Outlier Detection of Reservoir Water Level Data Using Artificial Neural Network Model

Maga Kim, Jin-Yong Choi †
Department of Rural Systems Engineering, College of Agriculture and Life Sciences, Seoul National University
Seoul, Republic of Korea
iamchoi@snu.ac.kr

Authors Name/s per 2nd Affiliation (*Author*)
line 1 (of *Affiliation*): dept. name of organization
line 2-name of organization, acronyms acceptable
line 3-City, Country
line 4-e-mail address if desired

The agricultural reservoirs determine the amount of water supply of irrigation according to water and environmental conditions of the reservoir. The reservoir water level data estimate the current water storage of the reservoir by capacity curve, to figure out the ability for irrigation and to manage agricultural water reasonably. In Korea, pieces of reservoir water level measuring equipment are installed for agricultural reservoirs having 100,000 tons storage capacity or more, and reservoir water levels are measured every 10 minutes. In spite of vast amount of available reservoir water level data, outlier detection systems for measured data is not properly equipped. The manual outlier detection and quality control requires time and labor consuming, and outliers and missing values create problematic causes in utilization of the reservoir water level data for irrigation planning appropriately. Therefore, it is necessary to detect outlier and improve the quality of reservoir water level data. This study was conducted to detect outliers of reservoir water level data using artificial neural network model. The artificial neural network model was trained with prepared training dataset as normal data (T) and outlier or missing data (F), and the artificial neural network model operated for identifying the outlier. The models are evaluated with reference reservoir water level data which were collected in daily by Korea Rural Community Corporation (KRC).

Keywords—outlier detection, artificial neural network; reservoir water level

I. INTRODUCTION

Reservoirs are artificial irrigation facilities which store the water stream (river, flowing water). It can be possible to supply water to where requiring water at when requiring water by efficient management of reservoirs (Jung and Kim, 2007). In particular, the reservoir water level data can be used to check the reservoir storage, and the reservoir water level data is used as a criterion for determining the reservoir storage according to purpose. In addition, reservoir water level data are used for various research fields such as optimization of hydrological model (Song et al., 2017) and groundwater flow analysis (Ji, 2014). In Korea, about 1,600 reservoir water level measurement equipment are installed for agricultural reservoirs with a capacity of 100,000 tons or more, and the water level data are collected and utilized for reservoir operation and research.

The water level measurement method uses a pressure type sensor and an ultrasonic type sensor and the reservoir water level are measured at intervals of 10 minutes (Bang et al., 2017). However, in the case of the pressure type sensor, the incorrectly measured values may occur due to the inflow of the internal sludge in inclined conduit or the accumulation of the deposit near the sensor. In the case of the ultrasonic sensor, the incorrectly measured values may occur due to the temperature and humidity, wave of water and plants near the sensor. In addition, there are causes that may occur the incorrectly measured values such as equipment error, calibration error, and so on. Although the incorrectly measured values are existed, there is not proper outlier detection method yet. The administrator of reservoir conduct outlier detection in subjective and manual way at present. Because manual outlier detection method takes a lot of time and labor, it is difficult to detect the incorrectly measured values and it makes utilization of reservoir water level to decrease.

Therefore in this study, we try to apply outlier detect algorithm to make automatic outlier detect possible and evaluate applicability. The methods employing in this study is artificial neural network (ANN) model. The ANN model make training data set with statistical method and train the model with training data set. Then ANN model detect the incorrectly measured values. The results of the models were evaluated with reference data which are daily reservoir water level offered from Korea Rural Community Corporation (KRC).

II. MATERIALS & METHODOLOGY

In this study, the reservoir water level data measured at 10-minute interval are used as raw data for outlier detection and daily reservoir water level data from KRC are used as reference data. The ANN model is applied to detect the incorrectly measured values. The results of the models are evaluated by comparison with the daily reservoir water level data from KRC. The reservoir water level data from KRC are measured daily and the administrator check the value. If there is problem with values, the administrator can change the value to correct value by eye observation. Fig. 1 is flow chart of this study.

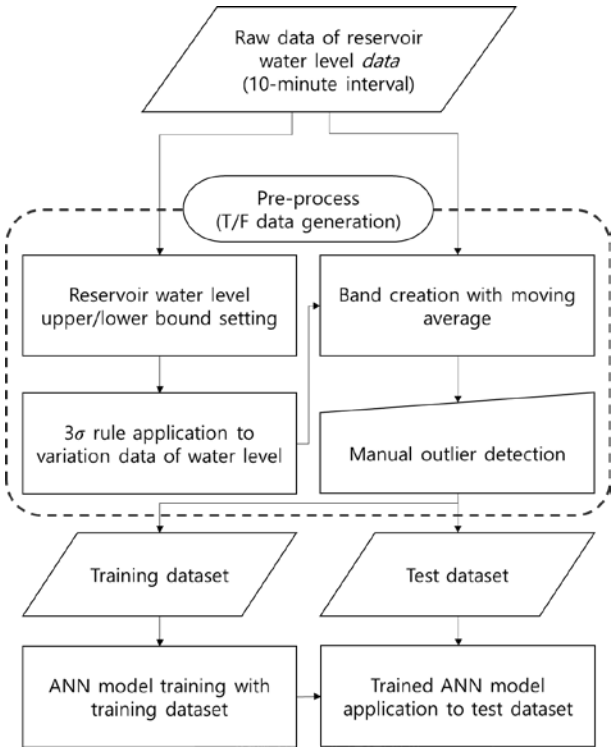


Fig. 1. Flow chart of the study

A. *Subject reservoir and water level data*

The subject reservoir of this study is the Gaeun reservoir which is in Gaeun-ri, Dong-myeon, Hongcheon-gun, Gangwon-do, Korea. The Gaeun reservoir has 474 ha of basin area and 1,636 thousand ton of effective storage. Table 1 is the properties of the Gaeun reservoir.

TABLE I. PROPERTIES OF THE GAEUN RESERVOIR

Dam properties	Contents	Dam properties	Contents
Dam type	fill dam	Intake works	main facility
Dam length (m)	224	Water-intake type	intake tower
Dam height (m)	39.7	Total storage (㎥)	1,649,375
Dead storage water level (m)	22.50	Effective storage (㎥)	1,636,475
Full water level (m)	245.80	Benefitted area (㎢)	103
Flood water level (m)	246.80	Basin area (ha)	474

The reservoir water level data of the Gaeun reservoir are used as raw data for outlier detection model, and daily reservoir water level data from KRC are us ed as reference data. The reservoir water level data of the Gaeun reservoir are measured by pressure type sensor in 10 minute interval and the period of data is 2011. 01. 01. 0:00~2018. 06. 12. 11:40. The reference data are measured by sensor first, and then adjusted by manager of the reservoir with eye measurement. The interval of measuring the reference data is a day and the period of the reference data is 2009. 08. 19.~2018. 05. 23. Fig. 2 is the reservoir water level data of the Gaeun reservoir.

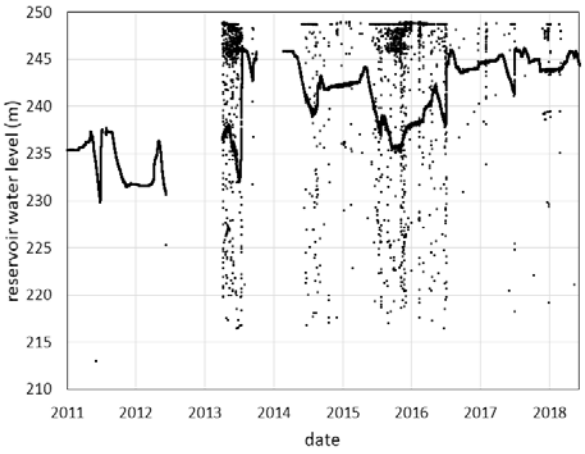


Fig. 2. 10-minute interval raw water level data of the Gaeun reservoir

B. *Artificial neural network (ANN) model*

The ANN model make the T/F dataset first, and then divide the dataset to training dataset for model training and test data for testing the model. T/F means normal water level data (T) and outlier or missing data (F). After the training, ANN model classify the reservoir water level data by using test data. Making the dataset is divided into four steps which are establishing upper/lower bound of reservoir water level, applying 3-sigma rule of thumb to variation data of water level, creating a band from reference value with moving average, and manual outlier detecting. When the reservoir water level value is classified as the outlier in the process of making training data set, the model exclude the value at the time and apply the result to next step..

1) Setting upper/lower bound of reservoir water level (step 1)

In first step, the reservoir water level data which are outside the bound are classified as the incorrectly measures values. The bound of reservoir water level data is determined with reservoir specification. The upper bound is established same as flood water level and the lower bound is established same as dead water level. When the reservoir water level value is classified as the incorrectly measured values in the process of making training data set, the model exclude the value at the time and the result are employed in second step.

2) Applying 3-sigam rule to variation data of water level (step 2)

In second step, the reservoir water level data are classified by applying 3-sigma rule of thumb to variation data of water

level. 3-sigma rule of thumb establish the bound using mean(m) and standard deviation(σ). If the variation data of water level is out of bounds(m ±3 σ), the reservoir water level data at the time are classified as the incorrectly measured values. In this step, if the reservoir water level value of previous point is missing value, the variation value of water level at the time is considered as zero. When the reservoir water level value is classified as the incorrectly measured values in this step, the model exclude the value at the time and the result are employed in third step.

3) Creating a band with reference vaule from moving average (step 3)

In third step, the reference value is determined by moving average from result of second step and then the raw reservoir water level data which are outside the band are classified as the incorrectly measured values. The moving average is calculated by using the previous 10 reservoir water level data including the value at the time and the constant width of band is 0.05 m.

4) Manual outlier detecting (step 4)

In fourth step, the misclassified data are reclassified as normal data or outlier. When there is sharp changes in variation data of water level, the normal data can have been misclassified as outlier (F). Thus, if there are F values for over 4 hours in a row, the classifications of reservoir water data were modified manually after checking. Also, some of the outlier can have been misclassified as normal data. They were appeared like spike noise in the graph. In that case, the misclassified outlier are reclassified manually. Fig. 3 is the dataset for ANN model which is eliminating the outliers (F).

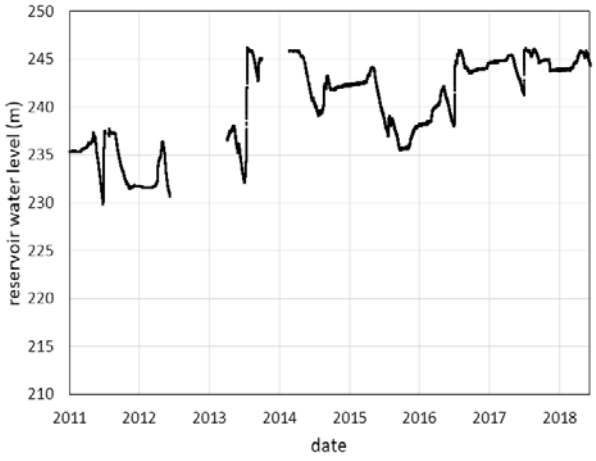


Fig. 3. T dataset after data pre-process

The ANN model is constructed with an input layer, one hidden layer, and an output layer. It applies sigmoid function as activation function for hidden layer, softmax function as activation function for output layer. The ANN model calculates error at output layer with cross entropy error function and adjusts the weights with the gradient descent method which is one of the back-propagation algorithm. The performance of the ANN model is affected by input data type which are demonstrated in Table 2.

TABLE II. INPUT DATA OF THE ARTIFICIAL NEURAL NETWORK MODEL FOR OUTLIER DETECTION

Data type	Contents
water level (wl_t)	wl_t
variation (v_t)	$v_t = wl_t - wl_{t-1}$
reservoir regularization (rr_t)	$rr_t = (wl_t - dl) / (fl - dl)$
limited regularization (lr_t)	lr_t = 1(if $rr_t > 1$) rr (if $0 < rr_t < 1$) -1 (if $rr_t < 0$)
regularization (r_t)	$r_t = (wl_t - \min(wl)) / (\max(wl) - \min(wl))$
variation regularization (vr_t)	$vr_t = (v_t - \min(v)) / (\max(v) - \min(v))$

*wl: measured reservoir water level, v: variation of reservoir water level, rr: regularized reservoir water level with dead water level (dl) and flood water level (fl), lr: regularized reservoir water level with dead water level (dl) and flood water level (fl) within range of -1 to 1, r: regularized reservoir water level, vr: regularized variation of reservoir water level

The hyper-parameters of the ANN model such as number of past-time input data, number of hidden layer nodes, and learning rate are determined by trial and error method. According to results of the trial and error method, number of past-time input data is 13, number of hidden layer nodes is 15, and learning rate is 0.005. Fig. 4 is the structure of the ANN model.

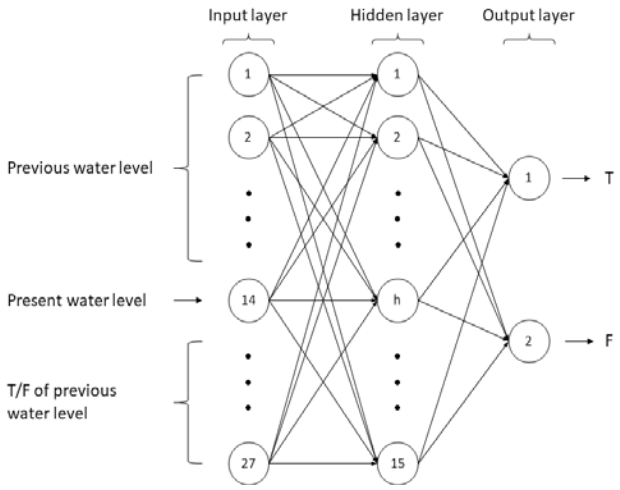
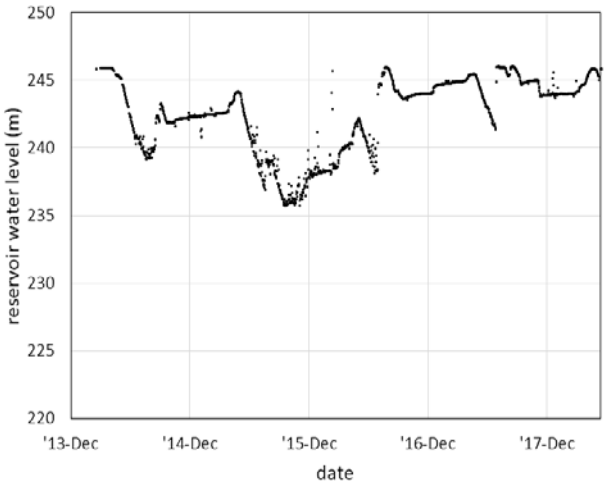


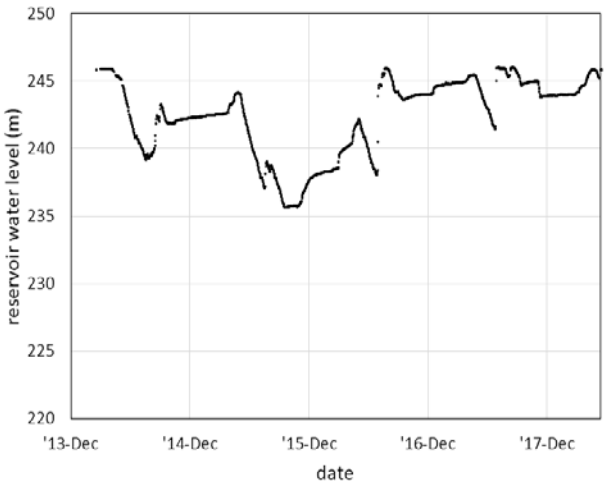
Fig. 4. The subject of the ANN model for detecting outlier

III. RESULTE

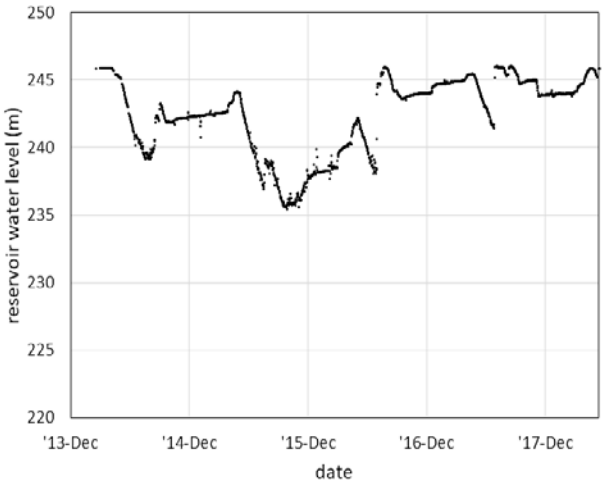
The ANN model applied 6 input data type such as water level (wl), variation (v), reservoir regularization (rr), limited regularization (lr), regularization (r), variation regularization (vr) and the result of the case of applying variation (v) as input data showed the best performance among the 6 input data type. Therefore, in this study, variation (v) data is used as input data. Daily mean data are used to compare the result of the model and reference data. Fig 5 is the graph of the (a) raw data, (b) target data of the ANN model, (c) result of the ANN model, (d) reference data.



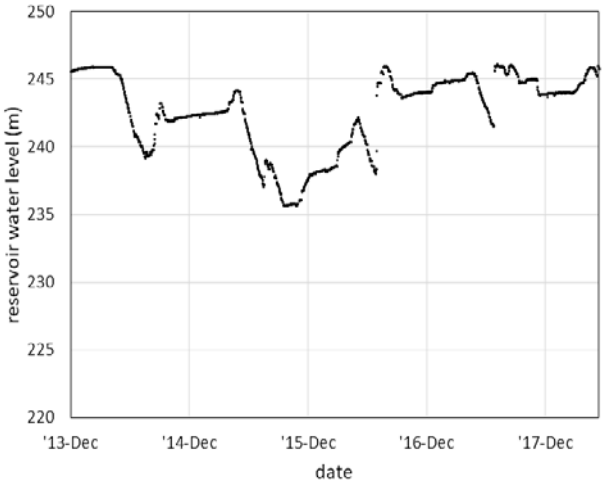
(a) raw data



(b) target data



(c) ANN model result data



(d) reference data

Fig. 5. Daily mean value of (a) raw data, (b) target data, (c) ANN model result data, (d) reference data after ANN model application

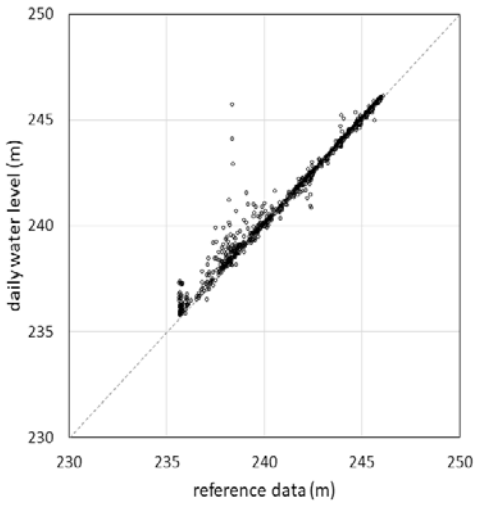
In Fig. 5, (c) ANN data showed improved result over raw data but it fall short of target data.

R^2, MAE, RMSE compared to reference data are in Table 3. R^2 of target data is 0.999 while R^2 of ANN data is 0.997. It means the ANN model detect the outlier and provide improved result over raw data, but it fall short of target data which has 0.999 of R^2.

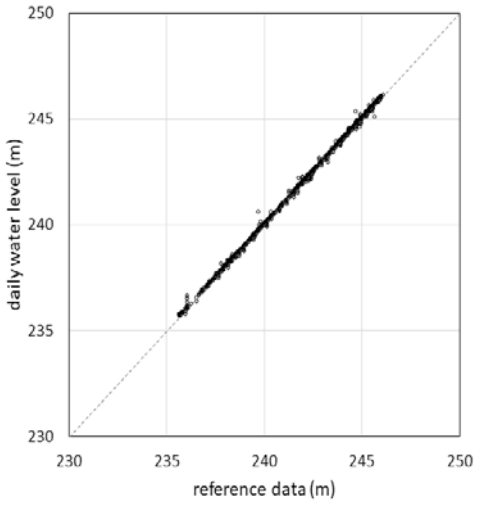
TABLE III. TABLE 1 THE STATISTICAL PARAMETERS (R^2, MAE, RMSE) COMPARED TO REFERENCE DATA

Statistical parameters	Raw data	Target data	ANN data
R^2	0.982	0.999	0.997
MAE	0.126	0.034	0.065
RMSE	0.396	0.067	0.149

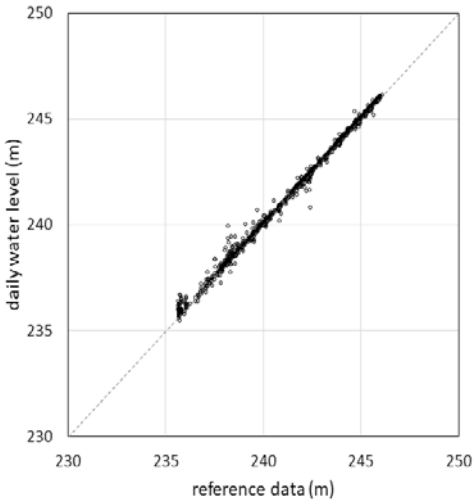
Fig. 6 is the scatter plot of the result compared with reference data.



(a) raw data



(b) target data



(c) ANN model result data

Fig. 6. The scatter plot of (a) raw data, (b) target data, (c) ANN model result data compared with reference data

REFERENCES

[1] J. Bang, Y. Lee, S. Jung, and J. Choi, "A study of outlier detection on time series of water level in agricultural reservoir", Korean Society of Agricultural Engineers Annual Conference, October 2017.

[2] W. Ji, "Analysis of groundwater flow in the reservoir water elevation" Keimyung University.

[3] G. Jung and T. Kim, "Comparison of water distribution model through reservoir and water system operation", Water for Future, vol. 41, pp. 38-43.

[4] J. Song, M. Kang, K. Kim and J. Ryu, "Estimation of daily reservoir inflow from water level observation using a hydrological model and an optimization method", Korean Society of Agricultural Engineers Annual conference, October 2017.