Water Supply

IWA PUBLISHING

Water Supply Vol 22 No 1, 278 doi: 10.2166/ws.2021.272

Sensitivity analysis and prediction of water supply and demand in Shenzhen based on an ELRF algorithm and a self-adaptive regression coupling model

Xin Liu [a,b], Xuefeng Sang[b,*], Jiaxuan Chang[b], Yang Zheng[b] and Yuping Han[a]

[a] North China University of Water Resources and Electric Power, Zhengzhou 450046, China
[b] China Institute of Water Resources and Hydropower Research, Beijing 100038, China
*Corresponding author. E-mail: sangxf@iwhr.com

XL, 0000-0003-0902-6277

ABSTRACT

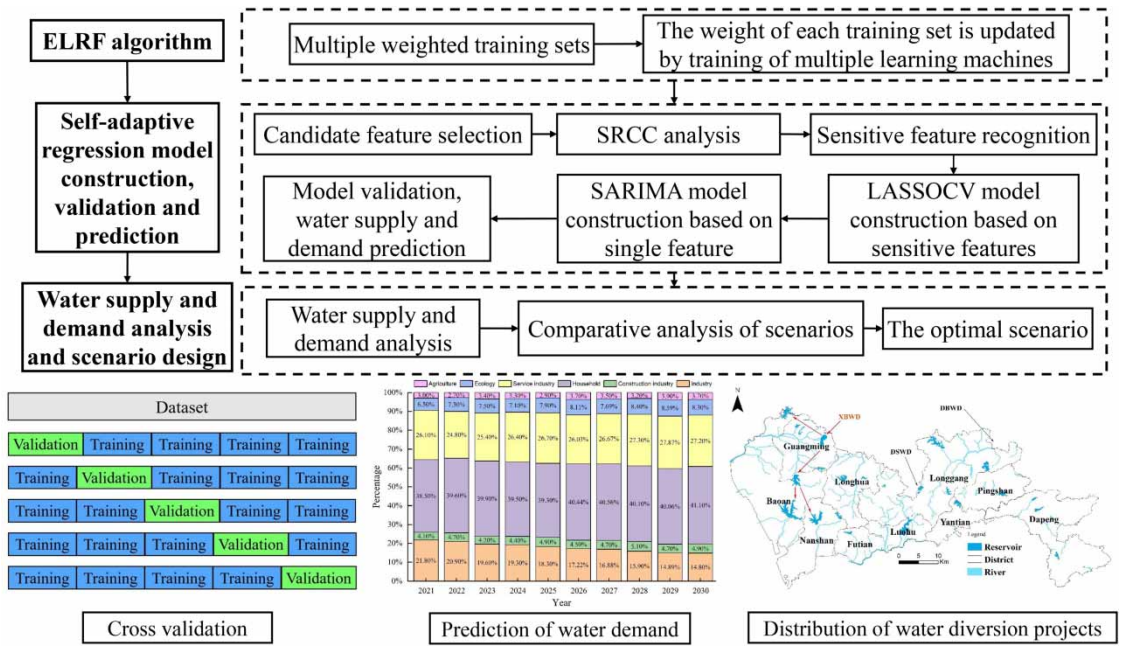
Given that sensitive feature recognition plays an important role in the prediction and analysis of water supply and demand, how to conduct effective sensitive feature recognition has become a critical problem. The current algorithms and recognition models are easily affected by multicollinearity between features. Moreover, these algorithms include only a single learning machine, which exposes large limitations in the process of sensitive feature recognition. In this study, an ensemble learning random forest (ELRF) algorithm, including multiple learning machines, was proposed to recognize sensitive features. A self-adaptive regression coupling model was developed to predict water supply and demand in Shenzhen in the next ten years. Results validate that the ELRF algorithm can effectively recognize sensitive features compared with decision tree and regular random forest algorithms. The model used in this study shows a strong self-adaptive ability in the modeling process of multiple regression. The water demand in Shenzhen will reach 2.2 billion m^3 in 2025 and 2.35 billion m^3 in 2030, which will exceeded the water supply ability of Shenzhen. Furthermore, three scenarios are designed in terms of water supply security and economic operation, and a comparative analysis is performed to obtain an optimal scenario.

Key words: ensemble learning random forest, multicollinearity, prediction of water supply and demand, self-adaptive regression coupling model, sensitive feature recognition, Shenzhen

HIGHLIGHTS

- An ensemble learning random forest algorithm was developed.
- Fivefold cross-validation was integrated into the model to solve the problem of multicollinearity.
- A seasonal autoregressive integrated moving average model was developed.
- The least absolute shrinkage and selection operator model with cross-validation was designed and developed.
- Three scenarios that can solve water resources shortage in Shenzhen were proposed.

GRAPHICAL ABSTRACT



INTRODUCTION

The prediction of water supply and demand for water resources planning is essential. If the water supply and demand in the future can be accurately predicted, then we can grasp the growth of water use and determine water resources shortage in time so that we can implement water resources planning more scientifically and effectively. There is an urgent need for accurate prediction of water supply and demand, especially for cities where local water resources are extremely scarce. However, rapid city development, modernization construction, frequent population flow, and water resources pollution make the physical environments of cities constantly change. These facts bring great challenges to the prediction of water supply and demand.

The mechanism model based on processes and phenomena depends on the external physical environment to some extent, and the model needs considerable measured data. In some cases, the mechanism model must also make some assumptions. Meanwhile, the measured data that can be collected generally have some missing values and outliers, and the external physical environment is always changing. These problems will have a great influence on the results of the mechanism model, and those assumptions can weaken the effectiveness of the results to some extent. Nonetheless, data-driven models, such as the regression model (Safari 2019; Reis *et al.* 2020), artificial neural network model (Praveen *et al.* 2020), and long short term memory recurrent neural network model (Nasser *et al.* 2020; Bai *et al.* 2021), do not depend on physical environments and assumptions. Additionally, they can find potential quantitative relations and correlation relationships between features and can learn valuable knowledge previously unknown. These models have also been widely applied in many fields.

Regression models can describe the relationship between variables, such as multiple linear regression (Cross *et al.* 2020), logistic regression (Zandi *et al.* 2016), Bayesian regression (Yang & Ng 2019), support vector regression (Ebtehaj & Bonakdari 2016), and polynomial regression (Chen *et al.* 2019). Although the coefficients of equations represent correlation relationships and correlation degrees between independent and dependent variables, regression models should determine independent variables in advance. Therefore, the performance of regression models is largely dependent on the selection of independent variables. When regression models need to cope with many independent variables, effectively learning enough potential quantitative relations and correlation relationships between variables is difficult. Once an independent variable that is more sensitive to a dependent variable is added to a regression equation, a previously sensitive independent variable is likely to become less sensitive. In addition, the parameters of regression models are solved on the basis of a fixed algorithm, and the process of parameter adjustment is not validated, which may make the analysis results unreliable. More importantly, according to several studies and experiments, serious multicollinearity (Kroll & Song 2013; Bassiouni

et al. 2016) among variables distorts regression models, and such models may generate spurious-regression equations. Quantitative relations and correlation relationships in spurious-regression equations are wrong, which is because these regression models lack the self-adaptive ability to autonomously adjust the parameters of models on the basis of new variables.

A neural network model has similar problems, and the performance of this model also depends on the selection of input features to a large extent. If the input and output features of a model have a strong correlation relationship, then the modeling effect will be good, and the input features can well simulate the changes of output features. Moreover, the relationship between input and output features established by a neural network model is transmitted through weight and bias parameters in a model. All features employ a set of parameters in a neural network model; hence, we cannot obtain the correlation degree between the input and output features. We also do not know whether the features are positively or negatively correlated. In addition, neural network models have a high probability of overfitting, which can weaken their generalization ability.

Recently, some machine learning methods, such as the autoregressive integrated moving average (ARIMA) (Nguyen 2020) model and random forest algorithm (Avanzi *et al.* 2019), have been widely utilized in runoff prediction, water demand prediction, sensitivity analysis, and feature recognition, and some good results have been obtained. Nevertheless, an ARIMA model can only construct an autoregressive equation of a single feature and cannot construct multiple regression models. Moreover, model parameters only include autoregressive, difference, and moving average orders; thus, models have certain limitations due to fewer parameters. Meanwhile, a random forest algorithm is weak and only includes single learning machine, which also has some limitations.

This study aims to recognize features that are sensitive to water supply and predict the water supply and demand in the future. In this study, an ensemble learning random forest (ELRF) algorithm, including multiple learning machines, was proposed for sensitive feature recognition. An adaptive regression model coupling the seasonal autoregressive integrated moving average (SARIMA) and the least absolute shrinkage and selection operator with the cross-validation (LASSOCV) was developed to predict water supply and demand. The coupling model can self-adaptively construct autoregressive or multiple regression models on the basis of the number of features and can self-adaptively adjust parameters in the modeling process of multiple regression. Shenzhen was selected as the study area to validate the feasibility and effectiveness of the algorithm and the model used in this study. In addition, three scenarios were proposed on the basis of the predicted results of the coupling model to balance the water supply and demand in Shenzhen to keep water resources shortage from becoming an obstacle to the development of the city. This research method can also provide reference for the sensitive feature analysis and prediction of water supply and demand in other megacities.

STUDY AREA AND DATA

Study area

Shenzhen (Figure 1) is the national economic and science and technology innovation center. It is a modern megacity with a total population of more than 20 million. It has also created more than 1,000 firsts in China and is one of the cities with the best business environment in China. Many internationally competitive enterprises, such as Huawei, Pingan, Merchants, and Tencent, are based in Shenzhen, and the entrepreneurial density of the city ranks first in China. Consequently, Shenzhen attracts numerous talents, which leads to the rapid increase of the floating population. Although the increase of population can promote the economic growth, it will also increase the amount of water supply.

However, Shenzhen faces a serious problem of water resources shortage. Affected by topography and geomorphology, most areas of Shenzhen are low hilly areas with no big rivers. Although small rivers are widely distributed, the mainstream of the rivers is short, and the runoff amount is small, thereby revealing that the amount of self-produced water is small. Therefore, local water resources in Shenzhen are extremely scarce, and the water supply mainly relies on water diversion. Presently, the water supply pressure of Shenzhen is mainly in the western region, and two water diversion projects exist in the city. The Dongbu Water Diversion (DBWD) transfers water from the eastern part of Shenzhen to reservoirs and waterworks in the west, while the Dongshen Water Diversion (DSWD) transfers water from the middle part of city to reservoirs and waterworks in the west.

Due to the economic development of the cities around the water diversion projects, water resources have been polluted to a certain extent, which poses a new challenge to the water diversion. The next ten years will be a critical period for Shenzhen to build an example of a socialist modernization megacity and a crucial period for the construction of Guangdong–Hong Kong–Macao Greater Bay Area. Shenzhen will attract more talents in the future. Although the Shenzhen government has been encouraging citizens to save water through some campaigns and propaganda, the growth of water supply is still relatively large. Therefore, systematically analyzing the sensitive features that promote the growth of water supply and predicting the

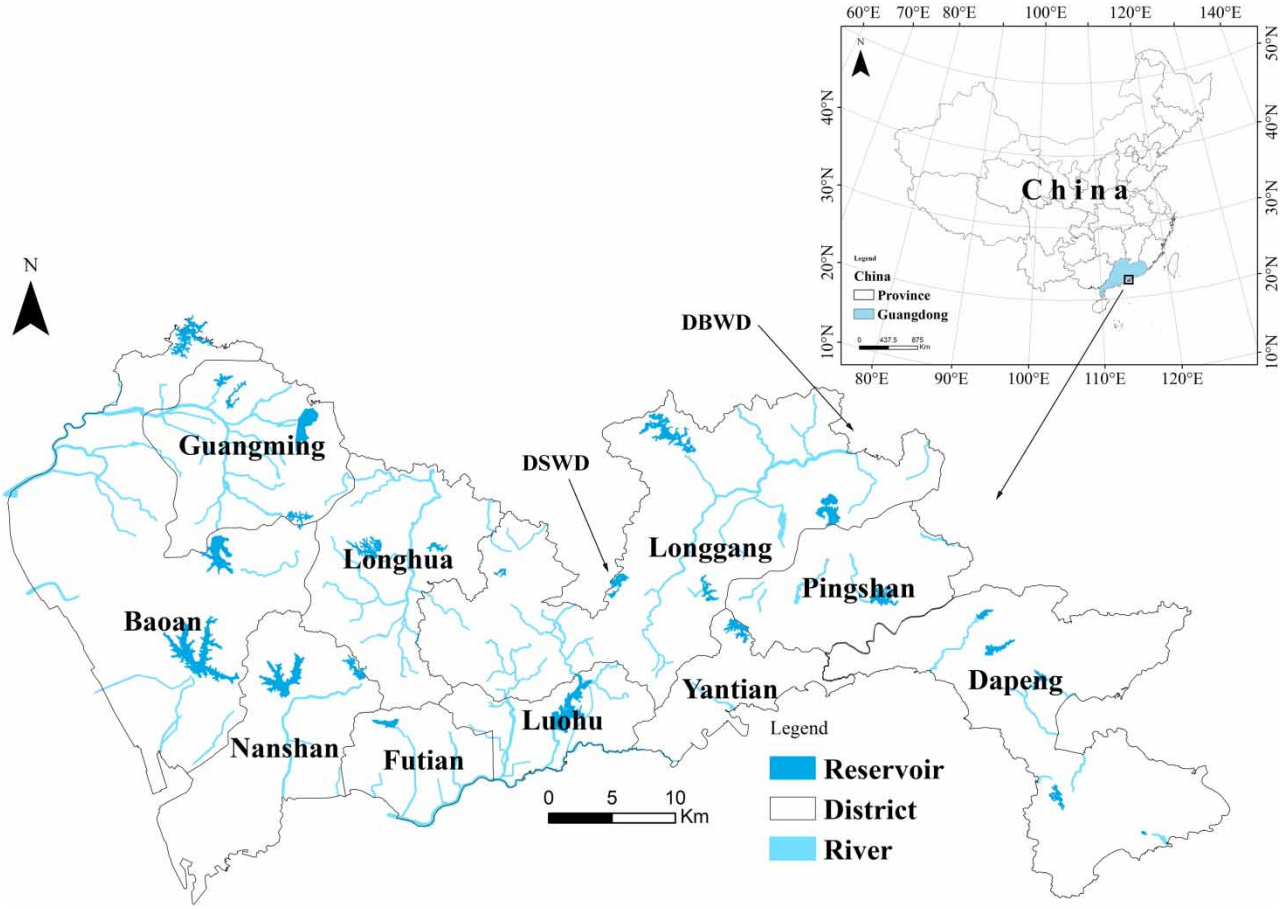


Figure 1 | Location and profile of Shenzhen. The Dongbu Water Diversion (DBWD) project and the Dongshen Water Diversion (DSWD) project are two existing water diversion projects in Shenzhen.

water supply and demand of Shenzhen in the future to determine the shortage of water resources early are necessary. Otherwise, the lack of water resources is likely to become a bottleneck in building a world-class bay area city in Shenzhen. In addition, the prediction of the water supply and demand helps in applying for the quota of the water diversion in the next year.

Data

Many variables can affect water supply. After the analysis of the actual situation of the water supply and demand and industrial structure of Shenzhen, the data of 18 variables from 2004 to 2019 were selected in this study. The water use data of different industries were obtained from the Shenzhen Water Group. Furthermore, the gross domestic product (GDP), rainwater, sewage, and added value of various industries data were acquired from the public statistical data of Shenzhen Statistics Bureau. The floating population data were also procured from the Shenzhen Statistical Yearbook, and the real-time data of water supply were derived from the measured data of the Shenzhen Digital Water System.

Table 1 shows the statistical description of these variables. Water sources in Shenzhen include surface water, groundwater, and other water source projects, including rainwater utilization and sewage reuse. According to the measured data of the Shenzhen Digital Water System, the amount of water supply in 2019 was 2.06 billion m^3. Affected by salt tides, the amount of groundwater supply has been declining year by year, accounting for only 0.16% of the water supply. The surface water supply is 1.93 billion m^3, accounting for 93.74% of the water supply, and the water diversion is 1.76 billion m^3, accounting for 85.4% of the water supply. Additionally, the domestic water use, service industry water use, and industrial water use are 784 million m^3, 519 million m^3, and 471 million m^3, respectively. The amount of agricultural water use, construction industry water use, and ecological water use is small. Hence, the growth of water supply in Shenzhen is mainly embodied in domestic water use, service industry water use, and industrial water use. The coefficient of the variation of sewage

Table 1 | Variable description analysis

Variable	Minimum	Maximum	Mean	Standard deviation	Coefficient of variation
Rainwater utilization rate	0%	0.61%	0.28%	0.22%	0.79
Sewage reuse/10^4 m^3	37.0	10,811.29	6,890.06	3,694.76	0.54
GDP/10^2million	4,282.0	26,927.09	13,301.72	7,191.17	0.54
Sewage discharge/10^4 m^3	108,000.0	142,263.81	131,545.75	9,617.91	0.07
Domestic sewage/10^4 m^3	49,695.9	61,251.9	55,017.7	3,177.68	0.06
Secondary industry sewage/10^4 m^3	39,680.74	54,145.2	46,375.88	5,277.43	0.11
Tertiary industry sewage/10^4 m^3	4,158.9	41,877.45	30,152.17	11,684.5	0.39
Floating population/10^4people	470.11	799.12	627.41	100.67	0.16
Ecological water use/10^4 m^3	363.0	13,275.41	9,092.6	4,271.25	0.47
Industrial water use/10^4 m^3	47,062.1	62,021.0	54,002.64	4,937.03	0.09
Domestic water use/10^4 m^3	61,417.0	78,363.93	69,487.81	4,572.6	0.07
Agricultural water use/10^4 m^3	5,369.0	9,910.0	7,620.44	1,262.04	0.17
Construction industry water use/10^4 m^3	4,509.35	8,800.67	6,561.88	1,274.32	0.19
Service industry water use/10^4 m^3	29,547.42	52,911.33	41,900.86	8,214.97	0.20
Water use of the added value of the tertiary industry/10^4 m^3	23,112.45	52,911.33	40,412.57	10,382.83	0.26
Added value of the tertiary industry/10^2million	2,073.65	16,390.0	7,527.84	4,557.88	0.61
Added value of the secondary industry/10^2million	2,264.31	10,495.84	5,979.72	2,657.11	0.44
Water use of per 10^4 yuan of industrial added value/m^3	4.91	24.16	12.29	6.18	0.50
Water supply/10^4 m^3	160,068.0	206,195.88	188,658.54	13,381.7903	0.07

discharge, domestic sewage, industrial water use, domestic water use, and water supply are all less than 0.1, indicating that the discrete degree of the five variables is smaller. Nonetheless, the coefficient of the variation of rainwater utilization rate, sewage reuse, GDP, the added value of the tertiary industry, and the water use per 10^4 yuan of industrial added value are all greater than or equal to 0.5, which reveals that the discrete degree of the five variables is larger.

METHODS

ELRF algorithm

Ensemble learning (Nourani *et al.* 2018) is an effective method to increase the reliability of machine learning algorithms. A machine learning algorithm that includes only a single learning machine is weak and may generate partially correct analysis results, which can decrease its reliability. The core of ensemble learning is that the error result of a single learning machine will not affect the analysis results of most learning machines. Even if one learning machine generates wrong results, other learning machines can revise such errors so as to increase the reliability and robustness of the algorithm. Therefore, machine learning algorithms that include multiple learning machines are strong. This strong algorithm is equivalent to assembling the advantages of weak machine learning algorithms, and the final analysis results only contain correct results. Compared with traditional machine learning algorithms, ensemble learning machine learning algorithms are not only training single learning machine but also training multiple learning machines. The training set is weighted, and multiple learning machines are trained iteratively. The training process is divided into three steps.

First, multiple training sets are generated, and the weight initialization of the training set (N_i, $i = 1,2,\ldots,n$) is conducted. The first weak learning machine (LM_1) is trained on the basis of N_1, and the weight of LM_1 is updated according to its learning performance. The data with poor performance of LM_1 are recorded, and their weight is set higher in N_2 to make LM_2 pay more attention to these data.

Second, LM_2 is trained according to the knowledge of the previous step, and this iteration continues until the number of learning machines reaches the number set in the study. The number of learning machines in this study is set to 100.

Third, multiple learning machines are integrated into a machine learning algorithm on the basis of ensemble learning, and we can obtain a machine learning algorithm with a stronger generalization ability.

A random forest algorithm is a machine learning algorithm with higher accuracy and lower calculation loads, and this algorithm can process input samples with high dimensional features without reducing dimensions. In this study, an ELRF algorithm was developed for sensitive feature recognition. We hope that the ELRF algorithm can obtain more reliable results so that regression models can be constructed more efficiently. Figure 2 exhibits the framework of this study.

Model development

SARIMA

The SARIMA model (Tomczak *et al.* 1990; Moeeni *et al.* 2017; Elganainy & Eldwer 2018) is a more advanced statistical machine learning model, and this model has a better generalization ability than the ARIMA model and has higher accuracy for the prediction of single features. For the ARIMA model (Equation (1)), fewer parameters hinder the improvement of the generalization ability of the model. In addition to the three parameters, the SARIMA model also includes seasonal autoregressive, seasonal difference, seasonal moving average, and seasonal orders. The SARIMA model is also equivalent to constructing the ARIMA model twice. The specific steps are as follows:

First, the centralized sliding average (CSA) method is performed twice (Equations (2)–(5)) to eliminate periodicity, which is equivalent to constructing an ARIMA model on the period interval.

Second, the original time series is transformed into a time series without periodicity, and an ARIMA model is constructed again.

Therefore, the SARIMA model can learn laws that the ARIMA model finds difficult to learn; thus, the generalization ability of the SARIMA model is stronger. The SARIMA model is developed for the prediction of single features in this study.

$$y_t = I + \sum_{i=1}^{p} r_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \tag{1}$$

$$a_t = \frac{y_t + y_{t+1} + \cdots + y_{t+s_1-1}}{s_1},\ t \in \left(\frac{s_1}{2}, n - \frac{s_1}{2}\right) \tag{2}$$

$$TS_{CSA1} = [a_1, a_2, \cdots, a_l],\ l = n - s_1 \tag{3}$$

$$b_t = \frac{a_t + \cdots + a_{t+s_2-1}}{s_2},\ t \in \left(a_1, l - \frac{s_2}{2}\right) \tag{4}$$

$$TS_{CSA2} = \left[b_1, b_2, \cdots, b_{l-\frac{s_2}{2}}\right] \tag{5}$$

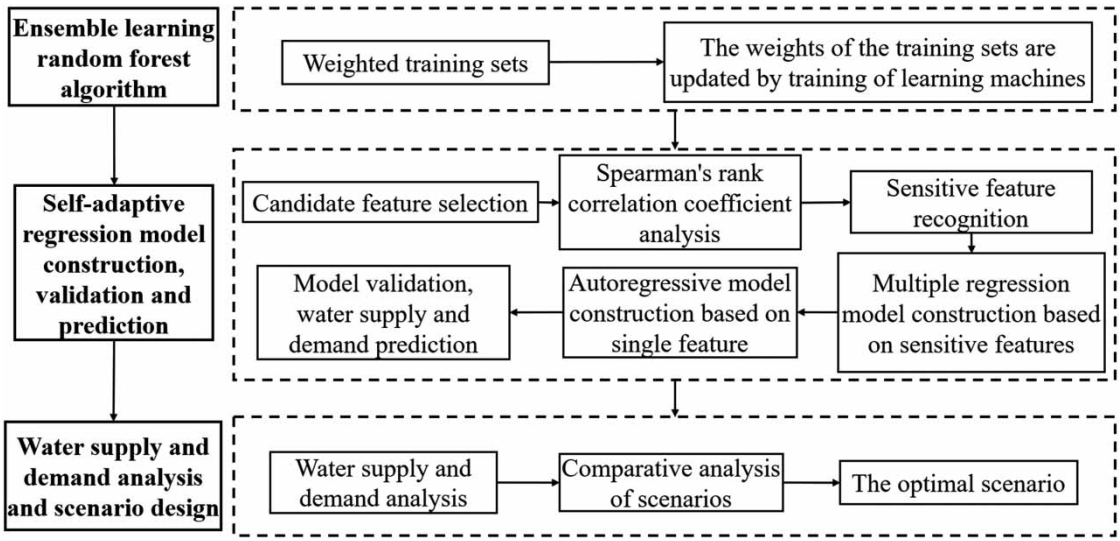


Figure 2 | Research framework.

where y, r, θ, and ϵ are the measured value, the autoregressive coefficient, the moving average coefficient, and the white noise at different times, respectively; I, p, q, t, and n are the intercept, the autoregressive order, the moving order, the time, and the length of the original time series, respectively; s_1 and s_2 are first-order and second-order CSA order; TS_{CAS1} and TS_{CSA2} are first-order and second-order CSA time series; a and b are the values of TS_{CAS1} and TS_{CSA2} at different times; and l is the length of TS_{CAS1}.

LASSOCV

The condition of multicollinearity among variables can be solved to apply multiple regression (Villarin 2019); however, the least absolute shrinkage and selection operator (LASSO) (Starn & Belitz 2018) model (Equation (6)) is a more advanced regression model that has a self-adaptive ability. The LASSO model can autonomously perform feature filtration and parameter punishment by regularization. Regularization is a common technique in machine learning, and the main purpose of this method is to control model complexity and reduce overfitting. Therefore, a more refined model can be finally obtained to solve the multicollinearity problem. The LASSO model forces the sum of the absolute value of the coefficients to be less than a fixed value (t). It compresses the large coefficients to be smaller and the smaller coefficients to 0 so that the features in the model do not have multicollinearity. Given that the LASSO model may excessively compress non-zero coefficients to obtain optimal parameters, the fivefold cross-validation (Figure 3) is integrated into the model.

$$LASSO = \min_{\theta} \left| y - \sum_{i=1}^{n} x_i \theta_i \right|^2 + \lambda \sum_{i=1}^{n} |\theta| \,, \; \sum_{i=1}^{n} |\theta_i| \le t \,, \; t > 0 \tag{6}$$

where y, t, and n are the dependent variable, the sum of the absolute value of parameters, and the number of independent variables, respectively; x is the independent variable; θ is the coefficient of x; and λ is the regularization factor.

RESULTS

Sensitivity analysis

In this study, Spearman's rank correlation coefficient (SRCC) (Zarei *et al.* 2016) (Equations (7) and (8)) is applied to filter out the features that have a weak correlation relationship with water supply. If the absolute value of SRCC between the feature and the water supply is less than 0.5, then the variable will be filtered out to reduce the calculation load of the ELRF algorithm. SRCC evaluates the monotone relationship of two variables, and it does not require prior knowledge. SRCC can accurately obtain the probability distribution of variables X and Y, which is more widely applicable than Pearson's correlation coefficient. SRCC can measure the dependence of two variables, and it is a nonparametric statistical method without any requirement on the distribution of variables.

$$d = \sum_{i=1}^{n} |R(X_i) - R(Y_i)|^2 \tag{7}$$

$$SRCC = 1 - \frac{6d}{n(n^2 - 1)} \tag{8}$$

where $R(X)$ and $R(Y)$ are the ranks of variables X and Y and n is the number of elements.

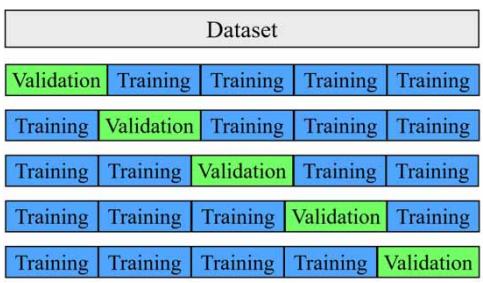


Figure 3 | Data set of fivefold cross-validation.

Table 2 shows the filtration results of SRCC. Industrial water use, agricultural water use, and construction industry water use are filtered out. This is because these three variables change little year after year, while the water supply increases year by year with an evidently increasing tendency; hence, the correlation relationship between these three variables and the water supply is weak. Thereafter, the ELRF algorithm is applied to analyze the sensitivity of the remaining variables, and five sensitive features, including the water use of the added value of the tertiary industry, floating population, domestic sewage, domestic water use, and rainwater utilization rate, are recognized.

The sensitive features are helpful to understand the change of water supply and demand. According to the measured data from the Shenzhen Digital Water System, the water supply decreases by approximately 10% during statutory holidays. The water supply during the Spring Festival is only 60% of normal levels, and the water supply returns to normal levels immediately after the holiday, which results from the floating population. The rapid growth of the floating population in Shenzhen every year will lead to the rapid growth of water use, which will lead to the growth of sewage discharge. Shenzhen has abundant rainfall, with a mean annual rainfall of 1,830 mm. If the rainwater utilization rate can be increased, then it can relieve the pressure of water supply to a certain extent. In 2019, the added value of the tertiary industry grew by 8.1%, showing a fast growth rate. Therefore, the water use of the added value of the tertiary industry is also highly sensitive to the amount of water supply in Shenzhen.

Model validation

In this study, the data from 2004 to 2014 are taken as the training set, and the data from 2015 to 2019 are taken as the validation set. The relative error (RE) (Equation (9)) between the measured and predicted values is selected as the evaluation criterion. Additionally, the LASSOCV model is constructed on the basis of sensitive features. In the process of modeling, the coefficients of domestic sewage, domestic water use, and rainwater utilization rate are compressed to 0. This is because the coefficient of rainwater utilization rate is the smallest, indicating that rainwater utilization rate has the least influence on water supply; thus, the coefficient of rainwater utilization rate is compressed to 0. The coefficients of domestic sewage and domestic water use are compressed to 0 because they have multicollinearity with the floating population. Nonetheless, the coefficient of the floating population is larger; hence, the coefficients of domestic sewage and domestic water use are also compressed to 0. The remaining variables include water use of the added value of the tertiary industry (x_1) and floating population (x_2) (Equation (10)). According to the coefficient of the equation, the more the water use of the added value of the tertiary industry and the more the floating population, the more the water supply (W). These results are consistent with the actual situation. Table 3 presents the validation results of LASSOCV model.

$$RE = \left| \frac{y_i - p_i}{y_i} \right| \tag{9}$$

$$W = 0.1 + 0.18x_1 + 0.64x_2 \tag{10}$$

where y and p are the measured and predicted values at different times.

Some major variables, such as the water use of the added value of the tertiary industry, floating population, domestic sewage, domestic water use, and water supply, are selected to validate the predicted results of the SARIMA model (Table 3).

Table 3 depicts that the RE of the predicted results is less than or equal to 0.07. Compared with the models in other researches (Wei *et al.* 2016; Sun *et al.* 2017; Li *et al.* 2019), the model proposed in this study shows a better generalization ability and can be applied for further prediction.

Table 2 | Filtration results of SRCC

Feature	SRCC
Industrial water use/10^4 m^3	−0.23
Agricultural water use/10^4 m^3	0.20
Construction industry water use/10^4 m^3	0.09

Table 3 | Model validation results

Feature	2015	2016	2017	2018	2019
Water use of the added value of the tertiary industry/10^4 m^3	0.06	0.06	0.01	0.07	0.03
Floating population/10^4people	0.02	0.01	0	0.03	0.05
Domestic sewage/10^4 m^3	0.04	0.04	0.02	0	0
Domestic water use/10^4 m^3	0.03	0.01	0	0	0.02
Water supply-SARIMA/10^4 m^3	0	0.01	0	0	0.01
Water supply-LASSOCV/10^4 m^3	0	0	0	0	0.01

Prediction

Water demand prediction and analysis

The main water use structure in Shenzhen includes industry, household, agriculture, the construction industry, the service industry, and ecology. Table 4 and Figure 4 display the water demand prediction results of the six features from 2021 to 2030.

According to Table 4, although the water demand of industries keeps a low speed decrease trend, the water demand is in third place. Water demand of households and the service industry continues to grow, and the water demand of the two features takes up the first and second places. Although the water demand of the construction industry and ecology continually grows, the growth rate is very small. The water demand of agriculture still fluctuates within a small value range.

From 2021 to 2030, the water demand in Shenzhen will increase by approximately 283 million m^3, and the increase of the water demand will mainly embody households and the service industry. In 2019, the water use proportion of households, the service industry, and industries accounted for 38%, 25.2%, and 22.8% of water supply. In 2030, the water demand proportions of households, the service industry, and industries will be 42.36%, 27.1%, and 14.2%, respectively. This shows that the water demand proportion of industries continues to decrease and that the water demand proportion of households and the service industry continues to rise. The water demand amount of households and the service industry will reach 866 million m^3 and 588 million m^3 in 2025 and 966 million m^3 and 638 million m^3 in 2030.

According to the sensitive features, the growth of water demand is mainly caused by the rapid growth of the tertiary industry and the floating population. In 2012, Shenzhen established a water supply red line indicator, and the industry replaced water-consumption equipment with water-saving equipment, and the industrial water use efficiency greatly improved. Therefore, industrial water use can continue to show a low speed decrease trend. Recently, citizen demands for ecological civilization and ecological environment have constantly been strengthened. In the meantime, with the increase of forestland areas in Shenzhen, the water demand of greening and landscape will continue to increase. However, Shenzhen has been promoting the construction of a sponge city, together with the construction of rainwater utilization and sewage reuse projects;

Table 4 | Water demand prediction results from 2021 to 2030 (10^4 m^3)

Year	Industry	Household	Agriculture	Construction industry	Service industry	Ecology	Water demand
2021	45,040.82	79,676.14	6,202.82	8,546.01	53,894.86	13,441.66	206,802.31
2022	43,813.99	82,833.64	5,633.60	9,819.30	52,004.05	15,303.68	209,408.26
2023	41,761.47	85,258.89	7,359.06	8,992.50	54,114.49	15,934.67	213,421.08
2024	41,849.56	85,764.22	7,063.4	9,564.64	57,176.86	15,469.94	216,888.62
2025	40,373.36	86,571.10	6,494.17	10,837.92	58,829.95	17,331.96	220,438.46
2026	38,250.47	89,728.60	8,219.64	10,011.12	57,683.09	17,962.95	221,855.87
2027	38,275.36	92,153.85	7,923.98	10583.26	60,506.76	17,498.21	226,941.42
2028	36,799.15	92,659.18	7,354.75	11,856.55	63,167.39	19,360.23	231,197.25
2029	34,676.26	93,466.06	9,080.22	11,029.75	64,972.24	19,991.22	233,215.75
2030	34,701.15	96,623.56	8,784.55	11,601.89	63,825.82	19,526.49	235,063.46

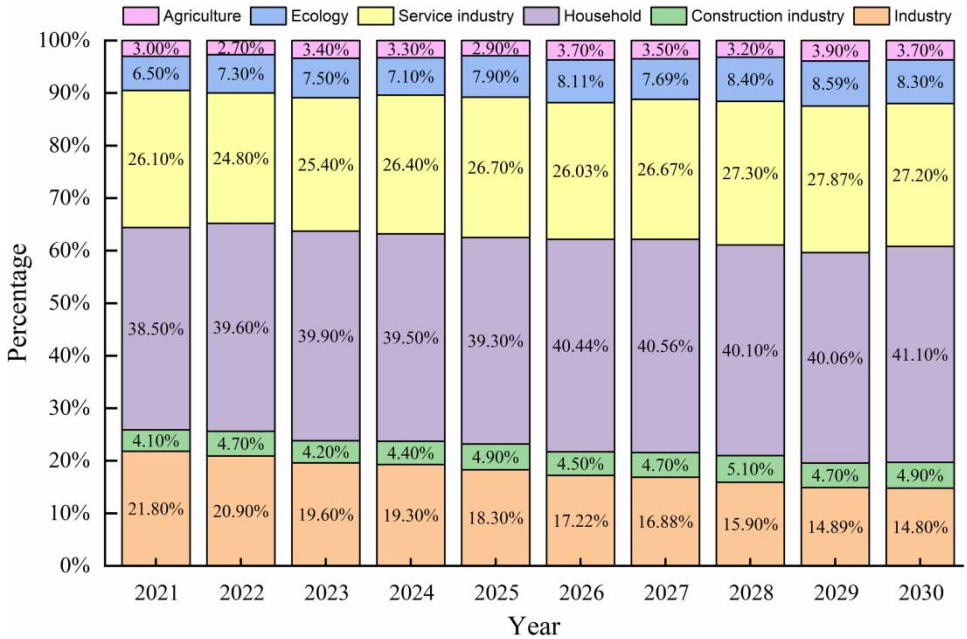


Figure 4 | Percentage of water demand in Shenzhen from 2021 to 2030.

hence, ecological water demand will not grow too fast. Shenzhen is a fully urbanized megacity, and the urban functions of Shenzhen are complete; thus, the water demand for the construction industry and agriculture will not increase greatly.

Water supply prediction and analysis

Table 5 presents the prediction results of surface water, sewage reuse, rainwater, and groundwater. The water supply will increase quickly from 2021 to 2030, which will exceed 2.2 billion m^3 in 2025 and 2.35 billion m^3 in 2030. The surface water is still the main water source of Shenzhen. A large amount of domestic water use and service industry water use has produced a large amount of sewage; therefore, sewage reuse is still the second major water source of Shenzhen. Meanwhile, if the rainwater utilization rate can be increased, then the water supply pressure of Shenzhen will be relieved. Groundwater supply is very small due to the influence of salt tide.

Table 6 shows the prediction results of water supply in Shenzhen according to the SARIMA and LASSOCV models from 2021 to 2030. The RE of the predicted results of the SARIMA and LASSOCV models is small, which reveals that the predicted results of the LASSOCV model correspond to the potential quantitative relationship of historical water supply data. This shows that the LASSOCV model constructed in this study can solve the problem of multicollinearity.

Table 5 | Water supply prediction results of water sources from 2021 to 2030 (10^4 m^3)

Year	Surface water	Sewage reuse	Rainwater	Groundwater
2021	198,958.67	11,722.66	1,941.57	365.02
2022	201,182.63	12,605.65	1,874.11	222.50
2023	202,656.52	13,565.93	2,181.41	386.51
2024	204,673.41	13,517.03	1,979.97	337.15
2025	209,999.38	14,400.02	1,722.56	213.21
2026	210,288.67	15,360.29	2,462.28	367.18
2027	212,512.63	15,311.39	1,907.10	323.24
2028	213,986.52	16,194.38	2,172.57	196.37
2029	216,003.41	17,154.66	1,957.16	351.92
2030	221,329.38	17,105.76	1,773.96	307.13

Table 6 | Predicted result comparison of water supply from 2021 to 2030

Year	SARIMA/10^4 m^3	LASSOCV/10^4 m^3	RE
2021	212,987.92	210,038.46	0.01
2022	215,884.89	213,261.79	0.01
2023	218,790.37	215,722.81	0.01
2024	220,507.56	219,435.95	0.00
2025	226,335.17	221,506.50	0.02
2026	228,478.42	225,117.56	0.01
2027	230,054.36	226,681.55	0.01
2028	232,549.84	230,394.69	0.01
2029	235,467.15	232,465.24	0.01
2030	240,516.23	236,076.30	0.02

Attracted by the economy and policy, many people will flock to Shenzhen. While floating populations bring sufficient labor force to Shenzhen, they also cause a large growth of water demand. In the meantime, the rapid growth of the population and the tertiary industry must lead to the growth of sewage discharge. In 2019, the sewage discharge in Shenzhen was 1.42 billion m^3, and the sewage reuse was only 0.1 billion m^3. The rainfall in Shenzhen in 2019 was nearly 1,866.56 mm, which is equivalent to 3.48 billion m^3 water. Therefore, Shenzhen should pay attention to sewage reuse and rainwater utilization. These water resources can be used for greening, road water spray, and toilet flushing, which can save the consumption of surface water to some extent. The concept of water saving should be continuously publicized to enhance the awareness of water saving, which is also conducive to relieving the water supply pressure of Shenzhen.

DISCUSSION

The results of different algorithms and models may be different. For the exhibition of the superiority of the algorithm and model developed in this study, the decision tree (DT) and regular random forest (RRF) algorithms are developed to compare their results with the results of the ELRF algorithm. We calculate the sensitivity score of these sensitive features on the basis of the idea of permutation (Breiman 2001). The calculation process is divided into the following three steps:

Step1: We employ these sensitive features to fit the water supply, and the Nash–Sutcliffe efficiency (NSE) (Equation (11)) coefficient is selected as the evaluation criterion to calculate the initial score.

$$NSE = \left\{1 - \left[\frac{MSE(\boldsymbol{M}, \boldsymbol{P})}{var(\boldsymbol{M})}\right]\right\} \times 100\% \tag{11}$$

where M and P are the measured and predicted value vectors; MSE is the mean square error between M and P; and var is the variance of measured values.

Step2: For each feature, we set five rounds of iterations. The data of the feature are randomly shuffled to generate new data with noise, and the NSE of each fitting result is calculated. Subsequently, the mean NSE of five fitting results can be obtained.

Step3: The gap between the initial NSE and the mean NSE is utilized as the sensitivity score.

The strategy of the calculation is that if a feature is not sensitive to water supply, then the data of this feature will not have a great influence on the fitting result. Conversely, when the data of the sensitive features are no longer accurate, there is a great influence on the fitting result. Table 7 presents the sensitive features recognized by the three machine learning algorithms and their scores.

Table 7 shows that the DT and RRF algorithms recognize the floating population and the rainwater utilization rate, which is consistent with the results of the ELRF algorithm. However, the DT and RRF algorithms recognize the ecological water use. According to the measured data, the ecological water use in 2006 was 363 10^4 m^3, and the ecological water use in 2007 was 8,849 10^4 m^3, which made ecological water use have a large coefficient of variation. In 2019, the amount of ecological water

Table 7 | Sensitive feature recognition results of three machine learning algorithms

DT	Score	RRF	Score	ELRF	Score
Ecological water use/10^4 m^3	0.16	Water use of per 10^4 yuan of industrial added value/m^3	0.16	Water use of the added value of the tertiary industry/10^4 m^3	0.09
Rainwater utilization rate	0.15	Floating population/10^4 people	0.10	Floating population/10^4 people	0.08
Added value of the secondary industry/10^2 million	0.11	Ecological water use/10^4 m^3	0.10	Domestic sewage/10^4 m^3	0.08
Floating population/10^4 people	0.10	Rainwater utilization rate	0.09	Domestic water use/10^4 m^3	0.08
GDP/10^2 million	0.09	Added value of the secondary industry/10^2 million	0.08	Rainwater utilization rate	0.07

use was 13,275.41 10^4 m^3; thus, the growth of ecological water use in the past 13 years was very small. Moreover, in the same year, the amount of sewage reuse in Shenzhen was 10,811.29 10^4 m^3, and the amount of rainwater utilization was 880.79 10^4 m^3. Shenzhen has been promoting the construction of a sponge city and has been paying attention to the capacity of sewage treatment and reuse. In addition, the growth of water supply mainly embodies the domestic water use, service industry water use, and industrial water use. Therefore, the ecological water use is not sensitive to water supply. On the basis of the above facts, it can be found that although the DT and RRF algorithms recognize the two correct features, these two algorithms also recognize the wrong feature. Only including a single learning machine is a limitation of algorithms. As we stated above, if an algorithm has multiple learning machines, other learning machines have a possibility to correct errors. These results prove the superiority of this research algorithm.

Among the multiple regression models, the support vector regression (SVR) and stepwise regression (SR) (Cáceres *et al.* 2018) models have better performance. The SVR and SR models are developed to construct multiple regression models on the basis of the sensitive features recognized by the ELRF algorithm. Table 8 shows the modeling results of the three models. The coefficients of rainwater utilization rates in SVR and SR are positive, indicating that the higher rainwater utilization rate, the more the water supply. Therefore, the results of the SVR and SR models are not consistent with the facts, and the LASSOCV model is superior to the SVR and SR models.

Meanwhile, the ARIMA model is also developed to conduct the validation of the single feature, including floating population, domestic sewage, domestic water use, and water supply. Table 9 exhibits the RE between the measured values and predicted results. The prediction accuracy of the two models is different in distinct years; thus, we use the mean RE of the five years as the evaluation criterion to compare the generalization ability of the two models. Evidently, the generalization ability of the SARIMA model is superior to that of the ARIMA model.

By the end of 2019, the permanent population of Shenzhen comprised more than 13 million people, and the floating population exceeded 8 million people. According to the above analysis, the fluctuation of water supply was caused by the floating population. Affected by the Corona Virus Disease 2019 (COVID-19), citizens were quarantined at home and unable to return to Shenzhen, and water supply in Shenzhen decreased significantly. After the nationwide quarantine was lifted in May 2020, many citizens chose to find new jobs in Shenzhen, resulting in a clear increase in the floating population of Shenzhen. In July 2020, the water supply in Shenzhen increased by 13.85% year on year with a large increase. At the same time, 2020 is a special low water year, with an annual rainfall of roughly 1,580 mm. Consequently, all pumping stations along the water diversion projects are running at full load, some waterworks are running at overload, and the water level of most water supply reservoirs is still falling slowly. The water level of reservoirs could not reach the water storage target; hence, the overhauling plan for the water diversion projects in 2020 had to be terminated.

Table 8 | Modeling results of three models

Model	Parameters of regression equation/{intercept, coefficients of five features}
SVR	{0.11, 0.35, 0, 0.42, 0, 0.25}
SR	{ − 0.02, 0.27, 0.24, 0.23, 0.22, 0.08}
LASSOCV	{0.1, 0.18, 0.64, 0, 0, 0}

Table 9 | Comparison of the predicted result of the ARIMA and SARIMA models

	Floating population/ 10^4people		Domestic sewage/10^4 m^3		Domestic water use/ 10^4 m^3		Water supply/10^4 m^3	
Year	ARIMA	SARIMA	ARIMA	SARIMA	ARIMA	SARIMA	ARIMA	SARIMA
2015	0.01	0.02	0.02	0.04	0	0.03	0.01	0
2016	0.03	0.01	0.02	0.04	0	0.01	0	0.01
2017	0.05	0	0	0.02	0	0	0.01	0
2018	0.06	0.03	0.03	0	0.04	0	0.02	0
2019	0.06	0.05	0.05	0	0.06	0.02	0.02	0.01
Mean/%	4.2	2.2	2.4	2	2	1.2	1.2	0.4

In 2019, the total amount of water diversion in Shenzhen was 1.76 billion m^3. The quota of the DBWD has all been run out, and there is still approximately 140 million m^3 left in the quota of the DSWD. Therefore, the annual water diversion of Shenzhen can reach nearly 1.9 billion m^3 at most; thus, the water supply in Shenzhen in the future can reach approximately 2.2 billion m^3. According to the predicted results of the model, the water demand in Shenzhen will reach nearly 2.2 billion m^3 in 2025 and 2.35 billion m^3 in 2030, which has already exceeded the water supply ability of Shenzhen. The Shenzhen government has been encouraging citizens to save water through some campaigns and propaganda, and several advertisements and posters have been put up on billboards in streets and restaurants. The government has been working to raise public awareness to decrease water consumption. However, Shenzhen possesses developed high-technology and emerging industries, and the economic increment in Shenzhen is dominated by advanced manufacturing, emerging, and modern service industries. Although these measures have indeed played a role in saving water, the annual amount of water supply is still growing rapidly. Therefore, some necessary measures must be taken to keep water resources shortage from becoming an obstacle to the development of Shenzhen.

Water supply in Shenzhen mainly depends on water diversion, and the above facts bring a new challenge to the water supply security in Shenzhen. Shenzhen will need to increase the quota of its existing water diversion projects or add another new water diversion project to meet the balance of water supply and demand in the future.

The pressure of water supply in Shenzhen is mainly concentrated in the western part of Shenzhen. Guangming District, Baoan District, Nanshan District, Longhua District, Futian District, and Luohu District (Figure 5) account for 76% of the water supply in Shenzhen. Accordingly, three scenarios are set with different objectives as follows:

Scenario I: The quota of the DBWD project can be increased. The water is transferred from the eastern part of Shenzhen to the western part of Shenzhen, and more water can be stored in the reservoirs to ensure the water supply of the eastern, central, and western parts of Shenzhen.

Scenario II: The quota of the DSWD project can be increased. The DSWD project supplies water not only to Shenzhen but also to Hong Kong and other cities along the project. The quota of the DSWD project is increased to ensure the water supply in Shenzhen, Hong Kong, and other cities along the project. It will also allow more water to be transferred to reservoirs and waterworks in the western part of Shenzhen.

Scenario III: Presently, the water supply pressure of Shenzhen is mainly concentrated in the west. Consequently, a new water diversion project can be constructed in the west to directly transfer water to the reservoirs and waterworks in the western part of Shenzhen. Guangming District has a large reservoir, the Gongming Reservoir, with a total storage of 139 million m^3, which can be employed as the main water storage reservoir. In addition, the high elevation of the Gongming Reservoir is higher than those reservoirs in Boan District and Nanshan District, and the water can flow by gravity to other reservoirs (see the red arrow in Figure 5). This not only has a lower cost but also can effectively reduce the risk of pipe explosion and the loss of evaporation leakage in long distance water transfer. It would relieve pressure on two existing water diversion projects.

In terms of water supply security and economic operation, Scenario III is the best one. In Scenario I, the DBWD is far away, and the water intake along the project requires pumping stations. Although more quota of the DBWD project can solve the current water resources shortage, more water intake means more electricity costs, which is bad for the economic

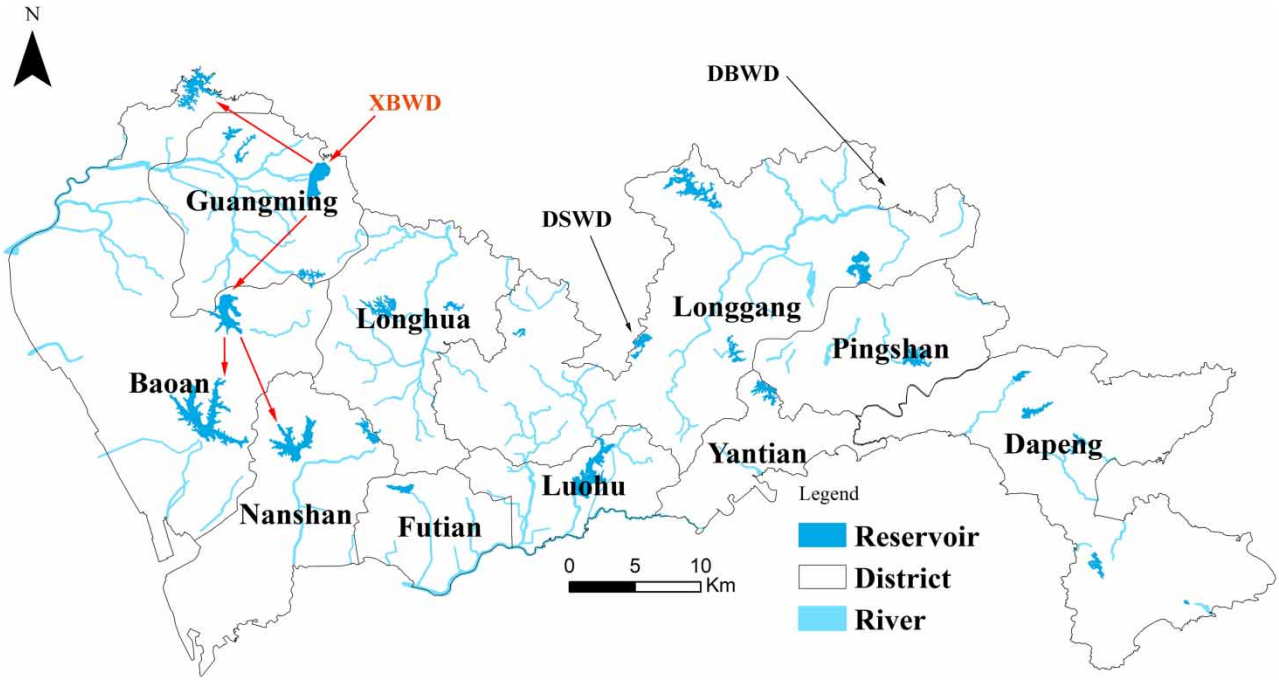


Figure 5 | Thumbnail of an optimal water diversion scenario.

operation in the long run. Currently, the quota of the DSWD has not been exhausted, and more water cannot be pumped because of the design scale of pumping stations, and the flow capacity limitation of pipelines. Therefore, expanding the scale of pumping stations and the flow capacity of pipelines and considering the water supply from other cities is necessary. The project is complicated and may affect the water supply in Hong Kong and other cities. Compared with Scenarios I and II, Scenario III can not only ensure the water supply in the western part of Shenzhen but also can transfer more water to the central and eastern parts of the city to form a pattern of bidirectional complementary water supply. The three water diversion projects can be coordinated to meet the balance of the supply and demand of Shenzhen to ensure that the backup water source of the city can meet the target of water supply for 90 days. In this way, the reservoir can easily meet the water storage target. No matter which water diversion project needs to be overhauled, the other two projects can easily meet the water supply in Shenzhen. Thus, the problems of water resources shortage and water diversion project overhauling can be completely solved, and the current embarrassing situation of insufficient backup water source can be changed.

CONCLUSIONS

In this study, SRCC is utilized to filter out features that are not strongly correlated with water supply, and the ELRF algorithm, including multiple learning machines, is proposed to conduct sensitive feature recognition. The self-adaptive regression model coupling the SARIMA and LASSOCV models is developed. The sensitive features are recognized by the ELRF algorithm, which is helpful for us to understand the change of water supply and demand. The ELRF algorithm can recognize sensitive features more effectively, and model validation results validate that the generalization ability of the model developed in this study is better. At the same time, the predicted results of the LASSOCV model correspond to the potential quantitative relationship of historical water supply, indicating that this model can solve the problem of multicollinearity.

According to the predicted results of water supply and demand in Shenzhen, the water demand in Shenzhen is increasing quickly. The water demand of agriculture is small, the water demand of ecology and the construction industry is basically stable, and the industrial water demand is decreasing at a low speed. With the rapid growth of the floating population and the tertiary industry, the water demand of household and the service industry keep growing, and the proportion accounting for the water demand is rising slowly. In the future, the main water supply sources of Shenzhen will still be surface water, sewage reuse, and rainwater. Therefore, Shenzhen should increase sewage reuse and rainwater utilization rate to reduce the consumption of surface water. Shenzhen must also pay attention to the growth of the floating population, continue to

enhance the water saving awareness of citizens, and improve efficiency of water use. These measures can reduce the consumption of surface water to a certain extent, and the water supply dispatching can be efficiently performed.

The water supply in Shenzhen will exceed 2.2 billion m^3 in 2025 and exceed 2.35 billion m^3 in 2030; hence, Shenzhen will face serious water resources shortage in the future. According to the current situation of Shenzhen, three scenarios are proposed to solve the balance of water supply and demand in Shenzhen. In terms of water supply security and economic operation, the scenario to build a new water diversion project in the west of Shenzhen (Scenario III) is the optimal solution. The scenario can achieve a bidirectionally complementary water supply pattern and completely solve the water resources shortage in Shenzhen to keep water resources shortage from becoming an obstacle to the development of Shenzhen.

ACKNOWLEDGEMENTS

This study was supported by the Scientific Research Projects of IWHR (01882103, 01882104), China Three Gorges Corporation Research Project (Contract No: 202103044), and the National Natural Science Foundation of China (51679089; 51709107).

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

Avanzi, F., Johnson, R. C., Oroza, C. A., Hirashima, H., Maurer, T. & Yamaguchi, S. 2019 Insights into preferential flow snowpack runoff using random forest. *Water Resources Research* **55** (12), 10727–10746.

Bai, Y., Bezak, N., Zeng, B., Li, C., Sapač, K. & Zhang, J. 2021 Daily runoff forecasting using a cascade long short-term memory model that considers different variables. *Water Resources Management* **35** (4), 1167–1181.

Bassiouni, M., Vogel, R. M. & Archfield, S. A. 2016 Panel regressions to estimate low-flow response to rainfall variability in ungaged basins. *Water Resources Research* **52** (12), 9470–9494.

Breiman, L. 2001 Random forests. *Machine Learning* **45** (1), 5–32.

Cáceres, L., Méndez, D., Fernández, J. & Marcé, R. 2018 From end-of-pipe to nature based solutions: a simple statistical tool for maximizing the ecosystem services provided by reservoirs for drinking water treatment. *Water Resources Management* **32** (4), 1307–1323.

Chen, G., Long, T., Bai, Y. & Zhang, J. 2019 A forecasting framework based on Kalman filter integrated multivariate local polynomial regression: application to urban water demand. *Neural Processing Letters* **50** (1), 497–513.

Cross, D., Onof, C. & Winter, H. 2020 Ensemble estimation of future rainfall extremes with temperature dependent censored simulation. *Advances in Water Resources* **136**, 103479.

Ebtehaj, I. & Bonakdari, H. 2016 A support vector regression-firefly algorithm-based model for limiting velocity prediction in sewer pipes. *Water Science and Technology* **73** (9), 2244–2250.

Elganainy, M. A. & Eldwer, A. E. 2018 Stochastic forecasting models of the monthly streamflow for the Blue Nile at Eldiem Station. *Water Resources* **45** (3), 326–337.

Kroll, C. N. & Song, P. 2013 Impact of multicollinearity on small sample hydrologic regression models. *Water Resources Research* **49** (6), 3756–3769.

Li, T., Yang, S. & Tan, M. 2019 Simulation and optimization of water supply and demand balance in Shenzhen: a system dynamics approach. *Journal of Cleaner Production* **207**, 882–893.

Moeeni, H., Bonakdari, H. & Ebtehaj, I. 2017 Integrated SARIMA with neuro-fuzzy systems and neural networks for monthly inflow prediction. *Water Resources Management* **31** (7), 2141–2156.

Nasser, A. A., Rashad, M. Z. & Hussein, S. E. 2020 A two-layer water demand prediction system in urban areas based on micro-services and LSTM neural networks. *IEEE Access* **8**, 147647–147661.

Nguyen, X. H. 2020 Combining statistical machine learning models with ARIMA for water level forecasting: the case of the Red river. *Advances in Water Resources* **142**, 103656.

Nourani, V., Elkiran, G. & Abba, S. I. 2018 Wastewater treatment plant performance analysis using artificial intelligence – an ensemble approach. *Water Science and Technology* **78** (10), 2064–2076.

Praveen, B., Talukdar, S., Mahato, S., Mondal, J., Sharma, P., Islam, A. R. M. T. & Rahman, A. 2020 Analyzing trend and forecasting of rainfall changes in India using non-parametrical and machine learning approaches. *Scientific Reports* **10** (1), 1–21.

Reis Jr., D. S., Veilleux, A. G., Lamontagne, J. R., Stedinger, J. R. & Martins, E. S. 2020 Operational Bayesian GLS regression for regional hydrologic analyses. *Water Resources Research* **56**(8), e2019WR026940.

Safari, M. J. S. 2019 Decision tree (DT), generalized regression neural network (GR) and multivariate adaptive regression splines (MARS) models for sediment transport in sewer pipes. *Water Science and Technology* **79** (6), 1113–1122.

Starn, J. J. & Belitz, K. 2018 Regionalization of groundwater residence time using metamodeling. *Water Resources Research* **54** (9), 6357–6373.

Sun, Y., Liu, N., Shang, J. & Zhang, J. 2017 Sustainable utilization of water resources in China: a system dynamics model. *Journal of Cleaner Production* **142**, 613–625.

Tomczak, M., Richard, A., Perrin, G. & Robert, M. 1990 Sarima modeling of an irrigation system. In: *Advanced Information Processing in Automatic Control (AIPAC'89)*, Pergamon, pp. 371–374.

Villarín, M. C. 2019 Methodology based on fine spatial scale and preliminary clustering to improve multivariate linear regression analysis of domestic water consumption. *Applied Geography* **103**, 22–39.

Wei, T., Lou, I., Yang, Z. & Li, Y. 2016 A system dynamics urban water management model for Macau, China. *Journal of Environmental Sciences* **50**, 117–126.

Yang, P. & Ng, T. L. 2019 Fast Bayesian regression kriging method for real-time merging of radar, rain gauge, and crowdsourced rainfall data. *Water Resources Research* **55** (4), 3194–3214.

Zandi, J., Ghazvinei, P. T., Hashim, R., Yusof, K. B. W., Ariffin, J. & Motamedi, S. 2016 Mapping of regional potential groundwater springs using logistic regression statistical method. *Water Resources* **43** (1), 48–57.

Zarei, A. R., Moghimi, M. M. & Mahmoudi, M. R. 2016 Parametric and non-parametric trend of drought in arid and semi-arid regions using RDI index. *Water Resources Management* **30** (14), 5479–5500.

First received 21 April 2021; accepted in revised form 10 August 2021. Available online 23 August 2021