IMPUTATION OF RAINFALL DATA USING IMPROVED NEURAL NETWORK ALGORITHM

Po Chan Chiu

Prof Ali Selamat

Prof Ondrej Krejcar

King Kuok Kuok

pcchiu@unimas.my

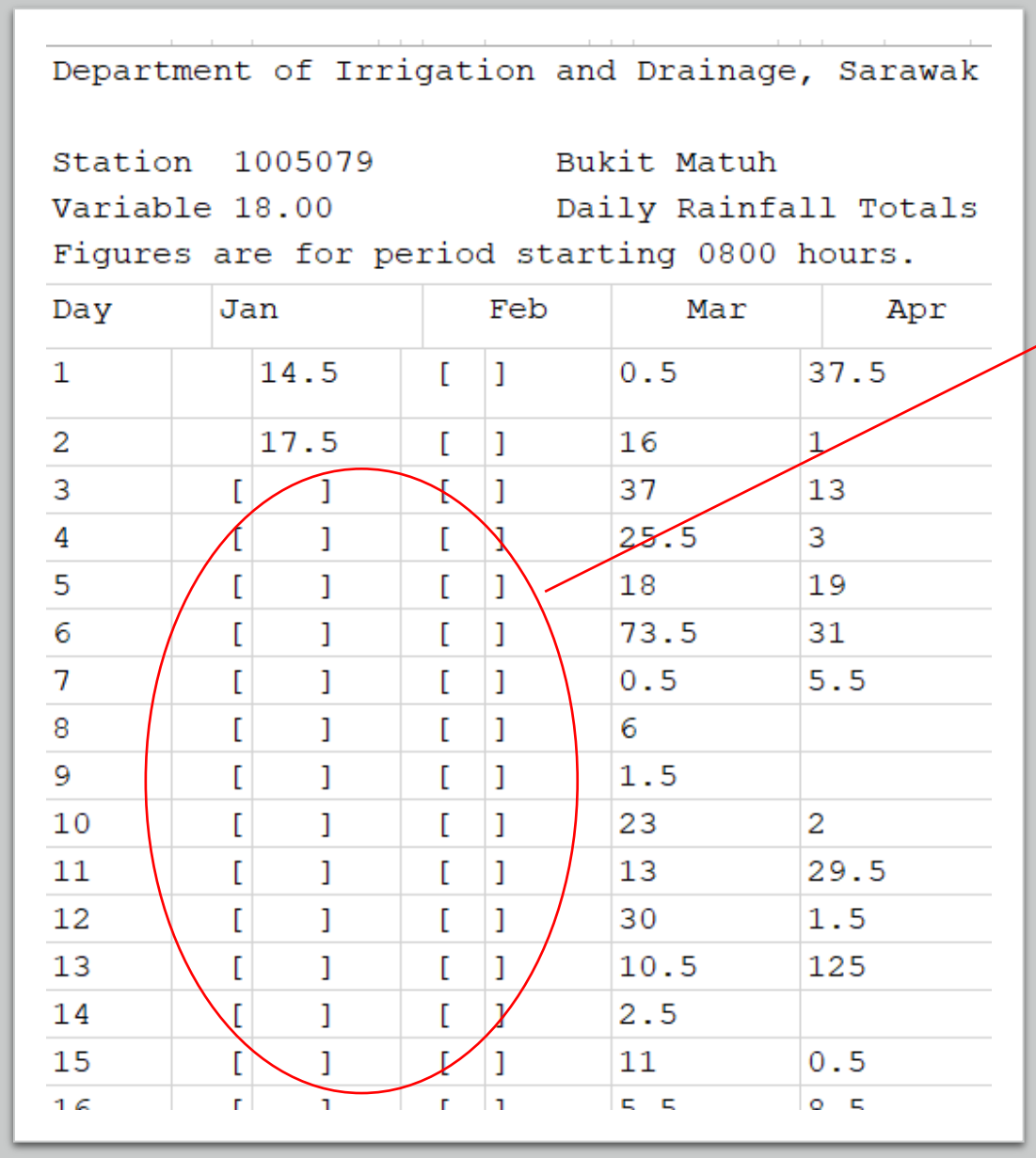
11 January 2021

PRESENTATION OUTLINE

- Introduction
- Related work
- Research methodology
- Imputation method
- Study area
- Result and discussion
- Conclusion
- References

INTRODUCTION

- Rainfall is the amount of rain that falls in an area over a period of time.
- Rainfall is an important component in hydrological research.
- Hydrological research uses rainfall data for flood forecasting, flood risk assessment, and water resources modeling.
- However, hydrologists are often encountered problem of missing values in a rainfall dataset.



Department of Irrigation and Drainage, Sarawak

Station 1005079 Bukit Matuh
Variable 18.00 Daily Rainfall Totals
Figures are for period starting 0800 hours.

Day	Jan	Feb	Mar	Apr
1	14.5	[ ]	0.5	37.5
2	17.5	[ ]	16	1
3	[ ]	[ ]	37	13
4	[ ]	[ ]	25.5	3
5	[ ]	[ ]	18	19
6	[ ]	[ ]	73.5	31
7	[ ]	[ ]	0.5	5.5
8	[ ]	[ ]	6	
9	[ ]	[ ]	1.5	
10	[ ]	[ ]	23	2
11	[ ]	[ ]	13	29.5
12	[ ]	[ ]	30	1.5
13	[ ]	[ ]	10.5	125
14	[ ]	[ ]	2.5	
15	[ ]	[ ]	11	0.5
16	[ ]	[ ]	5.5	8.5

INTRODUCTION

- Missing values in a rainfall database.

- Missing data happens when the values are **incomplete** or **not available** in the datasets.

- It affects **the accuracy and validity** of the hydrological **modelling analysis** result depending on the extent of the missing data.

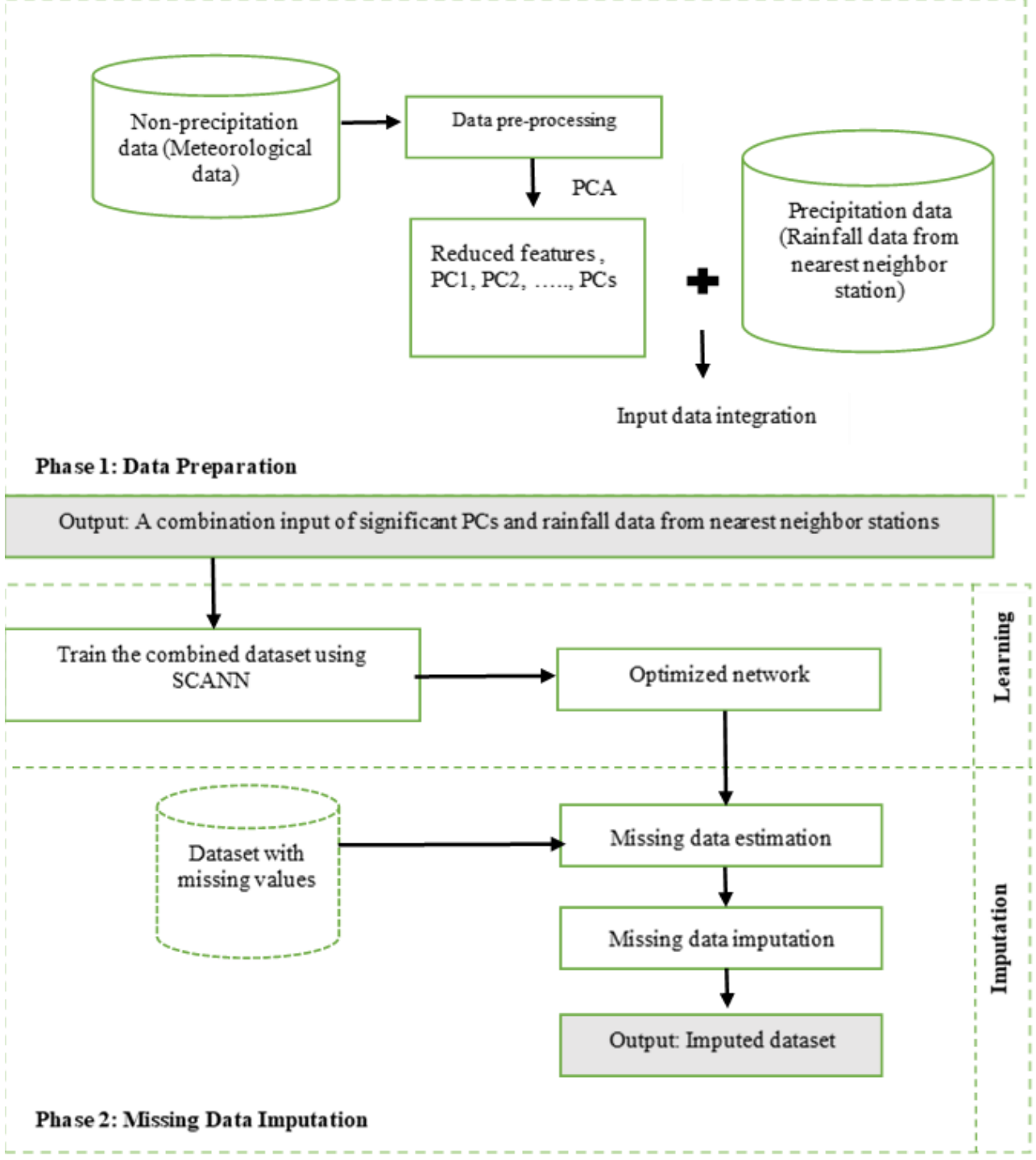
RELATED WORKS

- Imputation is a procedure that is used to fill in missing values with substitutes (Eskelson, 2009).
- Conventional missing data approach
  - McKnight et al. (2007) :Mean imputation
  - Mcdonald et al. (2000) :Listwise deletion
  - Lee and Carlin (2010) :Multiple imputation
- The conventional method is less accurate and lead to bias estimation.

RELATED WORKS

- Artificial neural networks (ANNs) have become one of the most promising tools for treating missing values in water resource engineering .
  - Levenberg-Marquadt back propagation algorithm
  - Gaussian mixture model-based K-nearest neighbor (GMM-KNN) algorithm
  - Bayesian principal component analysis (BPCA)
  - Adaptive neuro-fuzzy inference system (ANFIS)
- Much research has focused on rainfall data imputation.
- However, the compatibility of precipitation (rainfall) and non-precipitation (meteorology) as input data has received less attention.

The proposed methodology of missing data imputation



MISSING DATA IMPUTATION

- Propose an improved neural network imputation: Sine Cosine Algorithm Neural Network (SCANN) Imputation.
- Sine cosine algorithm (SCA) is a metaheuristic optimization technique introduced by Mirjalili (2016) to solve continuous optimization problems.
- SCA is employed to optimize neural network for infilling the missing rainfall data series.

SINE COSINE ALGORITHM (SCA)

- The SCA updates the best solutions obtained and denotes it as a destination point, P.

$$X_i^{t+1} = \begin{cases} X_i^t + r1 \times \sin(r_2) \times \left| r_3 P_i^t - X_{i,}^t \right|, & r_4 < 0.5 \\ X_i^t + r1 \times \cos(r_2) \times \left| r_3 P_i^t - X_{i,}^t \right|, & r_4 \geq 0.5 \end{cases} \quad \text{Eq (7)}$$

where X_i is the position vector of the current solution in the i^{th} dimension, t is the current iteration, P_i is the destination solution and r_1, r_2, r_3, r_4 are random variables; the r_4 value is between 0 and 1.

SINE COSINE ALGORITHM (SCA)

- The parameter r1 is the movement direction parameter that determines the region of the next solution

$$r_1(t) = a \times \left(1 - \frac{t}{t_{max}}\right) \qquad \text{Eq (8)}$$

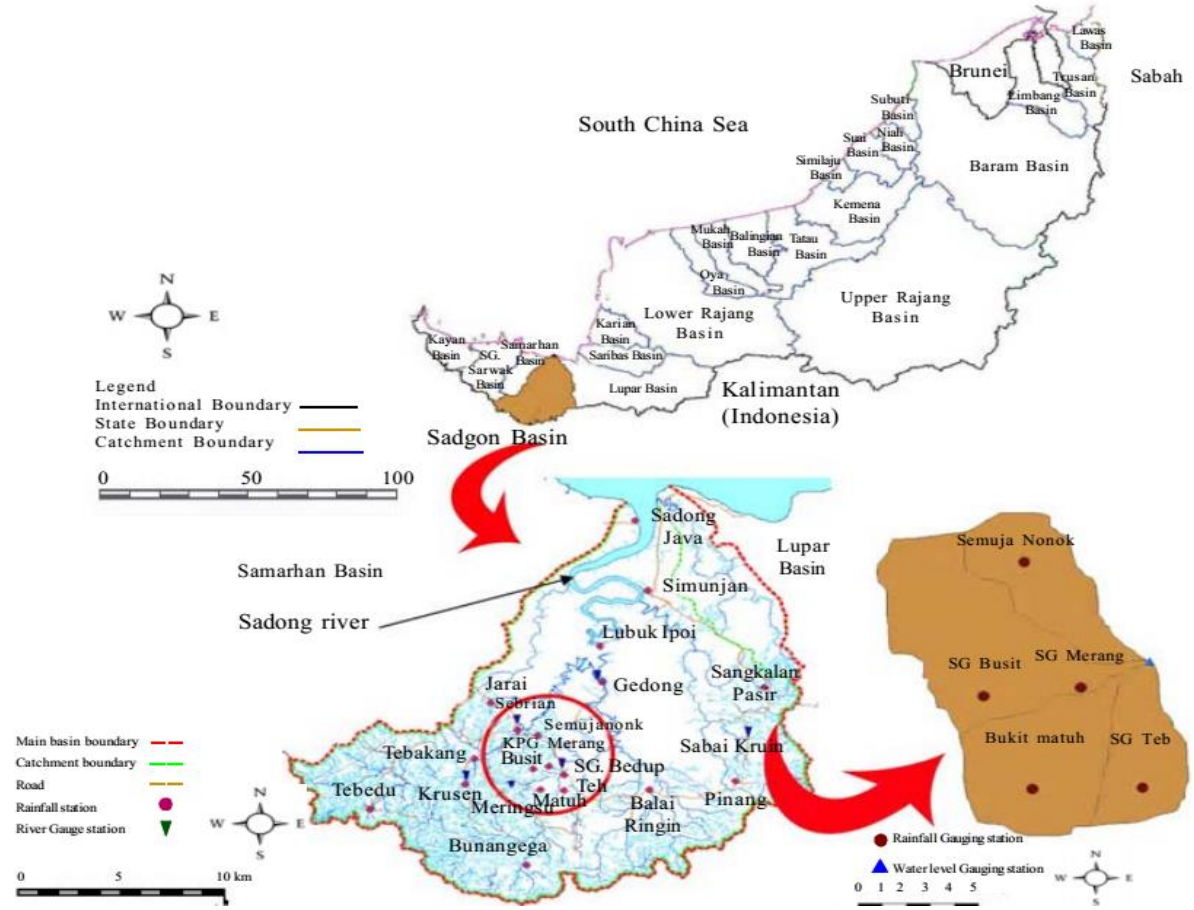
where t is the current iteration, t_{max} is the maximum iteration of SCA and a is a constant (a =2).

Algorithm 1: The proposed Sine Cosine Algorithm Neural Network (SCANN) imputation

Begin
Do
Load training dataset
Initialize SCANN parameters as in Table 2
Initialize a set of random search agents (solutions) (X) and SCA parameters (r_1, r_2, r_3, r_4)
    Do
        Evaluate each of the search agents by the objective function
        Update the best solution obtained so far (P)
        Update the parameters r_1, r_2, r_3, and r_4
        Update the position of search agents using Equation (7)
    While (t < maximum number of iterations)
    Return the best solution (P) obtained as the global optimum solution
Track the best network into net
Update training state into net
While (MSE > the minimum error or E < maximum number of epochs)
Use the optimized net
Train the optimized net for another dataset of the same format
Output: Estimated missing rainfall data
Do
    Impute the estimated values into the missing value
While (there is missing value)
End

where X is the search solutions, P is the destination solution, r_1 is the movement direction parameter, r_2 identifies the movement of forwards or outwards P within the value of 0 and 2π, r_3 is the random weights of P (value less than 1 or greater than 1) and r_4 is the random variables ($0 < r_4 < 1$).

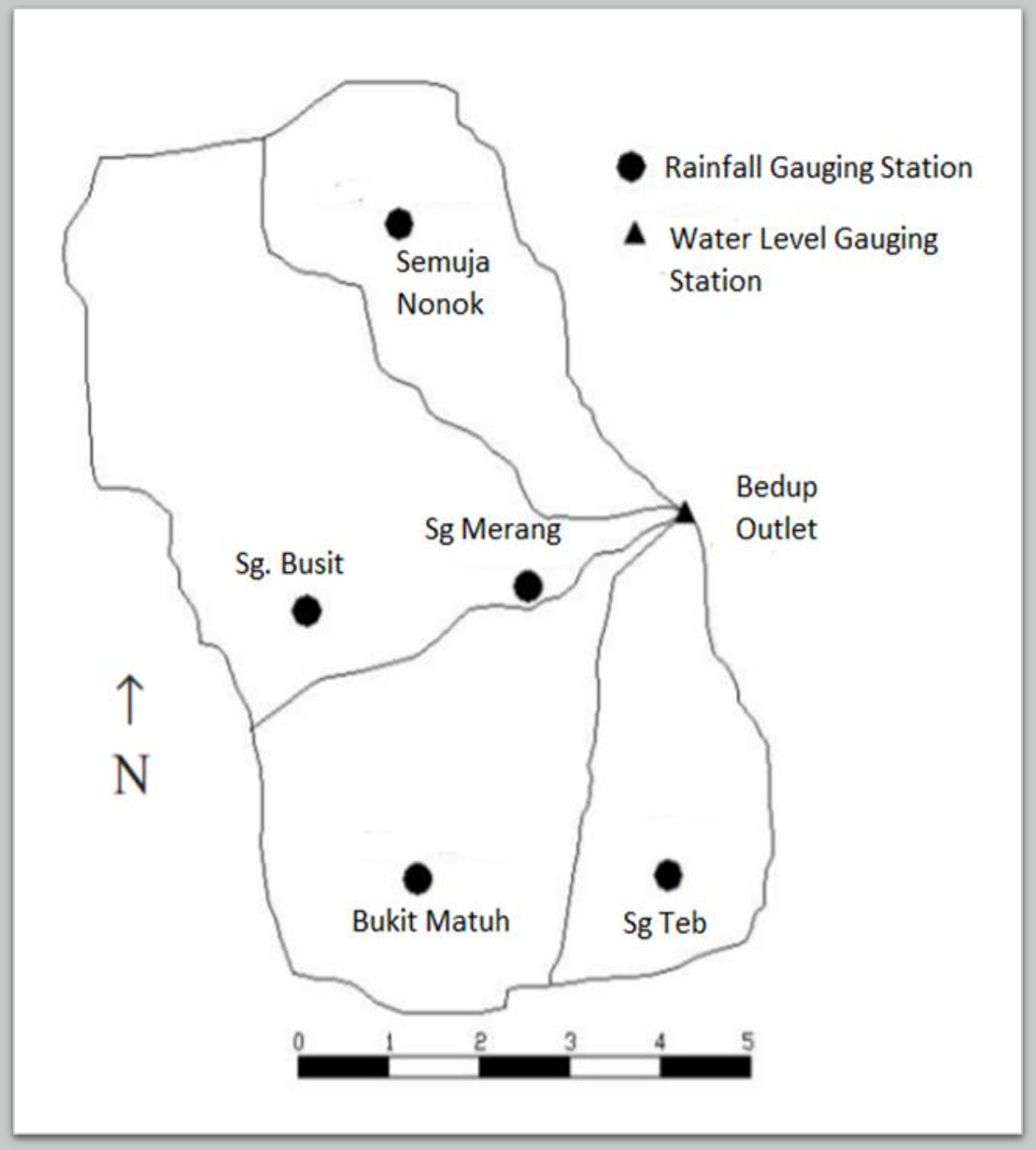
STUDY AREA



Sungai Merang station and its nearest gauging stations, Sarawak

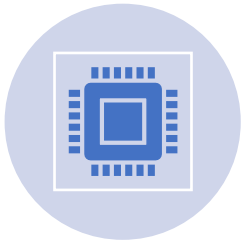
DATA

- The 24-month records of **hourly rainfall** datasets from rainfall gauging stations in Sarawak, Malaysia.
- **Meteorological data:** from the Malaysian Meteorological Department.
- **Rainfall data:** from Department of Irrigation and Drainage, Sarawak, Malaysia.



PRE-PROCESSING INPUT DATA

Principal component analysis (PCA) is used to extract the most relevant features from the meteorological data.



Output:

PC1, PC2,…, PCs

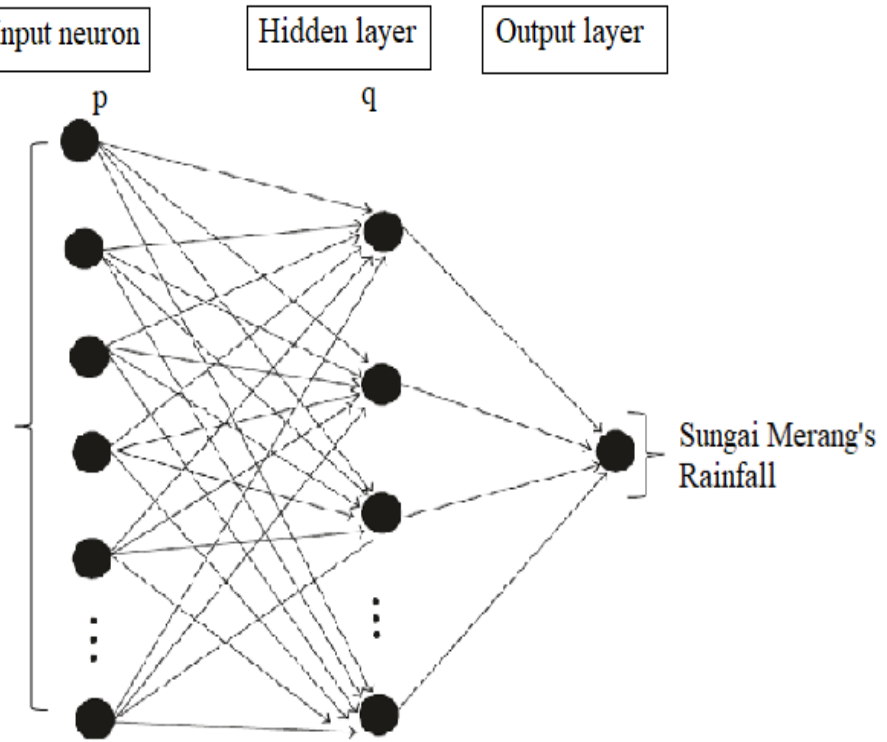
INPUT DATA

p1 = PC1 + Rainfall from nearest neigbhor station

p2 = (Cumulative PC1 + +PC2) Rainfall from nearest neigbhor station

p3 = (Cumulative PC1 + +PC2+PC3) Rainfall from nearest neigbhor station

p10 = (Cumulative PC1 +PC2+PC3+PC4 +PC5+ ...+PC10) + Rainfall from nearest neigbhor station



DATASET PREPARATION

- Rate-based schema (Oba, 2003): randomly remove a specific proportion of data from the dataset.
- Prepare two sets of missing datasets. Each set contains five missing rates: **10%, 20%, 30%, 40%,** and **50%**.
- This study falls under Missing Completely at Random (MCAR).
- The missing data in one gauging station do not influence by the other data in any gauging stations.

EXPERIMENT

- The proposed SCANN missing data imputation was compared with the FFNN missing data imputation.
- For each missing dataset, we execute 30 independent runs over each input p at different missing data rates.

PERFORMANCE MEASURES

Mean Absolute Error (MAE)
Provides the average error in the treated datasets

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|O_i - T_i|$$

Root Mean Square Error (RMSE)
Calculates the average square errors of the treated datasets

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

Correlation Coefficient (R)
Calculates the average square errors of the treated datasets

$$R = \frac{\sum(T - \bar{T})(O - \bar{O})}{\sqrt{\sum(T - \bar{T})^2 \sum(O - \bar{O})^2}}$$

N = total number of observations, O = actual values of observation
T = imputed values

Comparison of SCANN imputation and FFNN imputation

Input P	SCANN						FFNN					
	MAE (mm)						MAE (mm)					
	10%	20%	30%	40%	50%	Avg	10%	20%	30%	40%	50%	Avg
P1	**0.0718**	**0.1345**	**0.1626**	**0.2436**	**0.3010**	**0.1827**	0.1126	0.2169	0.2754	0.3981	0.4970	0.3000
P2	0.0866	0.1649	0.2137	0.3096	0.3878	0.2325	0.1198	0.2370	0.3155	0.4446	0.5724	0.3379
P3	0.1056	0.2102	0.2798	0.3998	0.5029	0.2996	0.1336	0.2711	0.3623	0.5138	0.6465	0.3855
P4	0.1161	0.2277	0.2962	0.4265	0.5293	0.3192	**0.1078**	**0.2124**	**0.2798**	**0.3970**	**0.4969**	**0.2988**
P5	0.1145	0.2211	0.2977	0.4198	0.5275	0.3161	0.1475	0.2862	0.3870	0.5329	0.6862	0.4080
P6	0.1133	0.2199	0.2961	0.4201	0.5224	0.3144	0.1526	0.3163	0.4244	0.5734	0.7624	0.4458
P7	0.1216	0.2411	0.3200	0.4451	0.5622	0.3380	0.1503	0.2976	0.4055	0.5705	0.7079	0.4264
P8	0.1420	0.2788	0.3814	0.5367	0.6647	0.4007	0.1478	0.2897	0.3909	0.5485	0.6897	0.4133
P9	0.1282	0.2496	0.3340	0.4684	0.5887	0.3538	0.1643	0.3451	0.4517	0.6104	0.7860	0.4715
P10	0.1232	0.2415	0.3219	0.4535	0.5679	0.3416	0.1662	0.3130	0.4244	0.5882	0.7424	0.4468

Note: The best results obtained are made bold.

Comparison of SCANN imputation and FFNN imputation

Input P	SCANN						FFNN					
	RMSE (mm)						RMSE (mm)					
	10%	20%	30%	40%	50%	Avg	10%	20%	30%	40%	50%	Avg
P1	0.8309	**1.1076**	**0.9515**	**1.3441**	**1.4733**	**1.1415**	1.0294	1.4542	1.3375	1.7405	1.9763	1.5076
P2	**0.8296**	1.1081	1.0136	1.3754	1.5341	1.1722	0.9943	1.5871	1.4953	1.8104	2.5168	1.6808
P3	0.8874	1.2137	1.1264	1.5138	1.7254	1.2933	1.0425	1.6286	1.5146	2.0166	2.3504	1.7106
P4	0.9571	1.3050	1.1400	1.5457	1.7385	1.3373	**0.8706**	**1.2102**	**1.1029**	**1.4353**	**1.6376**	**1.2513**
P5	0.8765	1.1700	1.0636	1.4376	1.6171	1.2330	1.1862	1.7574	1.6776	1.8868	2.4270	1.7870
P6	0.8687	1.1576	1.0384	1.4248	1.5910	1.2161	1.1599	2.6579	2.5854	2.3485	3.9494	2.5402
P7	0.9720	1.6611	1.5517	1.5820	2.1131	1.5760	1.0032	1.4244	1.3650	1.7236	1.9201	1.4872
P8	1.0190	1.4121	1.3236	1.7326	1.9196	1.4814	1.0702	1.5126	1.3925	1.7734	2.0606	1.5618
P9	0.9328	1.2399	1.1273	1.5015	1.6951	1.2993	1.2187	2.8911	2.7826	2.2853	3.5474	2.5450
P10	0.9092	1.1791	1.0882	1.4686	1.6404	1.2571	1.0564	1.4835	1.3806	1.7666	2.0243	1.5423

Note: The best results obtained are made bold.

Comparison of SCANN imputation and FFNN imputation

Input P	SCANN						FFNN					
	R						R					
	10%	20%	30%	40%	50%	AVG	10%	20%	30%	40%	50%	AVG
P1	0.9510	0.9124	**0.9360**	**0.8739**	**0.8420**	**0.9031**	0.9200	0.8342	0.8628	0.7891	0.7109	0.8234
P2	**0.9518**	**0.9139**	0.9295	0.8700	0.8371	0.9005	0.9289	0.8610	0.8816	0.8064	0.7613	0.8478
P3	0.9439	0.8973	0.9134	0.8488	0.8117	0.8830	0.9253	0.8553	0.8721	0.8025	0.7543	0.8419
P4	0.9334	0.8746	0.9052	0.8320	0.7850	0.8660	**0.9467**	**0.8961**	**0.9140**	**0.8558**	**0.8080**	**0.8841**
P5	0.9462	0.9023	0.9203	0.8554	0.8117	0.8872	0.9185	0.8645	0.8837	0.8177	0.7735	0.8516
P6	0.9467	0.9038	0.9245	0.8565	0.8148	0.8893	0.9200	0.8623	0.8769	0.8120	0.7642	0.8471
P7	0.9298	0.8632	0.8824	0.8261	0.7670	0.8537	0.9268	0.8554	0.8705	0.8079	0.7589	0.8439
P8	0.9289	0.8764	0.8968	0.8289	0.7858	0.8634	0.9169	0.8406	0.8681	0.7916	0.7233	0.8281
P9	0.9386	0.8885	0.9091	0.8408	0.7878	0.8730	0.8967	0.7946	0.8196	0.7522	0.6801	0.7886
P10	0.9421	0.8973	0.9170	0.8493	0.8087	0.8829	0.9189	0.8472	0.8675	0.7981	0.7266	0.8317

Note: The best results obtained are made bold.

RESULT AND DISCUSSION

The proposed SCANN imputation achieved an average accuracy of 90%, compared to FFNN ‘s 88% of accuracy.

The SCANN imputation has lower MAE and RMSE values for all missing rates than the FFNN imputation.

The proposed SCANN imputation method outperformed FFNN imputation in treating the missing values in the dataset.

CONCLUSION

- This study investigated the potential of using the combination input of significant PCs and rainfall data from nearest neighbor gauging stations for infilling missing rainfall data.
- This finding suggests the use of significant PCs values and nearest neighbor station variables that have high correlation coefficients as the input to the missing rainfall data imputation.

CONCLUSION

- The proposed SCANN imputation achieved an average accuracy of more than 90%.
- The proposed SCANN imputation has a higher capability in treating missing values in the dataset compared to FFNN imputation in terms of MAE, RMSE, and R.

FUTURE WORK

- To compare the SCANN imputation method with the existing state-of-art metaheuristic algorithms.
- To increase the number of neighboring stations in order to estimate better results for missing rainfall data.

REFERENCES

- Chiu, P.C., Selamat, A., Krejcar, O. Infilling Missing Rainfall and Runoff Data for Sarawak, Malaysia Using Gaussian Mixture Model Based K-Nearest Neighbor Imputation. In International Conference on Industrial, Engineering and Other Applications of Ap-plied Intelligent Systems, 27-38. Springer, Cham. (2019).
- Eskelson, B.N., Temesgen, H., Lemay, V., Barrett, T.M., Crookston, N.L., Hudak, A.T.: The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scand. J. For. Res. 24(3), 235–246 (2009).
- Lai, W.Y., Kuok, K.K. A Study on Bayesian Principal Component Analysis for Ad-dressing Missing Rainfall Data. Water Resources Management, 1-14 (2019).
- Lee, K.J., Carlin, J.B. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. American journal of epidemiology, 171(5), 624-632 (2010).
- Little RJ, Rubin DB. Statistical analysis with missing data. John Wiley & Sons (2014).
- Mcdonald, R.A., Thurston, P.W., Nelson, M.R.A Monte Carlo study of missing item methods. Organizational Research Methods, 3(1), 71-92 (2000).
- McKnight, P.E., McKnight, K.M., Sidani, S., Figueredo, A.J. Missing data: A gentle introduction. Guilford Press (2007).
- Mirjalili, S. SCA: A sine cosine algorithm for solving optimization problems. Knowledge-Based Systems, 96, 120-133 (2016).
- Mispan, M.R., Rahman, N.F.A., Ali, M.F., Khalid, K., Bakar, M.H.A., Haron, S.H. Missing River Discharge Data Imputation Approach using Artificial Neural Network. Methodology, 25, 20 (2015).

THANK YOU