Journal of Physics: Conference Series

PAPER • OPEN ACCESS

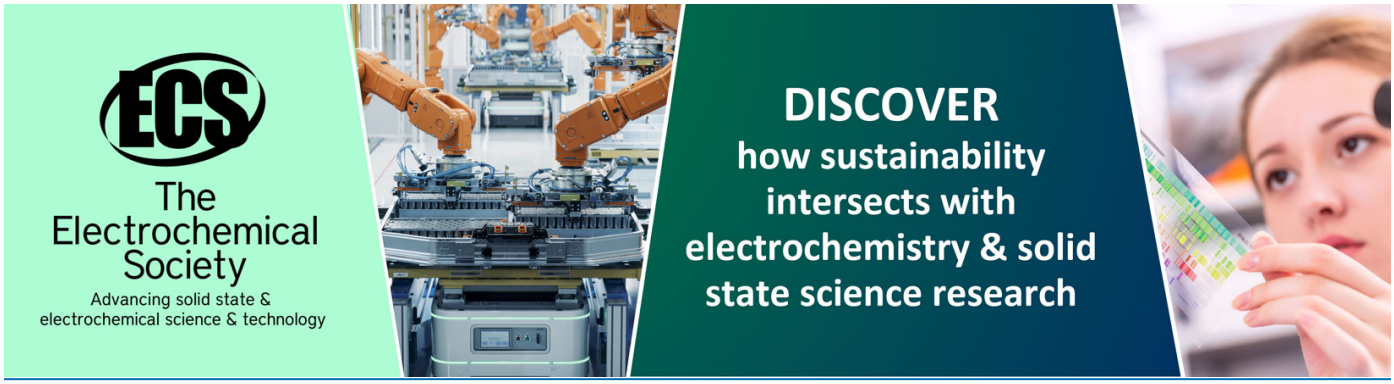
Correlation-Based LR Determination of Abnormal Data Collected with Rain Gauges at Remote Measurement Stations

To cite this article: Jian Xu and Yuan Chen 2021 *J. Phys.: Conf. Ser.* **2026** 012031

View the article online for updates and enhancements.

You may also like

- The Reliability of X-Band Multiparameter Radar Rainfall Estimates
Q R Fajriani, R Jayadi, D Legono et al.
- Bartlett Lewis Rectangular Pulse (BLRP) Approach with Proportional Adjusting Procedure in Rainfall Disagregation Method in Hidrology Laboratory of Brawijaya University Rain Station
Novita Putri Kurnia Dewi and Suci Astutik
- TRMM PR Observed Spatial Patterns of The Convective-Stratiform Rainfall Over Indonesia and Their Response to ENSO
Abd. Rahman As-syakur



This content was downloaded from IP address 18.118.95.222 on 14/05/2024 at 14:17

Correlation-Based LR Determination of Abnormal Data Collected with Rain Gauges at Remote Measurement Stations

Jian Xu and Yuan Chen

CHN Energy Dadu River Big Data Services CO., LTD, Chengdu, China
Email: xujian134@163.com; yuan.chen.ag@chnenergy.com.cn

Abstract. An analysis and evaluation model combining Pearson correlation and logistic regression with linear regression (LR) was proposed in this paper, to screen the remote measurement stations meeting the threshold requirements by Pearson correlation analysis based on historical rainfall of the related remote measurement stations; and the Pearson correlation coefficient was used to analyze the linear mapping relationship between the target remote measurement stations and the surrounding ones; finally, a data reliability determination model was finally constructed through analyzing the difference between the forecasted rainfall and measured rainfall. With the subjects of rain gauges, and based on the historical rainfall at 88 remote measurement stations (39 hydrologic stations, 49 rainfall stations) in the Dadu River Basin, the rainfall determination model and small rainfall identification model were respectively constructed according to the features of measurement points for rainfall correlation at different remote measurement stations. The experimental results showed that the logistic regression model performed well during the inspection period, with the average accuracy of positive and negative samples of 0.92, the average recall rate of 0.91, and the average F1_Score of 0.91; while in the small rainfall identification model, the linear regression R^2 was 0.927, and the accuracy rate was 0.60. The model provided a preliminary method for identifying abnormal rainfall, which can reduce artificial misjudgment and greatly improve the identification efficiency of abnormal rainfall.
Keywords. Rainfall forecast; rain gauges; logistic regression; linear regression.

1. Introduction

The automatic hydrologic remote measurement system in the Dadu River Basin is a comprehensive application system for automatic collection, transmission, and storage of hydrologic and rainfall data at specific locations and important sections of the Dadu River Basin with the sensor technology, communication technology and computer technology. This system can provide timely and accurate information for the production command center to improve its management level. Currently, there are 96 remote measurement stations in the Dadu River Basin, including 39 hydrologic stations, 8 gauging stations, and 49 rainfall stations [1].

As for the automatic hydrologic remote measurement system, the normal operation rate is a standard for measuring the data availability in actual operation. As required by intelligent scheduling in the Dadu River Basin, the annual normal operation rate of the system should be ≥99%. As shown in the survey, the annual average normal operation rate of the horizontal telemetry system from 2011 to 2019 was 98.2%, which was lower than the target requirement. The statistics data of faults of the hydrologic remote measurement system showed that among various remote measurement devices, rain gauges often resulted in records of abnormal rainfall or even missing records due to malfunctions or inadequate maintenance, thus affecting the quality of rainfall data. Due to the greatest number of

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

rainfall stations and the lowest frequency of reporting, the failures of rain gauges would be difficult to find in time, making a high proportion of failures among various devices; furthermore, the failure time of rain gauges would be quite long, and have the greatest impact on the normal operation rate of the related systems.

In respect of the issues above, the production command center assumed that if the failure duration of the rain gauges could be reduced by 50%, and the failure days by 180 days, the annual average normal operation rate of the remote measurement system would meet the target requirement (99.2%).

Based on what assumed by the production command center, the project team constructed a rain gauge data reliability analysis model for remote measurement stations in the Dadu River Basin focusing on improving the normal operation rate and identifying abnormal rainfall. This model involves 88 remote measurement stations (39 hydrologic stations, and 49 rainfall stations) in the Dadu River Basin, and it can be used to analyze the reliability rainfall data, and provide early warning regarding the gauges with possible failure.

Pearson correlation coefficient [2], as one of the most important indicators of linear relationship, has been widely used in the analysis of the strength of linear relationship between different things or characteristics. In recent years, logistic regression has been used for classification and identification, and by virtue of strong sparse applicability, strong linear classification ability, and fast convergence speed, it has been successfully applied to speed recognition, recommended classification, and spam mail identification. Furthermore, it can also be used to fit the measured rainfall data with the help of the traditional linear regression least square method. In this paper, a model combining Pearson correlation and logistic regression with linear regression (LR) was proposed, which, together with the linear regression model, can scientifically analyze the linear mapping relationship between rainfall and target rainfall in different regions based on the historical rainfall, and provide high-efficiency services for improving the forecasting accuracy and abnormal elimination based on the potential rainfall data [3].

2. Introduction of the Algorithm

2.1. Correlation Coefficient

Correlation coefficient, as a statistical indicator designed by the statistician Pearson, is an indicator for studying the linear correlation between variables, which is generally represented by "r". For different subjects, the correlation coefficient is defined in multiple ways, and the commonly used one is Pearson correlation coefficient. As a statistical indicator for reflecting the degree of correlation between variables, the correlation coefficient is calculated by the product moment method, which can be used to reflect the degree of correlation between the two variables by multiplying two deviations between the two variables and their average values. It is a single correlation coefficient focusing on linearity.

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var[X]Var[Y]}} \quad (1)$$

where, Cov(X, Y) is the covariance of X and Y, Var(X) is the variance of X, and Var(Y) is the variance of Y.

The rainfall-related data collected at 88 remote measurement stations from February 14, 2017 to June 30, 2020 were input in the model, with the time cycle of days. After pre-processing and deleting the missing data, Pearson correlation coefficients between two remote measurement stations were calculated; and the rainfall stations with correlation coefficients >0.5 based on table 1 as a group, to determine the regions with strong correlation.

2.2. Linear Regression

Linear regression is a statistical analysis method for determining the quantitative relationship between two or more than two variables with the regression analysis in mathematical statistics, which has been

widely used. Its expression form is y = wx+e, where, e is normal distribution with the mean error compliance value of 0 [4].

Table 1. Correlation coefficients.

Absolute value of correlation coefficients	Correlation
$0.7 \leq \lvert r \rvert \leq 1$	Extremely strong correlation
$0.5 \leq \lvert r \rvert < 0.7$	Strong correlation
$0.3 \leq \lvert r \rvert < 0.5$	Weak correlation
$0 \leq \lvert r \rvert < 0.3$	Extremely weak correlation or Uncorrelated

In statistics, linear regression refers to the regression analysis for modeling the relationship between one or more independent variables and dependent variables with the least square function of the linear regression equation, which is a linear combination of one or more model parameters (regression coefficients). If there is only one independent variable, it is called simple regression; otherwise, it is called multi-regression with more than one independent variable. (In turn, it should be distinguished by multiple linear regression forecasted by multiple dependent variables, rather than a single scalar variable.) In linear regression, the linear forecasting function is used for modeling, and unknown model parameters are also estimated with the data. These models are called linear models.

Linear regression is the first type of regression analysis that has been rigorously studied and widely used in practice, because the model linearly depending on unknown parameters is easier to be fit than other models, and the estimated statistical properties are easier to be determined. Linear regression models are usually fit with least square approximation, but they may also adopt other methods, such as the minimizing of the penalty of the least square loss function in bridge regression [5].

Generally speaking, the equation of linear regression can be determined by the least square method, which can be used to calculate the straight line "y=ax+b".

Assuming there is x1, x2, …., xk, where k refers to the number of factors, the following linear relation would be determined:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots\ldots + \beta_k x_k + \boldsymbol{\varepsilon}_n \quad (2)$$

where, ε_n is a random variable with the same distribution, meaning uncorrelated.

The final loss function should be minimized; however, due to the complex models with multiple parameters, and multiple optimization algorithms (such as locally weighted linear regression and ridge regression), L2 regularization and stochastic gradient descent [6] were adopted for iterative approximation based on the current mainstream machine learning method and the need of preventing over-fitting.

$$RSS = \sum_{i=1}^{n} (f(x_i) - y_i)^2 \quad (3)$$

2.3. Logistic Regression

Logistic regression [7], as a generalized linear model, has many similarities with multiple linear regression. Their models are basically the same, with wx+b, where, w and b are parameters to be calculated. What different are dependent variables, multiple linear regression directly takes wx+b as the dependent variable (y =wx+b), but logistic regression corresponds wx+b to a hidden state p with the function L (p=L(wx+b)), and then determines the dependent variable according to p and 1-p. If L is a logistic function, it is logistic regression, and L is a polynomial function, it is polynomial regression.

The dependent variable of logistic regression can be binary and multinomial, but the binary variables are more commonly used and easier to interpret; the multinomial variables can be processed by the softmax method. Actually, logistic regression only refers to the use of a logic function based on linear regression, but the logic function has made logistic regression excellent in the field of machine learning, and the core of computational advertising. As for multiple logistic regression, the following equation can be used for

similarity classification. As for the conversion of equation (4), the simplified formula for parameter estimation of the logistic regression model can bring many benefits; y= {0,1} is the classification result. The graph of the logistic regression function is shown below in figure 1.

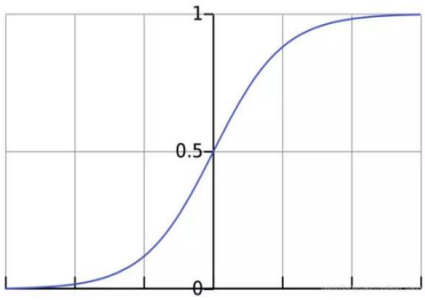


Figure 1. Logistic regression function.

Optimized objective function derivation:

$$p(y = 1|x, \theta) = \frac{1}{1+e^{-\theta^T x}} \tag{4}$$

$$p(y = 0|x, \theta) = 1 - \frac{1}{1+e^{-\theta^T x}} = 1 - p(y = 1|x, \theta) \tag{5}$$

It is the product of multiplying the probabilities of all samples; due to the complexity of continuous multiplication, the logarithms were determined, to change the manner into continuous addition. In order to prevent over-fitting, 2-Norm $|\theta|_2$ was used as the penalty term, namely:

$$L(\theta) = \ln(P_{total}) = \sum_{m=1}^{M}(y_n \ln(p) + (1-y_n)\ln(1-p) + \lambda\sum_{j=1}^{n}\theta_j^2) \tag{6}$$

The value of the loss function is the total probability of incidents, which is expected to be greater, but it is contrary to the meaning of loss; therefore, a minus sign can be used. Like linear regression, it should be minimized. The final optimization function can be obtained through iterative approximation to the minimum value by Stochastic Gradient Descent and taking the average log-likelihood loss of the entire data set:

$$\mathbf{J}(\theta) = -\frac{1}{m}\mathbf{lnL}(\theta) = \frac{1}{2m}\sum_{i=1}^{m}[(h_\theta(x^i) - y^i)^2 + \lambda\sum_{j=1}^{n}\theta_j^2] \tag{7}$$

where, $h_\theta = \theta^T x$; after calculating the parameters, the real data can be input into this model. With p = 0.5 as a threshold, a larger value means that there is rainfall, and a smaller one means that there is no rainfall.

3. Data Pre-processing and Evaluation Indicators

3.1. Survey Region

In this paper, the rainfall-related data collected by rain gauges of 88 remote measurement stations (39 hydrologic stations and 49 rainfall stations) in the Dadu River Basin were used, and their longitudes and latitudes, and rainfall data were taken as the subjects. Dadu River Basin covers an area of more than 700,000 square kilometers, with great variations in altitude, latitude and longitude, and complex topographic and ecological environment, which are irregular and make the climate, environment, and rainfall complicated. However, the historical rainfall can be used as the basic data to explore the potential correlation as the basis for pre-judging the current rainfall. The data mining algorithm can obtain their internal correlation. Therefore, the rainfall in various places must have certain inherent and implicit relationship under the action of the natural environment, and the region with a higher degree of correlation can directly reflect the changes in rainfall in the target region, and is of important reference significance to accuracy of the measured rainfall in the target region.

3.2. Sample Data

In this study, the rainfall-related data collected at 88 remote measurement stations from February 14, 2017 to June 30, 2020 were adopted, with the time cycle of days. At the same time, due to the different statistical date at each measurement point and missing data, date alignment between two measurement points was used for data synchronization, and complete structure of raw data shown in table 2 is:

POINT_CODE: Point code, each point is a rainfall station
DATA_TIME: Date (day)
VALUE: Value (ml)

Table 2. Complete structure of raw data.

POINT_CODE	DATA_TIME	VALUE
WW101YLW00004234	2020-03-15 08:00:00	0.00000000
WW101YLW00004234	2020-03-16 08:00:00	1.00000000
WW101YLW00004234	2020-03-17 08:00:00	1.50000000
WW101YLW00004234	2020-03-18 08:00:00	0.50000000
WW101YLW00004234	2020-03-19 08:00:00	0.00000000
WW101YLW00004234	2020-03-20 08:00:00	0.50000000
WW101YLW00004234	2020-03-21 08:00:00	3.00000000
WW101YLW00004234	2020-03-22 08:00:00	0.00000000
WW101YLW00004234	2020-03-23 08:00:00	0.00000000
WW101YLW00004234	2020-03-24 08:00:00	0.00000000
WW101YLW00004234	2020-03-25 08:00:00	0.00000000
WW101YLW00004234	2020-03-26 08:00:00	0.00000000
WW101YLW00004234	2020-03-27 08:00:00	0.00000000
WW101YLW00004234	2020-03-28 08:00:00	1.00000000
WW101YLW00004234	2020-03-29 08:00:00	1.00000000
WW101YLW00004234	2020-03-30 08:00:00	4.50000000
WW101YLW00004234	2020-03-31 08:00:00	0.00000000
WW101YLW00004234	2020-04-01 08:00:00	0.00000000
WW101YLW00004234	2020-04-02 08:00:00	0.00000000
WW101YLW00004234	2020-04-03 08:00:00	2.50000000
WW101YLW00004234	2020-04-04 08:00:00	4.50000000
WW101YLW00004234	2020-04-05 08:00:00	0.00000000
WW101YLW00004234	2020-04-06 08:00:00	0.00000000
WW101YLW00004234	2020-04-07 08:00:00	0.50000000
WW101YLW00004234	2020-04-08 08:00:00	0.00000000
WW101YLW00004234	2020-04-09 08:00:00	12.0000000
WW101YLW00004234	2020-04-10 08:00:00	0.00000000
WW101YLW00004234	2020-04-11 08:00:00	0.50000000

There was a total of 115, 920 articles of rainfall data, a small part of which can be visualized by visualization tools, as shown in figure 2.

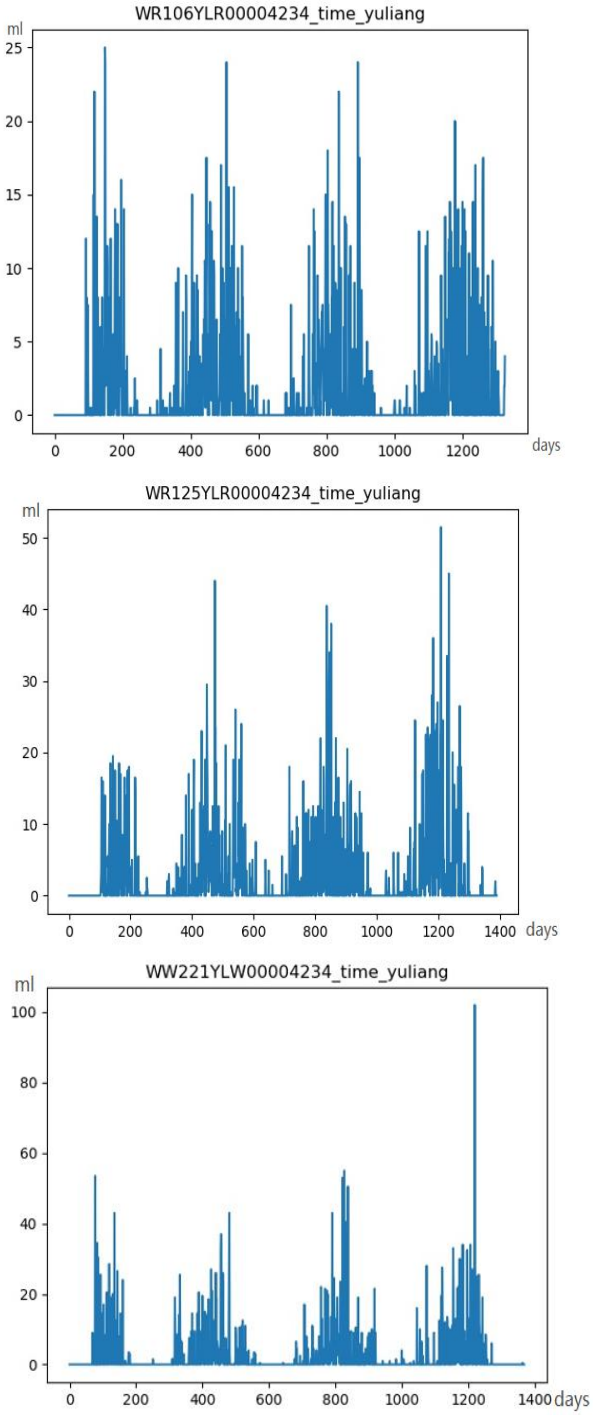


Figure 2. Characteristics of original data changing with time.

The visualized data showed that all regions basically conform to this law of distribution, which can be intuitively understood. The rainfall would generally rise, until the rainy season in spring and summer, and then drop in winter, forming a cycle each year.

3.3. Data Standardization

Data standardization [8] should be specially introduced: Since it is the basis of data mining, different evaluation indicators would have different measurement and dimensional units, thus affecting the results of data analysis. For eliminating the dimensional influence between indicators, all data should be standardized to resolve the problem of comparability between data indicators. After standardization of the raw data, the relevant indicators would be at the same order of magnitude, making it suitable for comprehensive comparative evaluation. The commonly used data standardization methods include Min-max standardization and z-score standardization [9], all of which intend to make different features have the same importance to the target, and prevent the unfair impact of large or small features on model parameters, thus avoiding the effect on accuracy of the model results.

But in this study, data standardization is not helpful to the model, and even has a negative effect. Based on the actual situation, all feature measurements are the same (ml). In some regions, since there is a high degree of correlation in rainfall, and the rainfall magnitudes are quite close, the non-standardization raw data may be more conducive to the fitting of the actual situation: Through iterative learning, the model can gradually enlarge the parameter β of the measurement point with a greater impact, and reduce the parameter β with a smaller impact. Therefore, in this case, no standardization was performed for the characteristic measurement points. The data pre-processing only involved the deletion of missing values, wrong data, and rainfall data of two measurement points synchronized by date. Subsequently, the data were split, 70% of data were taken as the training set and 30% the validation set.

3.4. Evaluation Indicators

In this paper, the logistic regression model was used to determine the daily rainfall of each remote measurement station. As for binary issues, the accuracy rate, precision rate, recall rate and F1 value were taken as the indicators to determine the quality of the results.

Where, accuracy rate [10] = the number of samples correctly classified/the total number of samples, but the accuracy rate would be lack of convincing power in the case of a data set with unbalanced data. In view of the unbalanced data in some data sets here, it would be not enough to evaluate the quality of a model based on accuracy rate only.

The forecasted results of a binary issue can be divided into the following four types:

- True Positive: forecasted value = 1, measured value = 1
- False Positive: forecasted value = 1, measured value = 0
- True Negative: forecasted value = 0, measured value = 0
- False Negative: forecasted value = 0, measured value = 1

Confusion matrices of the four results as shown in table 3:

Table 3. Confusion matrices of the four results.

Classification	Ture Value:1	Ture Value:0
Prediction value:1	TP (Ture Positive)	FP (False Positive)
Prediction value0	FN (False Positive)	TN (True Positive)

Precision rate: P = TP / (TP + FP)

Recall rate: R = TP / (TP + FN)

F1: Harmonic average of precision rate and recall rate

Furthermore, in order to evaluate the linear regression model for forecasting rainfall, RMSE and R^2 coefficient were used to represent the gap between the forecasted values and measured values, thus evaluating the quality of the forecasted values:

$$E_{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(R_{io} - R_{if})^2} \tag{8}$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (9)$$

R^2 is between 0 and 1, and if the value is closer to 1, it means that the regression fitting effect is better; generally, the model with a value exceeding 0.8 is high in goodness of fit.

The small rainfall is generally determined by the situation that the value added with the empirical threshold obtained by the model is smaller than the actual value for three consecutive days; then the collected values are deemed to be smaller.

4. Modeling and Training

In this paper, the data of daily rainfall collected at 88 measurement points from 2017 to 2020 were used for modeling, and 30% of the data were used for inspecting the model and evaluating the predictive accuracy. The accuracy of data collected by manual methods can only be determined based on manual observation and experience, which, as a time-consuming and labor-intensive method, may have many problems, with poor effects. In this paper, the logistic regression model was introduced to determine the daily rainfall at each remote measurement station. Based on historical data and experience, it was found that the rainfall data of strongly correlated remote measurement stations tended to be consistent in a three-day cycle.

In this study, data of the day and the previous two days of strongly correlated stations around the target remote measurement station were collected, and were input into the logistic regression model after binary processing (1=rainfall, 0=no rainfall), to speculate the theoretical rainfall of the target remote measurement station. Then the theoretical rainfall was compared with the measured rainfall, and discussed in the following situations:

(1) Transmitted data from the measurement point (measured rainfall) =0, theoretical rainfall=1: If the situation occurs for three consecutive days, an error is reported, and it can be determined that the rain gauges at the target measurement point are faulty;

(2) Transmitted data from the measurement point=0, theoretical rainfall=0: Normal, no processing required;

(3) Transmitted data from the measurement point=loss, theoretical rainfall=0: Error, there are faults of the rain gauges at the target measurement point;

(4) Transmitted data from the measurement point=loss, theoretical rainfall=1: Error, there are faults of the rain gauges at the target measurement point;

(5) Transmitted data from the measurement point=1, theoretical rainfall=1: Subsequent processing required, the rainfall should be estimated by the linear regression model, to determine whether the value is smaller.

Linear regression model:

The linear regression model was mainly used to calculate the theoretical rainfall of the target remote measurement station with the data collected at strongly correlated stations around the target remote measurement station, and determine whether the rainfall transmitted from the measurement point is smaller through comparing the theoretical rainfall with the transmitted data from the measurement point. In the case of 1,4 (data=0 or missing) after inputting the data in the logistic regression model, the theoretical rainfall should be calculated and filled with the linear regression model.

In the case of 5 the calculated theoretical value should be compared with the value transmitted from the measurement point. Based on the historical experience, it can be concluded that the data collected at the remote measurement stations with smaller rainfall are generally 30%-50% lower than the measured data of the surrounding remote measurement stations. In summary, as for the linear regression model with a 3-day cycle, if the data transmitted from a measurement point were lower than 70% of the calculated theoretical value, it can be determined that the rainfall data are smaller, and marked as abnormal; then an alarm would be released to indicate the failure in rain gauges of this measurement point. Logical flow of the logistic regression model is as shown in figure 3:

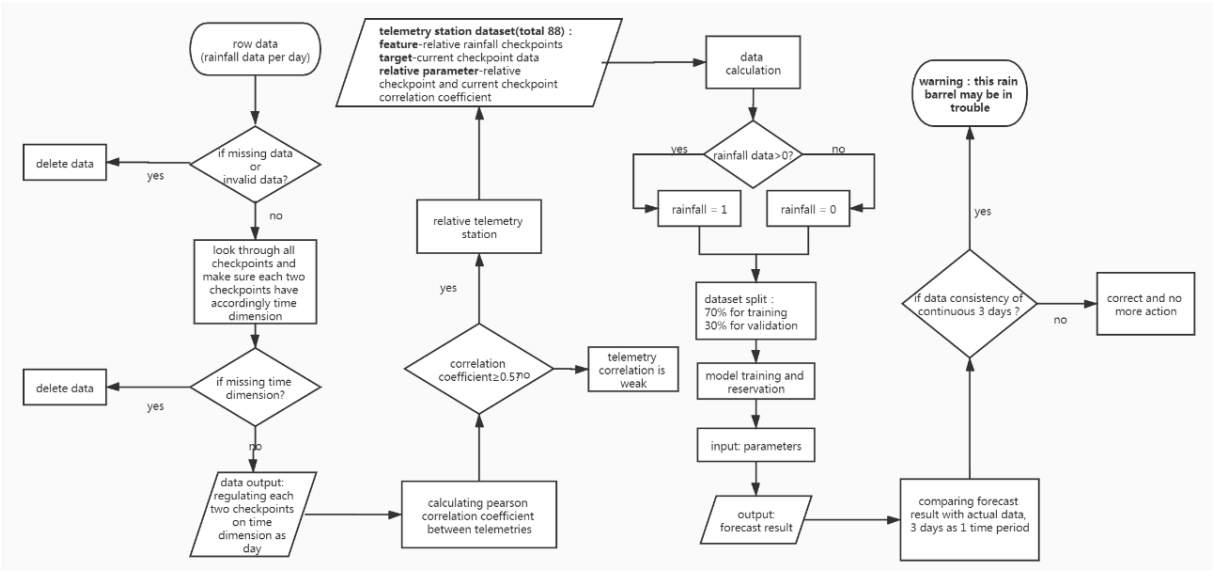


Figure 3. Logical flow of the logistic regression model.

In the experiment, each processed data set was input into the logistic regression model for separate training, testing and forecasting. The final test showed that the logistic regression model performed well in terms of the data set, with the accuracy rate r of over 90%. Logical flow of the linear regression model is shown in figure 4:

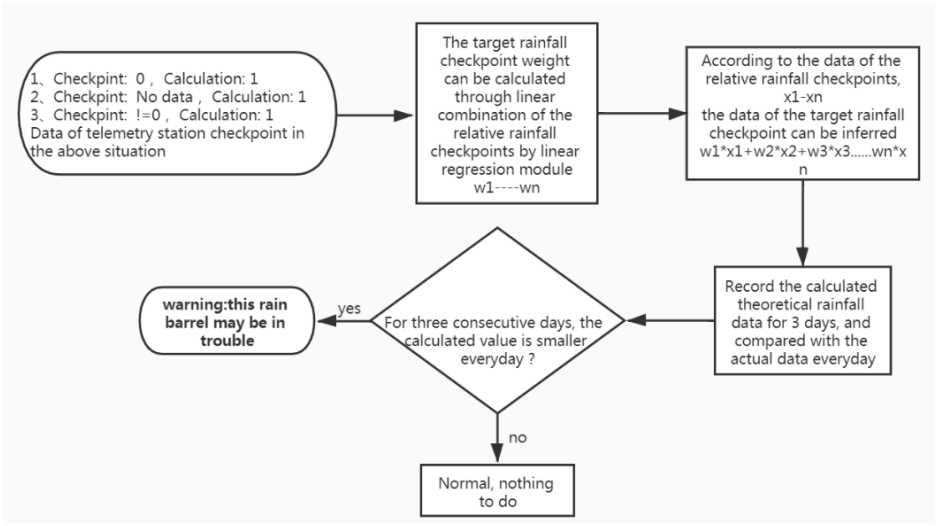


Figure 4. Logical flow of the linear regression model.

The linear regression model was operated after the logistic regression model. In the case of the above conditions 1, 4, 5 of rainfall at the target measurement points, the data were input into the linear regression model (figure 5); after completing the subsequent procedures, the theoretical rainfall at each target remote measurement point was calculated, to complement the error value and empty value 1,4, and then compared with the transmitted data, so as to determine whether the rainfall data were smaller 5.

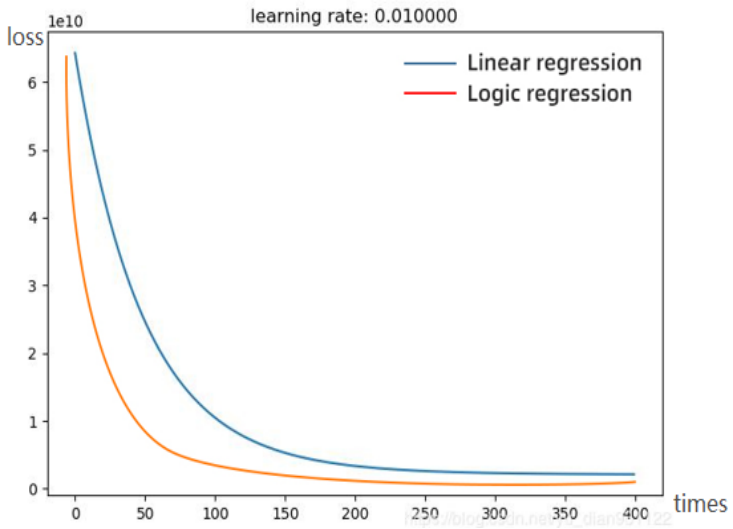


Figure 5. Iterative graph of linear regression and logistic regression mode.

It can be seen that after multiple iterations, the value of LOSS of linear regression and logistic regression was reduced to a small value. Considering the different complexity of models, the number of convergence iterations and the minimum value were not the same.

5. Experimental Analysis

5.1. Results of Logistic Regression Experiment

The results are shown below in table 4:

Table 4. Results of logistic regression experiment.

	Precision	Recall	FI Score	Number of samples	Accuracy
Labels_0	0.89	0.96	0.92	123	
Labels_1	0.95	0.86	0.91	111	
Average/Total	0.92	0.91	0.91	234	0.91

where, labels_0 means that there is no rainfall at a remote measurement station; labels_1 means that there is rainfall at a remote measurement station.

As for logistic regression, satisfactory results can be obtained as shown in data statistics; and logistic regression had a great advantage in the accuracy of sparse data, which was the reason for the use of this algorithm in the study. In actual, there was no rainfall on most days, there were many values of 0, and logistic regression could achieve a better fitting effect.

5.2. Results of Linear Regression Experiment

The results of linear regression experiment are shown above in table 5. The RMSE of the linear regression experiment is 115.397, the R^2 of the experiment is 0.927 and the accuracy of judgement of this experiment is 0.60.

Table 5. Results of linear regression experiment.

RMSE	**R^2**	**Accuracy of judgment**
115.397	0.927	0.60

5.3. Comparison with the Traditional Method

Based on the historical data, the traditional manual identification method was low in accuracy rate (only 0.2), and time-consuming and labor-intensive. The operators should inspect the equipment on site every time, and the failure of a rain gauge can only be identified through careful observation and even disassembling, which may completely use the data resources. Therefore, manual inspection is quite inefficient and error-prone. The data model can improve the work efficiency, reduce the pressure on the operators, and greatly increase the accuracy of determination of equipment conditions.

6. Conclusion

In this paper, correlation-based LR determination of abnormal data collected with rain gauges at remote measurement stations was proposed, and a forecasting model was constructed with LR and Pearson coefficient through making full use of the correlation between rainfall and natural environment. Of course, there are still some problems. In view of less error samples, intermittent rainfall data and inconspicuous fault characteristics, there is a great space for improvement of the model in identifying the conditions with small rainfall.

References

[1] Xu J, Luo W and Huang Y 2019 Dadu River Runoff Forecasting via Seq2Seq. *Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science (AICS 2019)* pp 513-517.

[2] Liu H S, Huang H Y and Xiao S H 2014 Calculation for power consumption of main transformer cooling system in hydropower plant based on Pearson correlation coefficient *Advanced Materials Research*. **3530** 263-267.

[3] Lefebvre M 2003 Short-term hydrological forecasts using linear regression/Prévisions hydrologiques à court terme obtenues en utilisant la régression linéaire *Revue des sciences de l'eau* **16** (2).

[4] Karlsson A 2007 Introduction to linear regression analysis *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **170** 856-857.

[5] Fang H B, Liu Z Q, Tan M T and Tian G L 2009 Regularized (bridge) logistic regression for variable selection based on ROC criterion *Statistics and Its Interface* **2** (4).

[6] Netrapalli P 2019 Stochastic gradient descent and its variants in machine learning *Journal of the Indian Institute of Science* **99** 201-213.

[7] Akinyemi I 2007 Comparative analysis of logistic regression and neural network for credit risk assessment *International Journal of Natural and Applied Sciences* **3** 441-447.

[8] Grannis S J, Xu H, Vest J R, Suranga K, Bo N, Ben M, Rita T and R Josh 2019 Evaluating the effect of data standardization and validation on patient matching accuracy *Journal of the American Medical Informatics Association*: *JAMIA* **26** (5).

[9] Ayatollahi S M T 2009 Age standardization of weight-for-height in children using a unified Z-score method *Annals of Human Biology* **22** 151-162.

[10] Lathrop Q N and Cheng Y 2013 Two approaches to estimation of classification accuracy rate under item response theory *Applied Psychological Measurement* **37** 226-241.