Unlike r and d, it only shows a little difference when the percentage of missing entries increases. On the other hand, the calculation of B_S serves as an estimation to judge whether the predicted values are overestimated ($B_S > 1$) or underestimated ($B_S < 1$). From the values tabulated in Table 1, it is observed that a slight overestimation ($B_S > 1$) of data occurred for both SKNN and KNN as the B_S values are not far from 1. This also shows that the tendency of both SKNN and KNN to overestimate the missing values is high. On the other hand, the tabulation of r and d serve as a basis to judge the similarities between the imputed dataset and the original dataset. However, the information that can be observed through r and d is limited. This is because the values of r and d are still very high (> 0.87) at high missing data entries of 20%.

Figure 2 and Figure 3 illustrate the RMSE achieved by both SKNN and KNN imputation model at different K values and percentage of missing entries. The graphs show that both KNN and SKNN are affected by the chosen nearest neighbor value, K. It is evident that the performance of both imputation models become better when larger value of K is chosen. It is reported that the range of K values that can be adopted by KNN imputation model is not defined. This then requires the user to identify the suitable range of K values via trial and error method. From Figure 3, it shows that the RMSE of KNN method at any missing entries will stay the same when the K values reaches 12. It is also reported that the same phenomenon occurred at the other evaluation methods where the performance stays the same when K reaches a value of 12 and above. This means the range of K values that should be adopted by KNN in this study should fall within the range of 1 to 12. On the other hand, the range of K values that can be adopted by SKNN model reduces as the missing entries increases. This can be observed as in Figure 2 where it is evident that the adoptable K values at 20% missing entries is the least. This is caused by the reduction of complete dataset that can be referred for carrying out the imputation as there will be more incomplete dataset at higher percentage of missing data entries. It is also significant that SKNN can execute faster than KNN as there is lesser K values that need to be tested and there will be no need for the users to determine the range of K values. Hence, it can be said that SKNN imputation model is more convenient when compared to the conventional KNN imputation model.

6. CONCLUSION

This paper demonstrates the application of SKNN imputation model to treat missing rainfall data. The results show that SKNN imputation model is reliable as the imputation performance of SKNN imputation model is about the same as KNN imputation model. The results also show that SKNN is superior to KNN imputation model in treating the missing rainfall data as the percentage of missing entries is at 10% and above. SKNN is more convenient than KNN as it reduces the time to identify the range of adoptable K values. B_S and RMSE are recommended to be used to evaluate the imputation performance of imputation models in hydrological application as they can illustrate the relationship between the imputation performance and percentage of missing data entries significantly. The proposed imputation model, SKNN is recommended to be executed in other hydrological applications to determine if the SKNN imputation model is capable in treating other missing hydrological data. Further comparison of SKNN with other data mining algorithms in treating the missing data is recommended to judge the performance of SKNN in treating other missing hydrological data. It is also recommended to seek for a better approach to improve the imputation performance of SKNN at high missing data entries.

ACKNOWLEDGEMENT

The authors would like to thank all the reviewers for their feedback that help to improve the quality of this manuscript.

REFERENCES

[1] I. Myrtveit, E. Stensrud, and U. H. Olsson, "Analyzing data sets with missing data: an empirical evaluation of imputation methods and likelihood-based methods," *IEEE Transactions on Software Engineering*, vol. 27, no. 11, pp. 999-1013, 2001. https://doi.org/10.1109/32.965340

[2] M. A. Malek, S. M. Shamsuddin, and S. Harun, "Restoration of hydrological data in the presence of missing data via Kohonen Self Organizing Maps," in *New Trends in Technologies*, B. Ramov, Ed. Croatia: InTech, 2010, pp. 223-243. https://doi.org/10.5772/7582

[3] M. Tawarish and K. Satyanarayana, "A Review on Pricing Prediction on Stock Market by Different Techniques in the Field of Data Mining and Genetic Algorithm," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 1, pp. 23-26, 2019. https://doi.org/10.30534/ijatcse/2019/05812019

[4] M. S. Gold and P. M. Bentler, "Treatments of Missing Data: A Monte Carlo Comparison of RBHDI, Iterative

Stochastic Regression Imputation, and Expectation-Maximization," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 7, no. 3, pp. 319-355, 2000. https://doi.org/10.1207/S15328007SEM0703_1

[5] S. M. Kueh and K. K. Kuok, "Precipitation downscaling using the artificial neural network BatNN and development of future rainfall intensity-duration-frequency curves," *Climate Research*, vol. 68, no. 1, pp. 73-89, 2016. https://doi.org/10.3354/cr01383

[6] S. S. Tfwala, Y.-M. Wang, and Y.-C. Lin, "Prediction of Missing Flow Records Using Multilayer Perceptron and Coactive Neurofuzzy Inference System," *The Scientific World Journal*, vol. 2013, pp. 1-7, 2013, Art. no. 584516. https://doi.org/10.1155/2013/584516

[7] N. Bennett, L. Newham, B. Croke, and A. Jakeman, "Patching and disaccumulation of rainfall data for hydrological modelling," in *Int. Congress on Modelling and Simulation (MODSIM 2007), Modelling and Simulation Society of Australia and New Zealand Inc., New Zealand*, 2007, pp. 2520-2526.

[8] I. F. Kamaruzaman, W. Z. W. Zin, and N. M. Ariff, "A comparison of method for treating missing daily rainfall data in Peninsular Malaysia," *Malaysian Journal of Fundamental and Applied Sciences*, vol. 13, no. 4-1, pp. 375-380, 2017. https://doi.org/10.11113/mjfas.v13n4-1.781

[9] K.-Y. Kim, B.-J. Kim, and G.-S. Yi, "Reuse of imputed data in microarray analysis increases imputation efficiency," *BMC Bioinformatics*, journal article vol. 5, no. 1, p. 160, October 26 2004. https://doi.org/10.1186/1471-2105-5-160

[10] H. Lee and K. Kang, "Interpolation of Missing Precipitation Data Using Kernel Estimations for Hydrologic Modeling," *Advances in Meteorology*, vol. 2015, pp. 1-12, 2015, Art. no. 935868. https://doi.org/10.1155/2015/935868

[11] K. K. Kuok and N. Bessaih, "Artificial neural networks (ANNS) for daily rainfall runoff modelling," *Journal-The Institution of Engineers, Malaysia*, vol. 68, no. 3, 2007.

[12] Department of Irrigation and Drainage (DID) Sarawak. (2018, 7 August 2018). *Rainfall Stations*. Available: https://ihydro.sarawak.gov.my/iHydro/en/map/rainfall-map.jsp

[13] C. J. Willmott, "Some comments on the evaluation of model performance," *Bulletin of the American Meteorological Society*, vol. 63, no. 11, pp. 1309-1313, 1982. https://doi.org/10.1175/1520-0477(1982)063<1309:SCOTEO>2.0.CO;2

[14] G. Wang, D. Wang, J. Yang, and L. Liu, "Evaluation and Correction of Quantitative Precipitation Forecast by Storm-Scale NWP Model in Jiangsu, China," *Advances in Meteorology*, vol. 2016, pp. 1-13, 2016, Art. no. 8476720. https://doi.org/10.1155/2016/8476720

[15] M. A. Malek, "Rainfall Data in-Filling Model with Expectation Maximization and Artificial Neural Network," PhD, Universiti Teknologi Malaysia, Skudai, 2008.

Table 1 : Result Summary of KNN and SKNN Imputations

Imputation Method	KNN					SKNN				
Missing Entries	1%	5%	10%	15%	20%	1%	5%	10%	15%	20%
B_S	0.9998	1.0059	1.0143	1.0383	1.0384	1.0004	1.0161	1.0195	1.0267	1.0110
RMSE (mm)	0.1700	1.6652	4.1106	4.8077	7.2478	0.3547	1.7093	3.6201	4.7087	7.0154
d	1.0000	0.9966	0.9785	0.9700	0.9292	0.9998	0.9964	0.9834	0.9715	0.9356
r	0.9999	0.9932	0.9578	0.9421	0.8653	0.9997	0.9929	0.9676	0.9446	0.8756

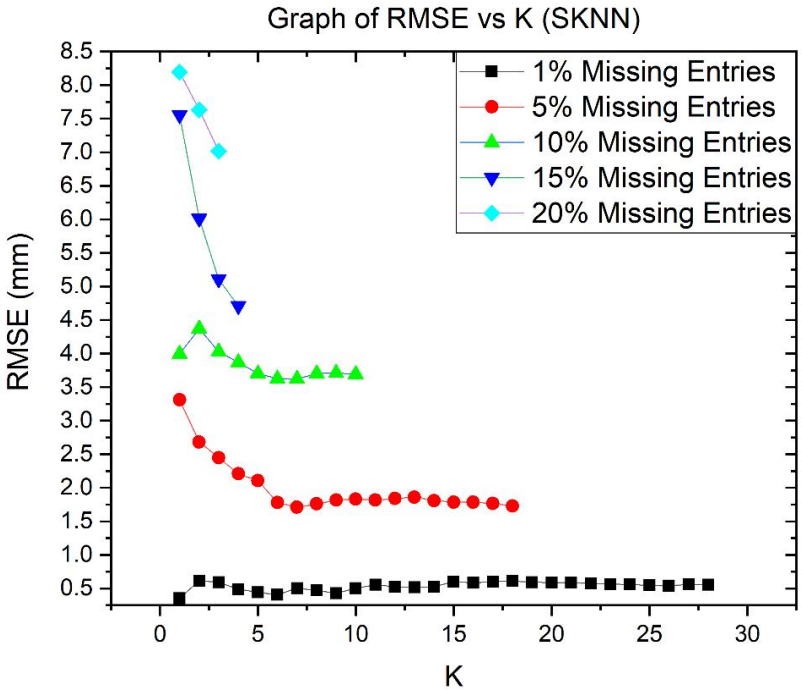

Figure 2: Graph of RMSE vs K of SKNN Imputation

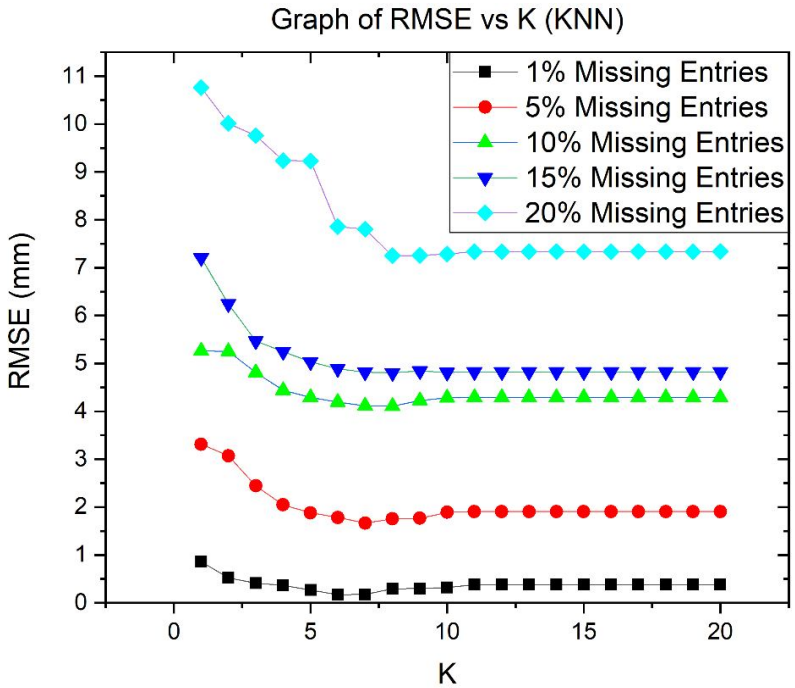

Figure 3: Graph of RMSE vs K of KNN Imputation