MULTIPLE IMPUTATION FOR HYDROLOGICAL MISSING DATA BY USING A REGRESSION METHOD (KLANG RIVER BASIN)

Mehrdad Habibi Khalifeloo[1], Munira Mohammad[2], Mohammad Heydari[3]

[1]B.Eng Student, Civil Engineering, SEGi University, Kuala Lumpur, Malaysia
[2]Assistant Professor, Civil Engineering Department, SEGi University, Kuala Lumpur, Malaysia
[3]PhD candidate, Faculty of engineering, University of Malaya, Kuala Lumpur, Malaysia

Abstract

Rainfall amounts and water surface elevation are considered as one of the most important climatic parameters. Because these two parameters will have a direct impact on water resources management decisions such as meet the water needs and prevent flooding. But in some cases, for some reason all time series data are not fully recorded. To fill the gaps in the data, several interpolation methods currently used. One of these methods is regression analysis as a statistical method. By using regression, we can determine the mathematical relationship coefficients between inputs and outputs. By achieving the equation, we can obtain the unknown quantities. In this research, the daily data between 2005 to 2015 for 5 Rain-gauge stations and 3 elevation measurement of water surface stations in the Klang River Basin were used. The main goal was to find the missing value of the water level in the mentioned three stations by rainfall and water level data. To evaluate the obtained results, Multiple R, R^2, and Standard Error were used. The results indicate that the standard error in normalized data was less than the regular data. Multiple r values for the Klang at Taman Sri Muda1, Klang at Jam, Sulaiman, WP and Klang at Emp Genting Klang, WP are 0.35, 0.42 and 0.28, respectively.

Keywords: *Filling gaps, Interpolation, Data mining, Statistical method, Rainfall, Water level*

----------***----------

1. INTRODUCTION

Recently, researchers attention has been drawn to analyzing the time series of rainfall [1],water level of river [2], temperature, sunshine hours, etc in order to investigate climate parameters [3]. Considerable changes in rainfall or water surface level directly effect on environmental concerns such as drought and flood [4]. Nonetheless, not all the hydrological data or climate data are available in most real issues, because data collection is not often fully carried out in the scientific research. Among the various reasons for the lack of some data can be pointed out to considering unimportant data while entering, being some flaws in the data recording equipment, lack of data entry because it is hard to be understood, incompatibility of the data with other data. There are various terms and often a synonym for the concept in the statistical literature. These terms include missing values, missing data, incomplete and unanswered data. But what must be considered, it is missing data is much better than wrong answers in the data. Each method that we use to analyze the missing data has the weaknesses and strength which depends on the factors such as the ratio and missing pattern, type and number of used variables, missing mechanism. Statistically, some missing data are completely independent from the data which has been observed yet, this data are called "Missing Completely at Random". In some cases, missing values are also "Missing at Random" and they are provided by a number of variables or data predictive class. Another set of missing data is considered "No Ignorable Missing Data". Many researchers have provided ways to deal with the problem of missing data.

Many researchers have provided ways to deal with the problem of missing data in their research, especially in the field of hydrology. Among these researches, we can be mentioned the inverse distance methods [5, 6], Kriging method [7, 8], the nearest neighbor method [9, 10], the linear interpolation [11], the arithmetic mean method [12], artificial intelligence techniques [13] and the regression method [14], as a well-known and powerful method. Regression analysis is one of the most popular methods among the mentioned methods above. There is a lot of research in the field of climate parameters prediction [15] such as rainfall [16, 17], solar radiation [18] and temperature [19] by regression method.

The purpose of this study is to estimate the missing data of water surface elevation for the three stations in Klang River Basin between 2005 and 2015 by regression analysis.

2. MATERIAL AND METHODS

2.1 Case Study (Klang River Basin)

The Klang River Basin flowing through Selangor and Kuala Lumpur has experienced flooding for more than a decade. As a capital city of a developing country such as Malaysia, it suffers from urbanization and a high rapid population. The catchment area of the Klang River Basin is 1288km^2 with a total stream length of approximately 120 km. Located at 3°17'N, 101°E to 2°40'N,101°17'E, it covers areas in Sepang, Kula Langat, Petaling Jaya, Klang, Gombak and Kuala Lumpur.

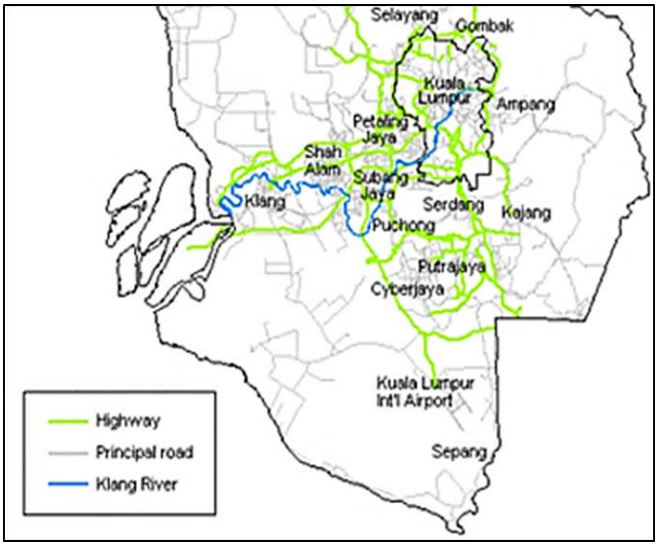

Fig -1: Klang river basin as a study area

Figure 2 shows a view of the data discussed in this study shows. As you can see, the missing data were highlighted in yellow. In Figure 2, we have:
R1:(Rainfall in the UITM Shah Alam at Selangor station)
R2:(Rainfall in the Bukit Kerayong at Selangor station)
R3:(Rainfall in the JPS Ampang at Ampang station)
R4:(Rainfall in the Empangan Kelang at Gombak station)
R5:(Rainfall in the Genting Sempah at Kala Lumpur station)
WL1:(Water level in the Klang at TmnSri Muda1, Selangor)
WL2:(Water level in the Klang at jam. Sulaiman, w.p)

WL3:(Water level in the Klang at Emp. Genting Klang, w.p)

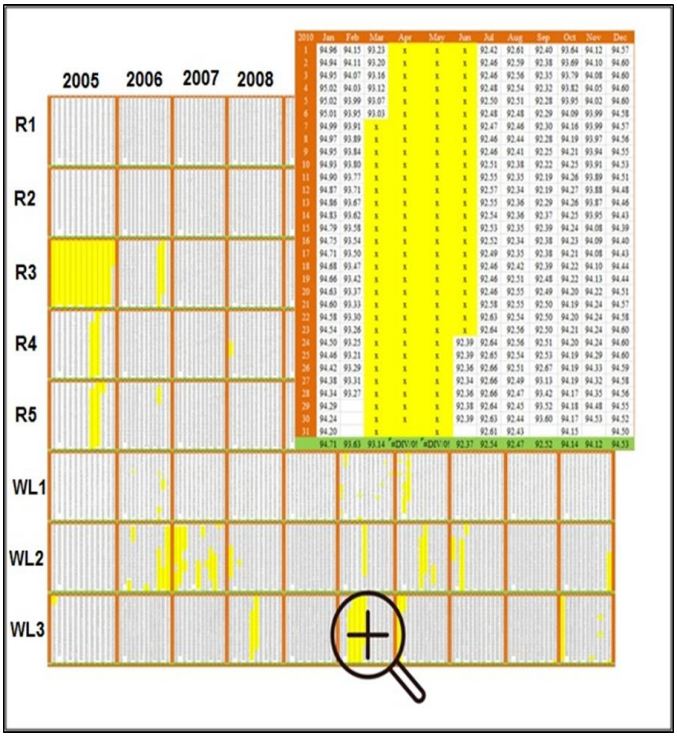

Fig -2: Our data and the missing data (in yellow color)

2.2 Regression Method

One of the most widely used statistical methods in different science is an implementation of regression techniques to determine the relationship between a dependent variable with one or more independent variables. The dependent variable, response and independent variables are also called explanatory variables. A linear regression model assumes there is a linear relationship (direct line) between the dependent variable and the predictor. Running a regression model is possible by defining the regression model. The linear regression model with the dependent variable Y and independent variable p x1, x2,..., xp is defined as follows [20]:

$$y_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip} + e_i \quad \text{(Eq 1)}$$

Where
y_i: is the amount of i^{th} dependent variable
p: is the number of predictors
b_j: is the amount of i^{th} coefficient, j = 0,..., p
X_{ij}: is the value of i^{th} of j^{th} predictor
ei: is the observed error of the value for i^{th}

The model is linear because the value of dependence of b_i is increased by increasing predictive value -i^{th}. b_0 is the intercept, that when any predictive value is zero, the value of predictor model b0 is the dependent variable. In order to test hypotheses about the values of model parameters, linear regression model also takes into consideration the following assumptions:

• The error term has a normal distribution with a mean of zero
• The variance of the error term is constant in all cases and it is independent of the variables in the model. (An error term with inconstant variance is called heteroscedastic).
• The amount of the error term for a given amount is independent of the variable values' s in the model and independent of the amount of error term or the other cases.

2.3 Methodology

2.3.1 Preprocessing

Cleaning incomplete data: Since, it is not possible to achieve real and effective results without having correct, reliable and effective input, before any analysis we must ensure the accuracy and appropriateness of the data and information. This vital issue led to preparing and the data are considered the basis for the subsequent analysis before their actual application. As we see in Figure (2), missing data or incomplete amounts in question has been marked by yellow. In the first step, all incomplete data (999 historical record) were deleted and the remaining data (2653 data) were used for regression and evaluating the obtained models.

Normalizing data: This is due to the presence of different ranges in the data, especially between the data of precipitation and water level data. Have different ranges

can result in ignoring the small range data range compared to the large range data. The problem will be solved by normalizing the attributes so that values are in the same range. This will minimize the effect of real scale and all entries are almost in a range. Therefore, using the following equation, all data were put between zero and one, respectively. Pictures (3) to (5) show the normalized data of this study between zero and one.

$$xN = (x - MinX) / (MaxX - MinX) \quad \text{(Eq 2)}$$

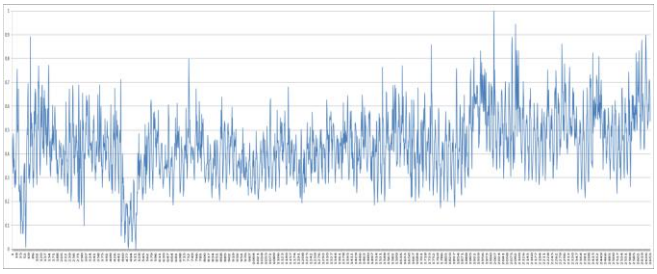

Fig -3: Water level 1 (Klang at Taman Sri Muda1, Selangor)

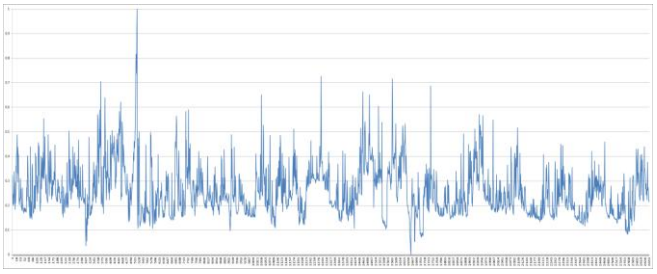

Fig -4: Water level 2 (Klang at Jam. Sulaiman, W.P)

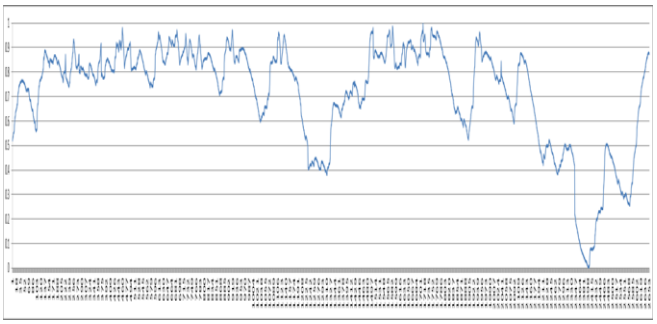

Fig -5: Water level 3 (Klang at EMP.Genting Klang, W.P)

2.3.2 Implementation

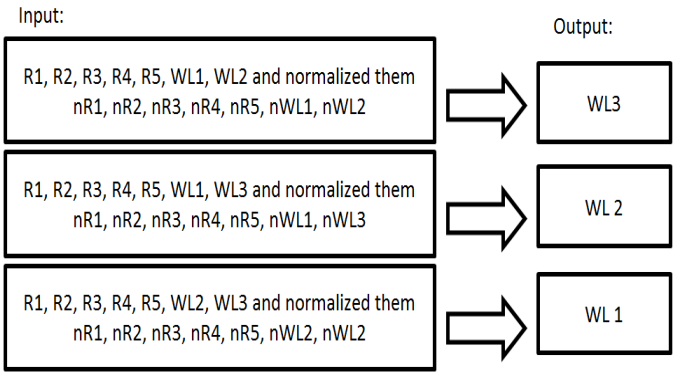

Fig -6: Input and output of the regression model

The phrase "n" in the first term of "nR1, nR2, nR3, nR4, nR5, nWL1, nWL2, nWL3" represents the normalized of the above data.

3. RESULTS AND CONCLUSION

Line fit plot for seven coefficients of the water level 1,2 and 3 equations are shown in Figure (7). The x1 to x5 are rainfall input coefficients and x6 and x7 are water level input coefficients. Input specifications and coefficients of the regression equation obtained for water surface elevation prediction are displayed in Figure (9).

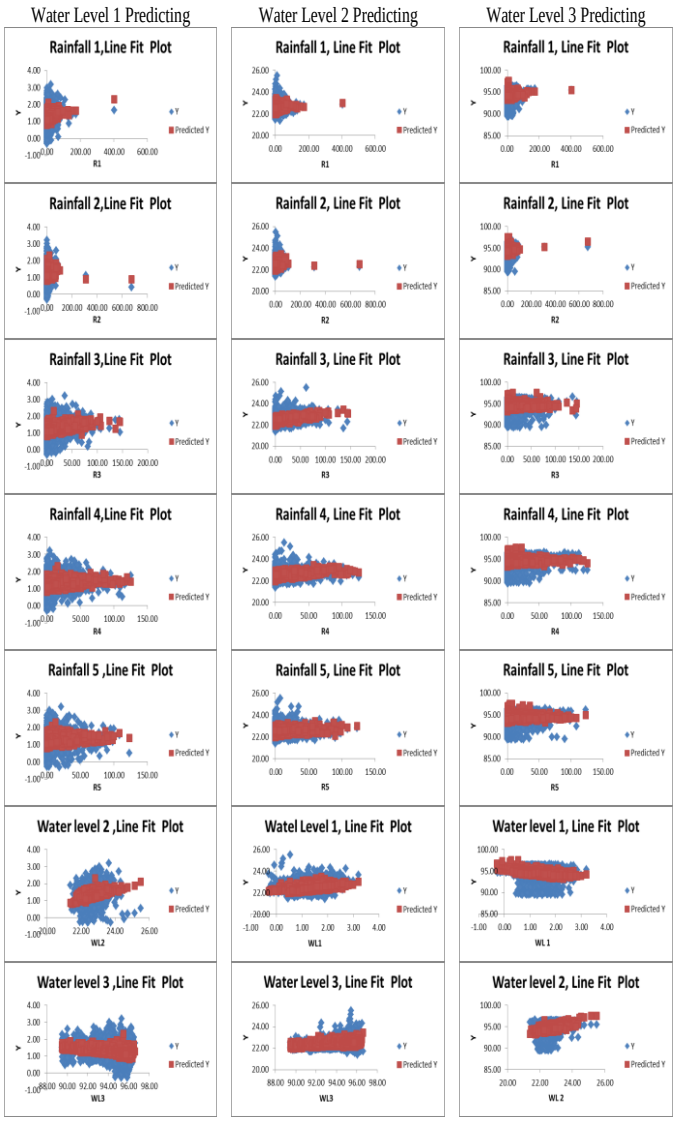

Fig -7: Line fit plot for seven coefficients of the water surface elevation equation

Table -1: Inputs specifications and coefficients of the regression equation obtained for water surface elevation in Klang at Taman Sri Muda1, Selangor

WL 1	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Interc ept	1.711	0.676	2.532	0.011	0.386	3.035
R1	0.003	0.001	4.740	0.000	0.002	0.004
R2	-0.001	0.001	-1.107	0.268	-0.002	0.000
R3	0.003	0.001	3.942	0.000	0.001	0.004
R4	0.001	0.001	1.445	0.149	0.000	0.002
R5	0.001	0.001	0.949	0.343	-0.001	0.002
WL2	0.261	0.022	11.737	0.000	0.218	0.305
Wl3	-0.068	0.006	-11.523	0.000	-0.079	-0.056

Table - 2: Inputs specifications and coefficients of the regression equation obtained for water surface elevation in Klang at jam. Sulaiman, w.p

WL 2	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Inter cept	16.6446	0.4764	34.9395	0.0000	15.7105	17.5787
R1	0.0009	0.0005	1.8852	0.0595	0.0000	0.0018
R2	-0.0001	0.0004	-0.1610	0.8721	-0.0009	0.0008
R3	0.0045	0.0005	8.4556	0.0000	0.0034	0.0055
R4	0.0032	0.0005	5.8463	0.0000	0.0021	0.0042
R5	0.0012	0.0006	2.1277	0.0335	0.0001	0.0023
WL1	0.1895	0.0161	11.7372	0.0000	0.1578	0.2211
Wl3	0.0580	0.0050	11.6056	0.0000	0.0482	0.0678

Table - 3: Inputs specifications and coefficients of the regression equation obtained for water surface elevation in Klang at Emp. Genting Klang, w.p

WL 3	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Inter cept	76.7015	1.5974	48.0169	0.0000	73.5693	79.8338
R1	0.0021	0.0018	1.1819	0.2373	-0.0014	0.0056
R2	0.0023	0.0016	1.4102	0.1586	-0.0009	0.0055
R3	-0.0015	0.0020	-0.7576	0.4488	-0.0055	0.0024
R4	-0.0009	0.0021	-0.4312	0.6663	-0.0049	0.0032
R5	-0.0012	0.0022	-0.5659	0.5715	-0.0055	0.0030
WL1	-0.7065	0.0613	-11.5227	0.0000	-0.8267	-0.5862
Wl2	0.8353	0.0720	11.6056	0.0000	0.6942	0.9764

From the results of the regression model (Table (1) to (3)) can be used the following relations to obtain the values of the unknown and missing data.

WL1 = 1.7105 + 0.0026 R1 – 0.0006 R2 + 0.0025 R3 + 0.0009 R4 + 0.0006 R5 + 0.2613 WL2 – 0.0677 WL3 (Eq 3)

WL2 = 16.6446 + 0.0009 R1 –0.0001 R2 + 0.0045 R3 + 0.0032 R4 + 0.0012R5 + 0.1895 WL1 + 0.0580 WL3 (Eq 4)

WL3 = 76.7015 + 0.0021 R1 + 0.0023 R2 - 0.0015 R3 - 0.0009 R4 - 0.0012 R5 – 0.7065 WL1 + 0.8353 WL2 (Eq 5)

As it is known, water level coefficients and the value of constant number is more important (maximum weight) than the precipitation coefficients in 5 to predict the amount of water in the three stations.

The t-test is one of the simplest and most common tests that are used for man comparison. The full name of this test is Student's t-test.

The P value indicates the probable level that the hypothesis under testing (the null hypothesis) is true. So if the p-value is 0.05, the probability of being true null hypothesis is 0.05. Since in most cases the null hypothesis is tested, we want a lower level of P to reject the null hypothesis. In the form of short, small amounts of p (p less than 0.05 indicates difference and the equal to or greater amount than 0.05 indicates that there is no difference. Obviously, the smaller to be obtained p, one can conclude with more confidence.

Confidence Limits: confidence limits show the accuracy of the computed average. Confidence limits indicate that if the re- sampling of the population is done, the possibility that samples are put in calculated average rang to be 95%.

Table - 4: Regression statistics for water level prediction

	WL 1		WL 2		WL 3	
Regression Statistics	Regular data	Normalized data	Regular data	Normalized data	Regular data	Normalized data
Multiple R	0.353	0.353	0.424	0.424	0.283	0.283
R Square	0.125	0.125	0.18	0.18	0.08	0.08
Adjusted R Square	0.123	0.123	0.177	0.177	0.078	0.078
Standard Error	0.445	0.127	0.379	0.092	1.437	0.201

R-value or correlation coefficient is another statistical tool to determine the type and degree of relationship of a quantitative or qualitative variable. Correlation coefficient shows the intensity of the relationship and also the type of relationship (direct or inverse). This coefficient is between 1 to -1 and if there is no relationship between two variables, it will equal to zero, a large amount of it also shows a strong correlation between the amounts.

The value of R2 is in fact a measure of how well the fitted regression line of the sample is measured. Small amounts show that the model does not comply with the data. The value of R^2 is calculated as follows:

$$R^2 = (1-\frac{Residual\ SS}{Total\ SS}) = \frac{Regress\ SS}{Totall\ SS} \qquad \text{(Eq 6)}$$

Adjusted R Squared has attempted to correct R square to reflect the highest rate of adaption of in the population. Use the coefficient of determination to determine which model is better:

$$AdjR^2=1-(\frac{\text{Total df}}{\text{Residual df}})(\frac{Residual\ SS}{Totall\ SS}) \qquad \text{(Eq 7)}$$

Standard Error is a measure that shows how much the estimated average obtained is accurate. So, if the SE is smaller, better estimation of population has been taking place better and vice versa. SE also is known as the standard deviation of the mean.

$$SE = \frac{SD}{\sqrt{n}} \qquad \text{(Eq 8)}$$

Table -5: ANOVA result

	df [a]	SS [b]			MS [c]		
		WL 1	WL 3	WL 2	WL 1	WL 2	WL 3
Regression	7	74.61	475.86	83.03	10.66	11.86	67.98
Residual	2645	522.77	5458.53	379.09	0.20	0.14	2.06
Total	2652	597.38	5934.39	462.12			
F [d]		53.93	82.76	32.94			
Significance F [e]		2.53E-72	4.77E-109	3.51E-44			

The analysis of variance or ANOVA table checks the acceptance of the statistically. The regression line shows information about a change in your model. Residual line also shows the information about the change that is not intended for your model. In other words, the residual of a product is equal to the observed the error term for the product. Total output also shows the total data related to regression and residual.

(a) The number of independent observations minus the number of estimated parameters is called the degree of freedom -regression. In other words, the degree of freedom -regression- is a dimensional unknown volume (complete model) minus the given volume(bound model).

(b) The Sum of square (SS) is composed of two sources of variance. In particular, it is obtained from the sum of the SSregression and the SSresidual. It shows total variability in the scores of the predicted variable Y.

$$\sum(Y-\overline{Y})^2=\sum(Y'-\overline{Y})^2+\sum(Y-Y')^2 \qquad \text{(Eq 9)}$$

(c) RegressionMS=(RegressionSS)/(Regressiondf) (Eq10)

(d) F ratio is a number which is obtained from dividing the average of Timar squares by Residuals mean.

(e) Based on the F probability distribution, If the Significance F is not less than 0.1 (10%) you do not have a meaningful correlation [21].

ACKNOWLEDGEMENTS

The authors would like to thank the reviewers for their valuable suggestions that have led to substantial improvements to the article.

REFERENCES

[1]. Noori, M., M.B. Sharifi, and M. Heydari, Comparison of the SDSM and LARS-WG weather generators in Modeling of Climate Change in Golestan Province of Iran.

[2]. Othman, F., et al., Prediction of Water Level And Salinity of Lakes By using Artificial Neural Networks, Case Study: Lake Uremia, in 35th International Association for Hydro-Environmental Engineering and Research (IAHR). 2013: China.

[3]. Noori, M., et al., Utilization of LARS-WG Model for Modelling of Meteorological Parameters in Golestan Province of Iran. Journal of River Engineering, 2013. 1.
[4]. Parsa, A.S., et al., Flood Zoning Simulation by HEC-RAS Model (Case Study: Johor River-Kota Tinggi Region). Journal of River Engineering, 2013. 1.
[5]. Di Piazza, A., et al., Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for Sicily, Italy. International Journal of Applied Earth Observation and Geoinformation, 2011. 13(3): p. 396-408.
[6]. Shen, S.S., et al., Interpolation of 1961-97 daily temperature and precipitation data onto Alberta polygons of ecodistrict and soil landscapes of Canada. Journal of applied meteorology, 2001. 40(12): p. 2162-2177.
[7]. Rossi, R.E., J.L. Dungan, and L.R. Beck, Kriging in the shadows: geostatistical interpolation for remote sensing. Remote Sensing of Environment, 1994. 49(1): p. 32-40.
[8]. Teegavarapu, R.S. and V. Chandramouli, Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. Journal of Hydrology, 2005. 312(1): p. 191-206.
[9]. Batista, G.E. and M.C. Monard, An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence, 2003. 17(5-6): p. 519-533.
[10]. Eskelson, B.N., et al., The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scandinavian Journal of Forest Research, 2009. 24(3): p. 235-246.
[11]. Paulhus, J.L. and M.A. Kohler, Interpolation of missing precipitation records. Mon. Wea. Rev, 1952. 80(5): p. 129-133.
[12]. De Silva, R., N. Dayawansa, and M. Ratnasiri, A comparison of methods used in estimating missing rainfall data. Journal of Agricultural Science, 2007. 3: p. 101-108.
[13]. Kuligowski, R.J. and A.P. Barros, Using Artificial Neural Networks To Estimate Missing Rainfall Data1. 1998, Wiley Online Library.
[14]. Haitovsky, Y., Missing data in regression analysis. Journal of the Royal Statistical Society. Series B (Methodological), 1968: p. 67-82.
[15]. Chu, P.-S., et al., Climate prediction of tropical cyclone activity in the vicinity of Taiwan using the multivariate least absolute deviation regression method. Terrestrial Atmospheric and Oceanic Sciences, 2007. 18(4): p. 805.
[16]. Sun, R., L. Chen, and B. Fu, Predicting monthly precipitation with multivariate regression methods using geographic and topographic information. Physical Geography, 2011. 32(3): p. 269-285.
[17]. Gowariker, V., et al., A power regression model for long range forecast of southwest monsoon rainfall over India. Mausam, 1991. 42(2): p. 125-130.
[18]. Goodale, C.L., J.D. Aber, and S.V. Ollinger, Mapping monthly precipitation, temperature, and solar radiation for Ireland with polynomial regression and a digital elevation model. Climate Research, 1998. 10(1): p. 35-49.
[19]. Balaghi, R., et al., Empirical regression models using NDVI, rainfall and temperature data for the early prediction of wheat grain yields in Morocco. International Journal of Applied Earth Observation and Geoinformation, 2008. 10(4): p. 438-452.
[20]. Tunçal, T., Evaluating drying potential of different sludge types: Effect of sludge organic content and commonly used chemical additives. Drying Technology, 2010. 28(12): p. 1344-1349.
[21]. Wilcox, W.R. Explanation of results returned by the Regression tool in Excel's Data Analysis. 2010; Available from: http://people.clarkson.edu/~wwilcox/ES100/regrint.htm.