Advances in Computer, Signals and Systems (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2022.060302
ISSN 2371-8838 Vol. 6 Num. 3

Data analysis of campus water supply system based on random forest algorithm and time series model

Changsheng Li, Shuo Liu, Sujun Liu

College of Information Processing and Control Engineering, Lanzhou Petrochemical University of Vocational Technology, Lanzhou, China, 730060

Keywords: Water supply system, Decision tree, Random forest (RF), Time series model (ARMA), Autocorrelation coefficient

Abstract: Firstly, this paper makes statistics and Analysis on the water use characteristics of different functional areas in the campus, excavates the change law of each water meter data with time dimension and the proportion of total water use, so as to excavate the water use characteristics of different functional areas in the campus (such as office area, living area, logistics service area, etc.). Then, this paper constructs a random forest (DRF) model based on the decision tree algorithm, analyzes the "caliber" and "water consumption" of the water meter through the python language design algorithm and the hierarchical relationship of the campus water meter, then excavates the relationship model between the water meter data, and analyzes the classification error of the model through the existing data provided in the annex; At the same time, according to the collection time of each water meter reading, this paper analyzes the leakage of the water supply pipe network in the campus based on the time series model (ARMA), models and excavates the change law of water consumption from the first level table to the fourth level table in four quarters, and then monitors the leakage of the campus water supply pipe network system in real time through the error threshold of the model. Finally, combined with the VC dimension theory in statistical theory, this paper analyzes the relationship between the leakage degree of water transmission pipe network and the cost (labor cost and material cost) and water price of pipe network maintenance, and gives a feasible decision-making scheme for pipe network maintenance of campus water supply system.

1. Introduction

Campus water supply system is an important part of campus public facilities. In order to ensure the normal operation of campus water supply system, the school needs to invest a lot of human, material and financial resources. With the development of scientific water supply system, a large number of real-time data can be obtained. Based on these data, the logistics department hopes to find and solve the problems existing in the water supply system in time through modeling and data mining, so as to improve the level of campus service and management.

To analyze the water use characteristics of different functional areas in the campus, we need to combine the "consumption" data provided in the first quarter, the second quarter, the third quarter and the fourth quarter to count and analyze the change law of each water meter data with the time dimension, so as to divide the different functional areas of water use in the campus (such as office area, living area, logistics service area, etc.), and then summarize the water use characteristics of each functional area.

Establish the relationship model between water meter data, build the random forest (DRF) model based on the decision tree algorithm, analyze the "caliber" and "water consumption" of water meter in combination with the hierarchical relationship of campus water meter, then mine the relationship model between water meter data, and analyze the classification error of the model through the existing data provided in the annex.

For the analysis of the leakage of the campus water supply network, based on the time series model (ARMA), this paper uses the water consumption of each quarter to statistically analyze the difference of the water consumption from the first level table to the fourth level table from the time dimension to analyze the leakage of the campus water supply network system.

Through the time series model (ARMA) analysis of the leakage of the campus water supply network, the water consumption law of each water meter can be excavated. According to the water consumption law of each water meter, when there is an abnormal water consumption of the water meter in the water supply network, the campus area position corresponding to the water meter can be further analyzed, and the leakage in the campus water supply network can be detected through the inspection of the area position.

For the formulation of the optimal maintenance decision-making scheme for the leakage of the campus water supply network, based on the VC dimension theory in statistics, this paper analyzes the relationship between the leakage degree of the water supply network and the cost (labor cost and material cost) and water price of pipe network maintenance, and finally formulates the optimal pipe network maintenance decision-making scheme.

2. Analysis of water characteristics in Campus

Firstly, the water meter data in the "water meter level" data set is filtered by layers, and the corresponding areas of water meters at different levels are classified and summarized by using the "water meter number" corresponding to the fields "primary meter number "," secondary meter number", "tertiary meter number" and "tertiary meter number" (table 1)

Table 1 Water meter number corresponding to primary water meter

Water meter No	Water meter level	Primary meter code	Water meter name	User number	User number	caliber
3620300500	1	416X	School Hospital South+	30089	School Hospital South+	100
3620300100	1	405X	XXX Garden+	30090	XXX Garden+	150
3620300300	1	403X	64397 sub table	30095	East Gate greenhouse (auxiliary table)	200
3620300200	1	401X	Breeding team 6721 Deputy tables+	30081	Breeding team 6721 Deputy tables	100
3620303000	1	419T	Temporary dormitory of fish culture group+	30093	Temporary dormitory of fish culture group	15
3620302900	1	418T	Toilet of fish culture group+	30092	Toilet of fish culture group	15
3620302800	1	417T	Greenhouse of agricultural test station+	30091	Greenhouse of agricultural test station	15
3030100102	1	413T	Property	30097	Property	15
3030100101	1	412X	Tennis court duty room	30096	Tennis court duty room	15
3312800100	1	411T	XXXs Hotel	30004	XXXs Hotel	80
3620302200	1	406T	East Gate reception room+	30066	East Gate reception room+	20

Then, in the "first quarter" data set, according to the water meter name of the "water meter level" data set, different functional areas in the campus are divided, and the total water consumption of each functional area is counted as follows:

Table 2 Total water consumption of each functional area

Functional area	Include architecture	Total water consumption	Proportion of water consumption
Living area	Dormitory No. 1, No. 2, No. 5789, dormitory of the security office of the attached room of the aquaculture hall +, temporary dormitory of the fish culture group +, canteens No. 1, No. 2 and No. 5 of the international student building (New), track and field field field / fish culture group / building / nano building / toilet of the aquaculture Hall (public, first floor and second floor)	7762.69	8.61%
Activity area	Retirement activity room, gymnasium, cadre training building, swimming pool, central pool	2304.14	2.55%
Water affairs and green area	3/4/5/8 community heat pump hot water, boiler room, tea garden, pump room of central building, flower garden +, East Gate greenhouse / greenhouse, botanical garden, greenhouse of agricultural test station	10154.99	11.26%
Logistics service area	Breeding house, breeding team, later building, property, East Gate / new gate reception room, motorcade, bungalow behind the hall, high-rise room, hotel, education supermarket, barber shop, bathroom, hotel, bookstore, school hospital south, School Hospital, old medical building (2 / 3), old medical room building	23940.24	26.54%
Administrative Area	East West building, science museum, library, international nano Research Institute, nano building (3/4/5), aerospace, dangerous goods warehouse, poison Research Institute, seed building, Judicial Expertise Center, sewage treatment	12727.05	14.11%
Other areas	Area (1 / 2 / 3 / 4)	33317.32	36.93%

Then it statistically analyzes the change law of water use in different functional areas in different quarters, and finally analyzes the water use characteristics of different functional areas in the campus. Then, the water consumption law of water meters at different levels of functional areas in the campus in four quarters is statistically analyzed. Figure 1 and the proportion of total water consumption in each functional area are shown in Figure 2.

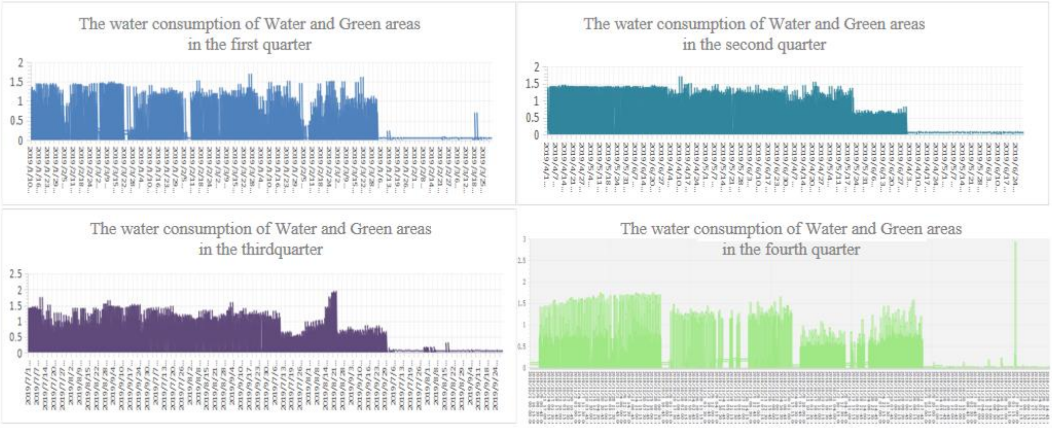

Figure 1 Change law of water use in water affairs and green areas in each quarter

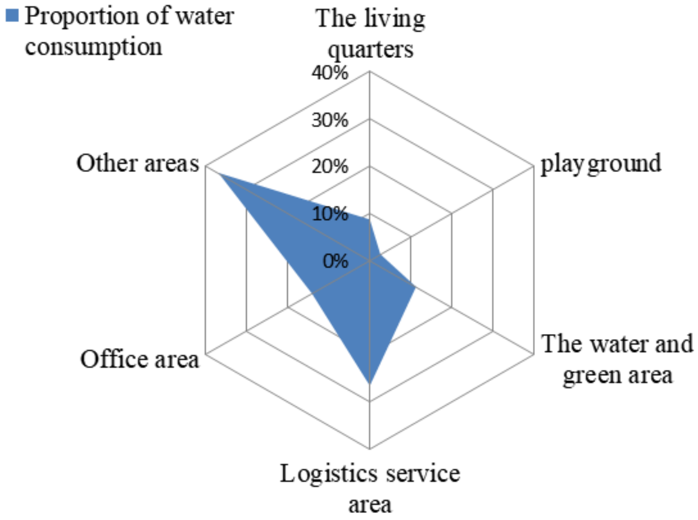

Figure 2 Proportion of total water consumption in each functional area in the first quarter

By analyzing the water consumption of each functional area in the campus in four quarters, it can be seen that the water consumption of each functional area presents different trends. For example, for the functional areas divided into water affairs and green areas, the frequency of water use in each quarter is high, especially in the first two months of each quarter; The total water consumption classified in other regions (1/2/3/4) accounted for 36.93% in the first quarter, with the most water consumption; The functional area classified in the office area uses water relatively frequently, and the total amount of water used every day is relatively stable. At the same time, the daily and monthly water consumption in this area is relatively stable, but the total amount of water used in January and February in the first quarter and August in the third quarter of the four quarters decreases, which is in line with the school calendar work and rest law.

3. Analysis of relationship model of water meter data based on random forest algorithm (DRF) constructed by decision tree

3.1 Build DRF model to analyze the relationship between water meter data

The four quarter water use data sets in this paper build a random forest model based on the decision tree algorithm in the machine learning algorithm. The "caliber" and "water consumption" of the water meter are important characteristics to establish a decision tree with the "water meter level" as the target variable, and then analyze the forest model composed of multiple decision trees, so as to mine the relationship model between the water meter data.

Random forest (RF)

Random forest is a set of attribute classifiers $\{h(\chi,\beta_k),\kappa=1,2,3\cdots\}$, in which the meta classifier $h(\chi,\beta_k)$ is a classification decision tree without pruning constructed by cart algorithm; χ is the input vector; β_κ is an independent identically distributed random vector. Determines the growth process of a single tree; The output of the forest adopts the simple majority voting method, or the simple average of the output results of a single tree. The simple majority voting method is mainly aimed at the classification model; The simple average of the output results of a single tree is mainly for the regression model. Its essence is an improvement of the decision tree algorithm, which combines multiple decision trees. The establishment of each tree depends on an independent sample. Each tree in the forest has the same distribution, and the classification error depends on the classification ability of each decision tree and the correlation between them.

Random forest is a random forest composed of K decision trees, which is randomly selected from the original training sample set n through the boot strap resampling technology, and then generated according to the independent sample set. The classification result of new data is determined by the score formed by the number of votes of the decision tree.

The Bootstrap method is based on multiple random training sets. Let the number of attributes of the sample be M, and m is an integer greater than zero and less than M. The flow of random forest algorithm is as follows:

Random forest algorithm (DRF) based on decision tree

The bootstrap method is used to resample and randomly generate N training sets S_1, S_2, S_n.

Using each training set, generate the corresponding decision trees C1, C2,..., Cr; Before selecting attributes on each non leaf node (internal node), M attributes are randomly selected from m attributes as the splitting attribute set of the current node, and the node is split in the best splitting way of these m attributes. The value of m remains unchanged during the growth of the whole forest.

Each tree grows completely without pruning. For the test set samples, each decision tree is used to test, and the corresponding categories C (1), C (2), * and C (n) are obtained.

Using the voting method, the category with the most output in the n decision trees is taken as the category of the test set sample

This article is based on Python language, in Python 3 6.5. Pycharm 2018.2.4 x64 software environment mainly realizes the random forest algorithm based on decision tree through the machine learning scikit learn Library in Python. In the constructed random forest, the construction parameters of one sub decision tree are as follows:

```
DecisionTreeClassifier(
        ccp_alpha=0.0, class_weight=None, criterion='gini',
          max_depth=4, max_features='auto', max_leaf_nodes=None,
          min_impurity_decrease=0.0, min_impurity_split=None,
          min_samples_leaf=1, min_samples_split=2,
          min_weight_fraction_leaf=0.2, presort='deprecated',
          random_state=1343987037, splitter='best')
```

Using the existing data, combined with the hierarchical relationship of the campus water meter, taking the "caliber" and "water consumption" of the water meter as the eigenvalues, a random forest model is constructed based on the decision tree algorithm. Through the analysis of the classification error of the model through the existing data provided in the annex, the constructed random forest RF model predicts each water meter level on the first quarter water consumption data set, and the final ROC is obtained_ The AUC value is 0.87 for the random forest RF model, while the single decision tree decision classifier is 0.67. For the confidence interval of the accuracy of machine learning algorithm analysis, the ROC of RF model_ The AUC value of 0.87 has high reliability and good performance in hierarchical classification of water meter data.

3.2 Analysis of water use based on time series model (ARMA)

3.2.1 Introduction to time series (ARMA) model

The full name of ARMA (auto regressive and moving average model) model is autoregressive moving average model. It is not only the most commonly used model to bridge the stationary series, but also an important method to study the time series. It can be subdivided into AR model, MA model and ARMA. Can be regarded as multiple linear regression model.

AR: abbreviated as AR (P), autoregressive. A model that uses dependencies between observations and some lag observations. That is, the value of random variable is the multiple linear regression of the previous P period, which is mainly affected by the sequence value of the past P period. The error term is the current random interference, which is a zero mean white noise sequence;

MA: abbreviated as MA (q), moving average. A model using the dependence between observations and residuals in the moving average model applied to lag observations. That is, the value of the random variable at time t is the multivariate linear function of the random disturbance in the previous q period. The error term is the random interference of the current period, which is a zero mean white noise sequence and is the mean of the sequence. It is considered that it is mainly affected by the error term in the past q period.

ARMA model: stationary series fitting model. ARMA (p, q) model can be expressed as:

$$y_t = \theta_1 y_{t-1} + \ldots + \theta_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_q \varepsilon_{t-q} \qquad (1)$$

Where: p is the order of autoregressive model; q is the order of the moving average model; ϕ i(i=1, 2, …, p), θj (j = 1, 2,..., q) is the undetermined coefficient of the model; εt is the residual error; yt is the observed value. The value of random variable is related not only to the sequence value of the previous p period, but also to the random disturbance of the previous q period.

3.2.2 Analysis of water use based on time series model (ARMA)

Obtain the time series data of the observed system,Plot the data and observe whether it is a stationary time series; For non-stationary time series, d-order difference operation should be carried out first to convert it into stationary time series； After the previous processing, the stationary time series has been obtained. The autocorrelation coefficient ACF and partial autocorrelation coefficient PACF of stationary time series are obtained respectively. Through the analysis of autocorrelation diagram and partial autocorrelation diagram, the optimal level p and order q are obtained;From the above d, q and p, the ARMA model is obtained. Then start to test the model.

According to the data given in this paper, in the well maintained public water supply network, the average water loss is about 5%; In the older pipe network, there will be more water loss. Therefore, the confidence interval of the model set in this paper is 95%:

$$\Pr\left(-\frac{2}{\sqrt{n}} \le \hat{\rho}_k \le \frac{2}{\sqrt{n}}\right) \ge 0.95$$
$$\Pr\left(-\frac{2}{\sqrt{n}} \le \hat{\phi}_{kk} \le \frac{2}{\sqrt{n}}\right) \ge 0.95 \qquad (2)$$

After analyzing the water consumption in the first quarter, it can be seen that the test results of ARMA model in this paper are as follows:

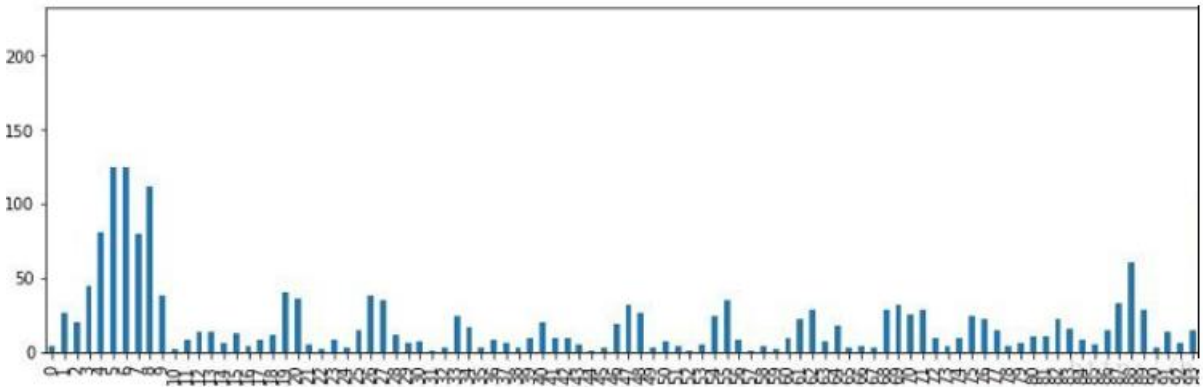

Figure 3 Statistics of water consumption by day

Inspection results:

Test Statistic	-2.709577
p-value	0.072396
Lags Used	12.000000
Number of Observations Used	119.000000
Critical Value (1%)	-3.486535
Critical Value (5%)	-2.886151
Critical Value (10%)	-2.579896

By observing its statistics, it is found that the autocorrelation of the sequence has the characteristics of rapid attenuation when the confidence level is 95%, and the T statistics is significant when the confidence level is 95%.

$$ACF(k) = p_k = \frac{Cov(y_t, y_{t-k})}{Var(y_t)} \qquad (3)$$

The changes of ACF and PACF are as follows:

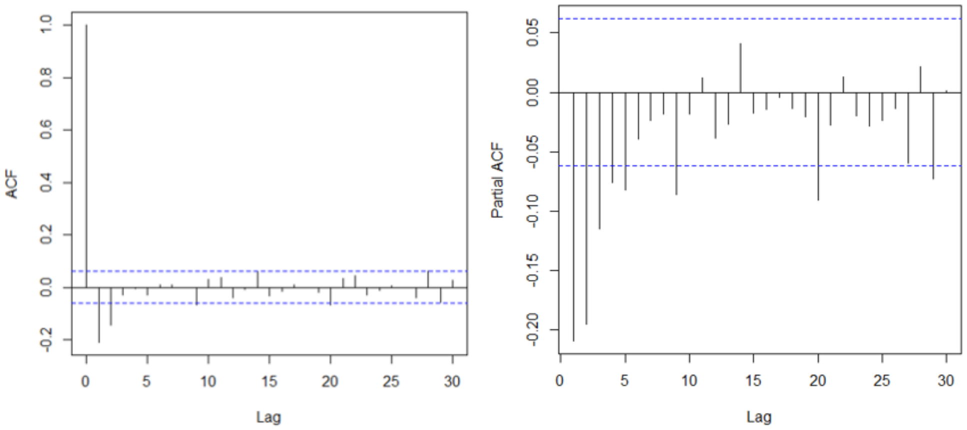

Figure. 4 Variation trend of ACF and PACF values

In this paper, the root mean square error (RMSE) is used to evaluate the quality of fitting in the model samples. The final root mean square error is 0.072396, and the prediction effect is good. In the water supply pipe network system, it is the allowable error range, which has certain guiding significance for the leakage detection of campus water supply pipe network system in the future. At the same time, through the verification of some data in the first quarter, the accuracy between the predicted value and the real value obtained in this paper is 96.27%, indicating that there is leakage in the campus water supply pipe network system.

4. Real time monitoring of leakage position in water supply network in Campus

Based on the relationship model of water meter data constructed by random forest algorithm based on decision tree, this paper analyzes the leakage of campus water supply network, excavates some water consumption laws of water meters in various functional areas of the campus, and designs a system that can find and determine the leakage position of water supply pipeline in time according to the water consumption laws of water meters.

When the water consumption of the water meter is abnormal in the water supply network (the threshold of the set time series (ARMA) model is 5%, and the water loss in the water meter area is calculated in combination with the historical data of the upper and lower water meters and the readings of the currently collected data. If the preset threshold is exceeded, the leakage occurs in the area), the level value of the campus area corresponding to the water meter can be further analyzed, The leakage of the water supply network in the campus is detected through the inspection of the regional position, so as to help the school solve the problem of detecting the leakage position of the water supply network.

5. Determination of optimal maintenance decision scheme for leakage of water supply network

For the formulation of the optimal maintenance decision-making scheme for the leakage of the campus water supply network, based on the VC dimension theory in statistics, this paper analyzes the relationship between the leakage degree of the water supply network and the maintenance cost (labor cost and material cost) and water price. When the maintenance cost is increased and the maintenance intensity is increased, the leakage degree of water transmission pipe network can be reduced, that is, the more serious the leakage degree of pipe network is, the higher the maintenance cost is; However, when the maintenance cost increases to a certain range, the leakage degree of water transmission network will not be reduced. Therefore, the best maintenance decision-making scheme provided in this paper is: when the maintenance cost investment is too large and the leakage reduction is limited, it is not advocated to blindly seek to reduce the leakage level. At this time, it is necessary to compare and analyze the water price and maintenance cost of water loss in reality, find the balanced cost threshold, and finally formulate the optimal pipe network maintenance decision-making scheme.

6. Conclusion

When constructing the random forest model (DRF) based on the decision tree algorithm, this paper constructs the model with the characteristics of "caliber" and "water consumption" of the water meter combined with the hierarchical relationship of the campus water meter, and then excavates the relationship model between the water meter data. The characteristic data used are only two characteristics of "caliber" and "water consumption", which has too little information input for the division of the hierarchical model of the water meter, To some extent, it will cause the wrong division of water meter hierarchy. Therefore, when establishing the water meter relationship model, we should increase the stability of water meter reading and other relevant factors as much as possible to the decision-making basis of RF model.

For the analysis of the leakage of the water supply network in the campus, based on the time series model (ARMA), when mining the water use law of each water meter, the default water meter reading is stable. However, when the data is not stationary data, the difference method needs to be used for processing. When the adjacent values of two columns in the sequence are equal, the previous column may be removed, so the processed data may not be distributed according to the data of each day.

Acknowledgement

Foundation Project: The teaching and scientific research project of Lanzhou Petrochemical Vocational and Technical College (KJ2020-2), The teaching and scientific research project of Lanzhou Petrochemical Vocational and Technical College (KJ2021-14)

References

[1] Theodoris et al., Li Jingjiao, Wang Aixia, Wang Jiao, et al., pattern recognition (Fourth Edition) Beijing: Electronic Industry Press, February 2010

[2] Li Ying, Qian Jianguo, Wang Xiao, Mo Jianguo, Shi zhengchai. Research on Intelligent Fault Diagnosis System of power grid system based on data mining [J]. Automation and instrumentation, 2020 (07): 205-208

[3] Shao Qi, Chen Yunhao, Yang Shuting, Zhao Yifei, Li Jing. Hyperspectral image identification of Maize Varieties Based on random forest algorithm [J]. Geography and geographic information science, 2019,35 (05): 34-39

[4] Wu Qinghua. Random forest algorithm and its application in metabolic fingerprint [D]. Central South University, 2013

[5] Lan Hua, Liao Zhimin, Zhao Yang. Output prediction of photovoltaic power station based on ARMA model [J]. Electrical measurement and instrumentation, 2011,48 (02): 31-35

[6] Yang Mao, Xiong Hao, Yan Gangui, Mu gang. Research on real-time prediction of wind power based on data mining and fuzzy clustering [J]. Power system protection and control, 2013,41 (01): 1-6

[7] Feng pan, Cao Xianbing. Empirical research on stock price analysis and prediction based on ARMA model [J]. Practice and understanding of mathematics, 2011,41 (22): 84-90