Water Supply

Water Supply Vol 22 No 8, 6833 doi: 10.2166/ws.2022.243

Evaluating the performance of feature selection techniques and machine learning algorithms on future residential water demand

Marziyeh Pourmousavi[a,†], Hossein Nasrollahi[b,†], Abdolhamid Amirkaveh Najafabadi[c] and Ahmad Kalhor[d,*]

[a] Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan 8415683111, Iran
[b] Department of Energy Systems, Mechanical Engineering Faculty, K. N. Toosi University of Technology, Tehran 158754416, Iran
[c] Department of Engineering Science, College of Engineering, University of Tehran, Tehran 1417935840, Iran
[d] School of Electrical and Computer Engineering, University of Tehran, Tehran 1417935840, Iran
*Corresponding author. E-mail: akalhor@ut.ac.ir
[†]Authors contributed equally to this work.

ABSTRACT

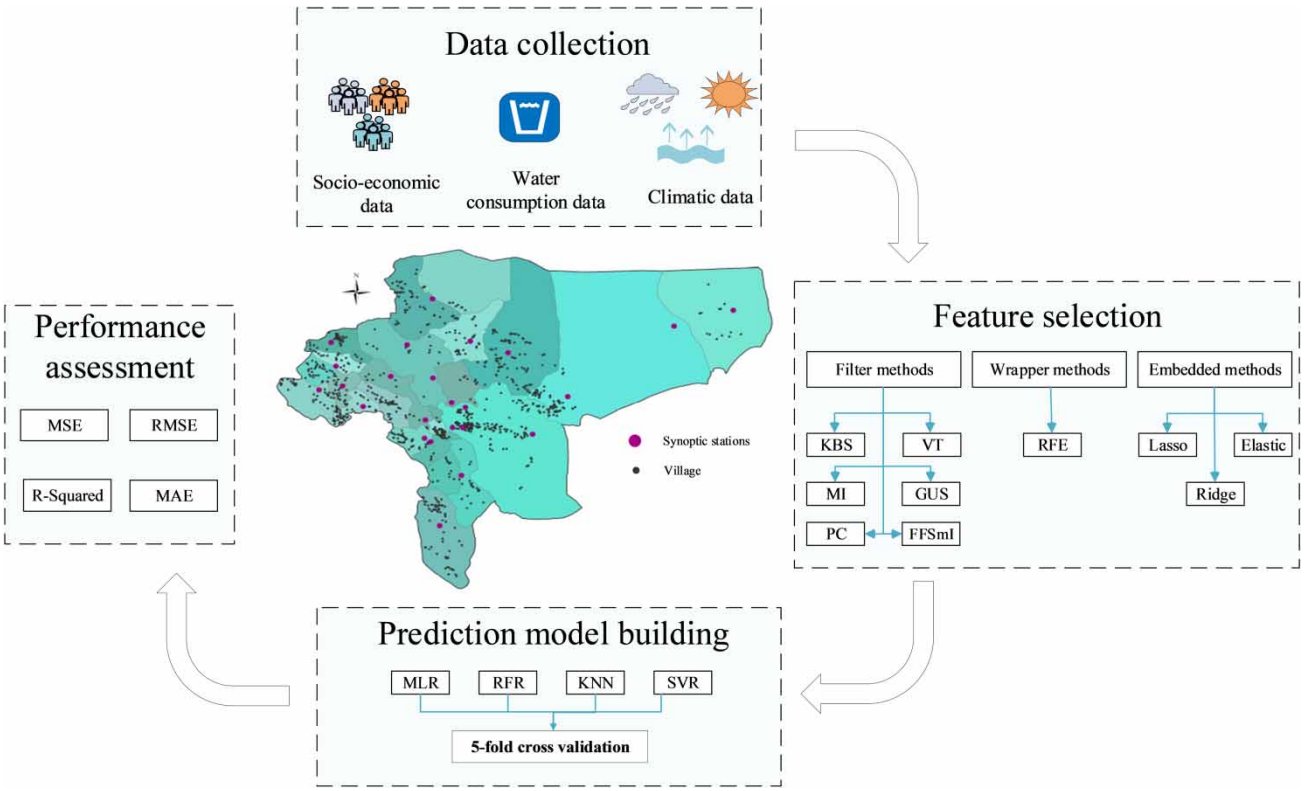
The provision of potable water as a severe challenge has engaged many people worldwide. So, identifying influential factors in water demand forecasting (WDF) for the residential sector performs a vital role in water crisis management. Nowadays, long-term macro-planning for vast geographical areas helps policymakers to achieve sustainable development goals. This study uses the same perspective to present a pattern for water consumption behavior and prediction. For this purpose, yearly residential water consumption data, along with climatic characteristics, and socioeconomic factors of rural areas of Isfahan, Iran, are aggregated. The feature selection task is conducted on the collected data using various machine learning (ML) methods along with a novel approach, forward selection based on smoothness index (FSSmI). Posterior to selecting features influencing residential water demand (WD), the raw data are analyzed using regression techniques, including multiple linear regression, support vector regression, and random forest regression. The employed methods show an improvement in the feature selection procedure and coefficient of determination as a result of implementing the FSSmI method. Based on the results, multiple linear regression and support vector regression gain 96% and 95% accuracy and less than 11% and 13% error respectively; it demonstrates the validity of forecasting methods.

Key words: feature selection, machine learning, mutual information, regression model, water demand forecasting

HIGHLIGHTS

- Machine learning approach is applied to predict residential long-term water demand for a new dataset.
- A novel feature selection method is developed based on the smoothness index criterion and forward selection.
- Testing and comparing machine-learning-based methods are carried out for feature selection and prediction.
- The proposed improvements resulted in better water demand prediction performance.

GRAPHICAL ABSTRACT



INTRODUCTION

Water is considered one of the most important primary resources of any settlements, whether village, town, city, or superci-ties. Water shortage as a common problem can be tackled by expensive strategies such as dams and desalination plants. However, limitations such as high cost and environmental impacts have encouraged decision-makers to search for more sustainable solutions. The novel understanding of the management of water resources has brought about a new paradigm in water supply and demand management and incorporates new policy frameworks for administration. Integrated water resource management has always been a complex issue due to the presence of multiple variables such as climate change, rapid population growth, and urbanization. As water resource management mainly concerns competing water demands and allocating equitable bases to satisfy them, water demand forecasting (WDF) has become an essential tool for optimal scheduling of water resource management (Villarin & Rodriguez-Galiano 2019).

Traditionally, calculating water demand (WD) for different applications, including potable, agricultural, industrial, and ecological sectors, relies on considering socioeconomic (including population, gross domestic product, income, education, water price, family or household size) and environmental (including precipitation, temperature, wind, evapotranspiration) changes.

The estimation of WD based on lots of involving parameters, conditions, factors, and uncertainties is challenging for decision-makers. Modeling WD is a complex exercise but vital for utilities to meet consumer demand while managing the available water resources. Determining the impact of changes in any socioeconomic or environmental parameters on WD is crucial for implementing a range of strategies to ensure the water security of urban and rural residents. Thereby water authorities will have practical and realistic investment plans for the future (Makki *et al.* 2015).

Water utilities use WDF for various types of purposes, whether for design, planning, operation, or management. Utilities' perspectives, objectives, and planning levels are reflected in the forecast horizon, which is the temporal resolution of modeling and forecasting (Nasseri *et al.* 2011). Although there are no generally accepted time-frames for these horizons, Nasseri *et al.* (2011) recommended three classes of time-frames: (1) long-term forecasting, which has a resolution of one year or

greater, deals with large-scale planning and management; (2) middle-term forecasting for middle-time management, which has a resolution equal to or greater than one month; (3) short-term forecasting, which has a resolution of less than one month, even to one hour, and concerns low-scale planning and management. Thus, reviewing the literature shows that the forecast span is related to management perspective and goal; in addition, it determines the method to some extent (Donkor *et al.* 2014).

Forecasting, although it is not a new subject, for its multi-dimensional nature, quality of available data, and uncertainties, is still considered a scientific gap. Various proposed methods and models attempt to improve forecasts despite these conditions. Forecasting with qualitative methods, unit WD analysis, several regression methods (such as multiple regression, piecewise linear regression, and geographically weighted regression), moving average, Bayesian maximum entropy, artificial neural networks, time series, and scenario-based and composite forecasts are examples of forecasting methods (Donkor *et al.* 2014). The choice of forecasting method depends on the available data and technical–analytical complexities such as modeling strategy and criteria selection.

WDF methods are categorized into four classes: temporal extrapolation, unit WD, end-use forecasting, and multivariate statistical model. These methods are described in Table 1.

Regression methods as a type of multivariate statistical model were split in two categories: parametric regression and non-parametric regression. The former is considered a special function for the relationship between independent and dependent variables, where the type of relationship is also known. Fitting only one equation to a large number of data that represent complex behavior leads to undesirable results (Cleveland *et al.* 1988). This method also requires time-consuming calculations. Schleich & Hillenbrand (2009) studied about 600 water supply areas of Germany to investigate the impact of several

Table 1 | WDF methods comparison

WDF method	Description	Strengths	Weaknesses
Temporal extrapolation or time-series extrapolation models	These models include simple time trends, exponential smoothing, and moving-average. They are based on the claim that WDF can be conducted using historical water consumption data (Lee *et al.* 2010).	Rapid execution and wide application when major changes are not expected and causal factors are expected to remain constant (Glantz & Mun 2011).	Not appropriate for long-term forecasting as it does not take into account possible future changes (Donkor *et al.* 2014).
Unit WD	Forecasting future demand based on the number of water subscribers, and by relying on the per capita consumption of water subscribers and calculating their unit of water consumption coefficients (such as residential, commercial, and industrial) WD is estimated (Donkor *et al.* 2014).	Although it is not feature-extensive and all it requires is to estimate consumption per unit of a customer category and number of units of those categories, it is the simplest approach used by most water utilities (Billings & Jones 2008).	Not appropriate for long-term forecasting as it does not take into account possible future changes (Donkor *et al.* 2014).
Micro-component or end-use forecasting	Considering all places where water is used, such as behavioral patterns (shower, bathtub, bathroom, faucets, and toilet) and data obtained from water-consuming appliances (washing machine and dishwasher) (Levin *et al.* 2006).	Simulating the long-term effects of technological change on WD management (Levin *et al.* 2006).	The entire detailed information of each subscriber is required.
Multivariate statistical models	In order to consider the impact of socioeconomic (water price, income, population, urban density, and etc.) and climatic (precipitation, temperature, humidity, wind, etc.) factors on WD, these models can be used (Donkor *et al.* 2014).	A powerful method as it examines the effect of a set of independent variables on per capita consumption (dependent variable) and shows a statistical relationship between them (Fullerton & Molina 2010).	The performance of these models rely on the volume of the dataset (Kindler & Russell 1984).

economic, environmental and social factors (such as price, income, household size, age, rainfall, and temperature) on per capita WD using the ordinary least squares method. According to the results, both rainfall pattern and household size have a negative relationship with WD. On the other hand, age has a positive relationship with WD.

In non-parametric regression, the regressor does not take a predetermined functional form; i.e., there are no assumptions about the distribution of data (Cleveland *et al.* 1988). This method divides the data into several regions to fit a curve to the data in each region, then turns them into appropriate and useful estimators. Eslamian *et al.* (2016), based on climate parameters and socioeconomic factors, predicted the daily WD of Quebec City in Canada using the multiple regression method. According to the results, WD decreases on rainy days and increases on weekends. Other non-parametric methods are neural networks and fuzzy logic. In these estimation methods, learning relationships between variables in the model leads to a prediction. These methods are more challenging to use and hard to explain to potential users, which prevents many water analysts from using them (Billings & Jones 2008).

On the same mathematical and theoretical basis, machine learning (ML) techniques are useful in many of the previously mentioned methods. ML applies to a wide variety of parts of the water sector, from forecasting the future WD and early flood warning to monitoring and alarm handling of distribution systems, identifying factors and features that influence actions in related systems, determining leakage in the water supply, optimizing management, and classifying types of consumers (Villarin & Rodriguez-Galiano 2019).

In this study, a framework for forecasting residential WD is developed based on various ML methods, aiming to improve previous methods and suggest a novel approach. A new feature selection method is proposed in this study named forward selection based on smoothness index (FSSmI).

The FSSmI method is modified to fill the following gaps of previous research:

- Some of the feature selection methods have the constraint of linear dependency, such as the Pearson correlation coefficient which measures the linear dependency between two random variables.
- Some of the feature selection methods are based on estimating the probability density function, such as mutual information. In this method, the extent of difference between the joint distribution of (X, Y) and product of marginal distribution of X and Y is determined.
- Some of the feature selection methods are dependent on the error of the regression model, such as recursive feature elimination. In this method, the set of features selected based on the error of the regression model determines the importance of each feature.

The developed FSSmI method has provided a short run time and straight framework for feature selection, which addresses the abovementioned problems. This method along with other methods such as principal component analysis (PCA), generic uni-variant selection (GUS), recursive feature elimination (RFE), and mutual information (MI) are implemented to determine the most profitable features. Thereby, this phase reduces problem dimensions and computational costs, resulting in a more feasible forecasting procedure. Afterward, multiple ML methods, such as support vector regression (SVR), multiple linear regression (MLR), and random forest regression (RFR), are conducted on the reduced set of features. Their results are analyzed with performance evaluation metrics, including mean square error (*MSE*) and R^2. Eventually, the best model based on minimum error or maximum accuracy, simplicity, and runtime cost is presented. This study tries to employ diversified data for modeling, including climatic and socioeconomic data. The remainder of this paper proceeds as follows. Section METHODS offers an overview of the method, including aggregated data, data preprocessing, feature selection methods, the proposed method, forecasting methods, and performance evaluation metrics. Section RESULTS AND DISCUSSION presents the dataset's details, selected features, parameter settings, results, and analysis of the method's performances. Finally, the conclusion is presented in Section CONCLUSIONS.

METHODS

An ML approach usually contains some stages, including data preparation and preprocessing, data transformation and feature selection, modeling, and deployment and test. This section represents the methodology and materials borrowed from the ML context and employed on the WDF issue. An overview of the methodology used in this research is available in Figure 1.

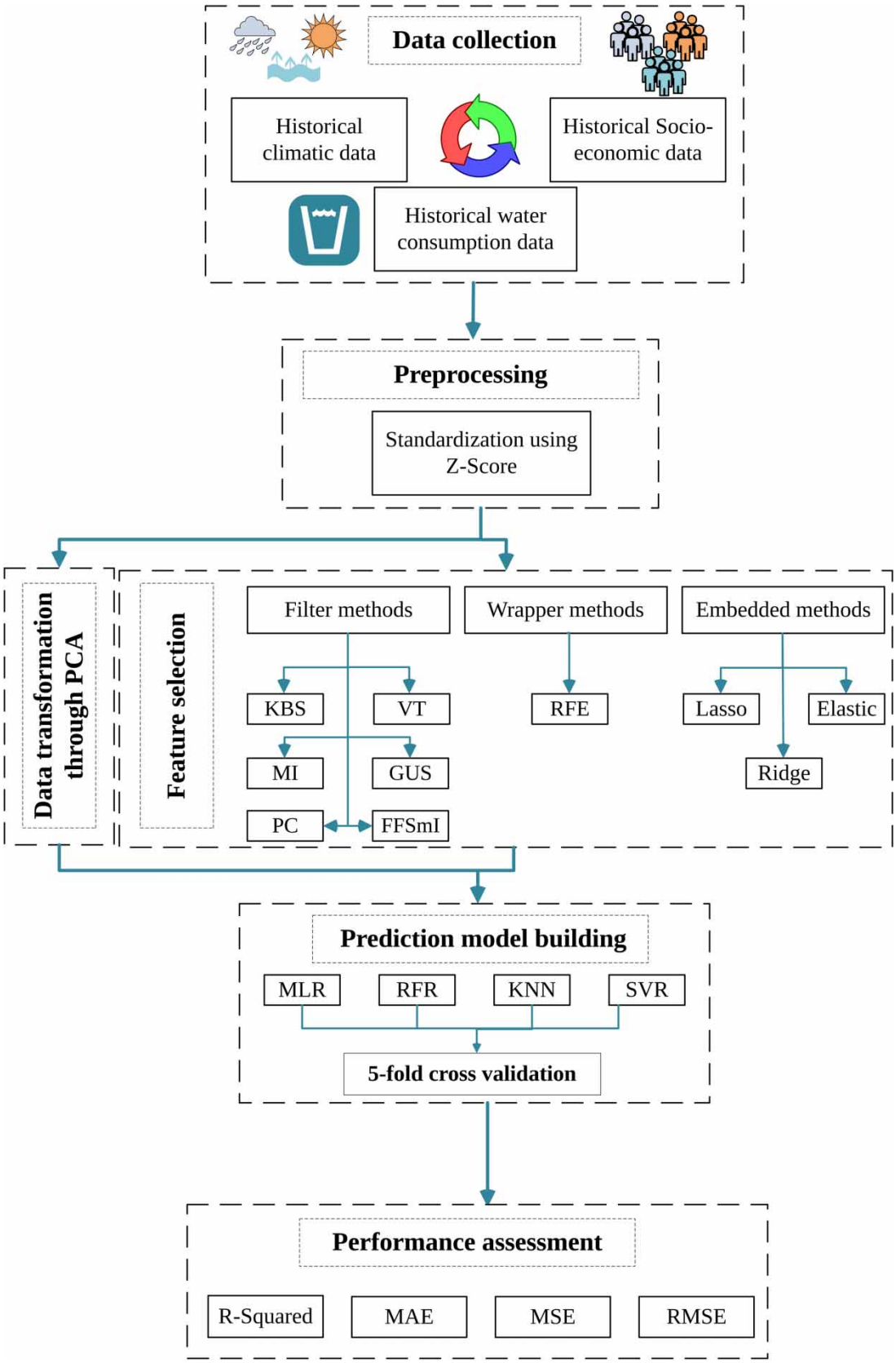


Figure 1 | Diagram of the methodology used in this research.

Case study design

Data collection

The implemented data in this paper is gathered from socioeconomic, climatic, and water consumption data of Isfahan province, Iran. Isfahan's area is about 107,000 square kilometres, which is larger than 89 countries, including Iceland, South Korea, and Portugal. Its population is about 5 million based on 2016 demographic data, which is roughly the population of Norway, Finland, or Ireland and is greater than 75 other countries. The data contain annual potable water consumption of 871 rural areas.

An enormous number of factors may be considered as influential on WD by the literature, from management intervention and socio-economic variables to demographic data, household characteristics, and weather–climate variables. This paper studies the most recent census of the area collected and published in 2016 (Statistical Centre of Iran 2016). It includes population, gender distribution, age distribution, the employment rate, home-ownership rate, house-type distribution, and housing area distribution in each rural area. To include weather, the most commonly measured weather parameter, air temperature, is used. The data provided by Iran Meteorological Organization (2016), which includes daily mean air temperature for each area, are aggregated to some parameters of a year-scale.

The potable water consumption data, along with the number of water subscriptions, in this paper are provided by Rural Water and Wastewater Company of Isfahan Province, Iran, which are measured every other month (Isfahan Province Rural Water and Wastewater Co 2016). The water consumption data are aggregated to an authentic feature of a year-scale.

Data description

The dataset used in this study consists of data per annum of 871 rural areas located in Isfahan province, Iran. The demographic data of these rural areas assembled with their potable water consumption data and weather information compose the dataset. The total population and number of households of the dataset are 508,111 and 163,213 respectively. All the measurements were made in 2016, as the most recent year that the demographic data were collected. Target water consumption data, along with 29 features, define each case of the dataset. For all features the data related to each rural area were available. Water data include a feature that represents the number of potable water subscriptions and water consumption in m^3.

Demographic data represents the household number, average family size, female ratio in the form of a female-to-population ratio, age structure in the form of distribution ratio among age groups, literacy rate, employment rate, number of owner-occupied and non-owner-occupied housings, non-apartment housing ratio, and housing area distribution.

The daily measured average temperature for each rural area is converted to several annual features. These features are maximum temperature, average summer temperature, and cooling degree days (CDD). A concise description of this paper's dataset is presented in Tables 2 and 3.

Table 2 includes all the features and their units. Table 3 shows statistical information about some of the features. It can be inferred from Table 3 that this case study is fairly distributed on every parameter.

Data Preprocessing

In a prediction problem, a dataset is divided into two subsets. A subset is used for training (training set) and a subset of data not seen by the model is used for testing (called validation or testing data). The purpose of cross-validation is to examine the model's ability to predict new data that was not used in the training phase.

Suppose the data of a dataset is available to be used in regression modeling. The goal in cross-validation is to achieve a model whose number of parameters is optimal, i.e., a model that does not suffer from overfitting. To achieve this aim, data is usually divided into two parts:

- Training data (training set): This part of the data is dedicated to creating a model and estimating its parameters.
- Test data (test set): This part of the data is devoted to testing the performance of the model. This part of the data includes the values of the independent and output variables that are not included in the training of the model but make the comparison of predicted and actual values possible.

Since test data do not affect the model, they have no role in determining the parameters of the model and are employed for model evaluation. However, in the cross-validation method, during a repetitive process, the training set is divided into two

Table 2 | List of features

Feature	Type/Unit	Feature	Type/Unit
Subscriptions	count	Non-owner-occupied housings	count
Households	count	Non-apartment housings	%
Average family size	number	Area −50- m^2	count
Female ratio	%	Area 51 to 75 m^2	count
Age 0 to 9	count	Area 76 to 80 m^2	count
Age 10 to 19	count	Area 81 to 100 m^2	count
Age 20 to 29	count	Area 101 to 150 m^2	count
Age 30 to 39	count	Area 151 to 200 m^2	count
Age 40 to 49	count	Area 201 to 300 m^2	count
Age 50 to 59	count	Area 301 to 500 m^2	count
Age 60 to 69	count	Area 501 + m^2	count
Age 70 +	count	Max temperature	°C
Literacy rate	%	Summer temperature	°C
Employment rate	%	CDD	count
Owner-occupied housings	count		

parts. Each time the cross-validation process is repeated, part of the data is used for training the model, and another part for testing the model. Therefore, this process is a resampling approach for estimating model error.

Since water consumption data and prediction span are annual, the daily temperature data should be transformed to appropriate parameters. CDD is the number of days with an average temperature above 18°C (65°F), and is typically used in the energy consumption context, and its influence on water consumption is examined in this study. Variables like CDD, maximum temperature, and average summer temperature are derived from daily air temperature data for each annum.

The demographic data represents age distribution as groups of five-year spans, from 0 to 74, and a group for above 75. In order to reduce problem dimensions, these groups are transformed into eight categories: 0 to 9, 10 to 19, 20 to 29, 30 to 39, 40 to 49, 50 to 59, 60 to 69, and 70+ .

Data standardization is conducted using the Z-score method. The Z-score shows the relative position of each value based on the mean, and is normalized by the standard deviation. It is a necessary procedure prior to data transformation using PCA. Equation (1) represents the Z-score formula (Kreyszig 2010):

$$z = \frac{(x - \mu)}{\sigma} \tag{1}$$

where x is the observed value, μ is the mean of the sample, and σ is its standard deviation.

Data transformation

It is usually required to use data transformation to convert data into a new coordinate system of a lower dimension. Since the transformation of data reduces the number of input variables, and selected suitable variables contain most of the information of the main data, the learning process and data analysis are performed more quickly, results are more reliable and data processing will be more efficient. For this purpose, PCA is defined as the orthogonal projection of the data in a way that maximizes the projected data's variance (Bishop 2006). The first coordinate (called the first principal component) is the greatest variance of the data's projection; the second coordinate is the next greatest variance, and so on.

Consider X as the data matrix with zero empirical mean and the observations $x_{(i)}$ as its rows. The $x_{(i)}$ are variables in p-dimensional space. Therefore, each of the X's p columns is a feature. Equation (2) represents the PCA transformation formula, that maps each row of X to a new vector of principal component scores $t_{(i)} = (t_1, \ldots, t_m)_{(i)}$, so that the individual

Table 3 | Feature characteristics

Feature	Mean	Std	Min	25%	50%	75%	Max
Subscriptions	188.54	245.80	1	30	92	243	1,833
Water consumption	36,122.32	45,960.00	150	5,177	18,444	48,500	258,516
Households	187.39	233.51	4	37	93	247.5	1,678
Average family size	2.96	0.52	1	2.7	3.0	3.3	5.5
Female ratio	284.10	361.70	0	50	135	382.5	2,548
Age 0 to 9	0.13	0.06	0	0.10	0.14	0.17	0.36
Age 10 to 19	0.11	0.04	0	0.08	0.12	0.14	0.36
Age 20 to 29	0.16	0.06	0	0.14	0.17	0.19	1
Age 30 to 39	0.16	0.05	0	0.14	0.17	0.20	0.53
Age 40 to 49	0.12	0.04	0	0.11	0.13	0.15	0.40
Age 50 to 59	0.11	0.04	0	0.08	0.10	0.12	0.33
Age 60 to 69	0.09	0.06	0	0.05	0.07	0.10	0.78
Age 70+	0.12	0.12	0	0.04	0.07	0.14	0.88
Literacy rate	0.75	0.15	0	0.69	0.79	0.85	0.95
Employment rate	0.50	0.09	0.11	0.47	0.50	0.54	1
Owner-occupied housings	148.00	184.57	0	29	75	192	1,349
Non-owner-occupied housings	21.86	42.13	0	2	7	23	488
Non-apartment housings	0.97	0.05	0.49	0.96	0.98	1	1
Area −50- m^2	5.99	10.17	0	1	2	7	125
Area 51 to 75 m^2	11.25	18.21	0	1	5	13	187
Area 76 to 80 m^2	11.30	18.63	0	1	4	12	149
Area 81 to 100 m^2	42.69	59.05	0	7	18	52.5	415
Area 101 to 150 m^2	69.81	97.35	0	8	30	89	652
Area 151 to 200 m^2	25.26	41.13	0	2	9	32	358
Area 201 to 300 m^2	8.36	14.96	0	0	3	9	159
Area 301 to 500 m^2	2.14	4.34	0	0	0	2	54
Area 501 + m^2	0.52	1.51	0	0	0	0	18
Max temperature	32.98	3.17	27	30	34	35	40
Summer temperature	28.86	2.43	24	27	29	31	34
CDD	164.13	25.23	121	140	170	185	221

variables of t successively inherit the maximum possible variance from x (Jang *et al.* 2018):

$$t_{k(i)} = x_{(i)} \cdot w_{(k)} \quad \text{for } i = 1, \ldots, n; \quad k = 1, \ldots, m \tag{2}$$

where $w_{(k)} = (w_1, \ldots, w_p)_{(k)}$ is a set of p-dimensional vectors of loadings which is constrained to be a unit vector, $t_{k(i)}$ is elements of $t_{(i)}$, and $x_{(i)}$ is rows of X.

Feature selection methods

Selecting relevant features that represent and fit data better is done by feature selection methods. This paper uses three categories of feature selection methods: filter, wrapper, and embedded methods. Filter methods act independently of learning models to select features. These methods select features by assigning scores to features based on evaluation metrics. This paper uses K-best selection (KBS), GUS, MI, Pearson correlation coefficient (PCC), variance threshold (VT), and FSSmI.

The proposed FSSmI feature selection method also belongs to filter methods, which works based on the smoothness index (SmI) (Kalhor *et al.* 2019). This index is mostly used for classification problems and has improved layer performance in deep neural networks. In this paper, this index is customized to select the regression problem's relevant features using forward selection. The combination of the forward selection and SmI proposed in this paper is expected to select features affecting regression. Details of the proposed method are given in Section METHODS.

Feature selection with smoothness index

The SmI is a generalized and extended version of the separation index (Cano 2013), as the quantized separation index is the SmI. The separation index is a measure of data evaluation and data complexity for classification problems, as it shows the separability between samples of different classes. After counting all the closest data points of the same label and normalizing them relative to all data points, a number ranging from 0 to 1 is obtained. The closer this number is to 1, the better the separation of the data of different classes is. That is, each new sample with a particular class is closer to samples of the same class than samples of other classes.

The proposed feature selection method can be used for any type of data and there is no limitation in this regard as long as that data is quantitative (can be expressed in numbers). This can be applied to both signal data (both 1-dimensional and multi-dimensional) and discrete data with no specific temporal–spatial order.

The SmI determines how close is the output of the nearest input to the input x^q and the output y^q. More close outputs result in SmI close to 1, i.e., data are smooth, and nearby inputs result in nearby outputs. More distant outputs result in decreasing SmI. SmI can be calculated in two ways: the linear SmI that uses linear punishment for an increasing distance of nearby outputs compared with the mean of output distances, and the exponential SmI that uses exponential punishment for increasing distance. Therefore, the exponential SmI is more strict and goes faster to 0 for more distant outputs.

The first step for calculating the smoothness of a set of vector points is to normalize the scattered data (anomalous data do not hurt the z-score). The input is assumed as a matrix with Q rows and n columns (each of the vector points appears as the row of a matrix), and the output will be a matrix with Q rows and m columns. The SmI for each of the inputs (x^q) considers the nearest input (x^{q^*}) and calculates their outputs (y^q and y^{q^*}). Then it calculates the L_p norm of outputs divided by the mean of output distances (d^q). Averaging distances between the output of instance q and the other outputs results in d^q. If the ratio of the distance between y^q and y^{q^*} to the mean of the output distances is low, then its *Relu* (rectified linear unit) gets close to 1. *Relu*, is a non-linear activation function in many types of multi-layer neural networks or deep neural networks. *Relu* will output the input directly if it is positive, otherwise, it will output 0. It has become the default activation function because a model that uses it is easier to train and often achieves better performance.

Assume there is Q with n-dimensional input–output data points $\{(x^q, y^q)\}_{q=1}^{Q}$ where x and y are two arbitrary dimensional vectors. Linear SmI is defined in Equation (3):

$$SmI = \frac{1}{Q}\sum_{q=1}^{Q} \mathrm{Relu}\left(1 - \frac{\left\|y^q - y^{q^*}\right\|_{L_p}}{d^q}\right)$$
$$q^* = \underset{i\in\{1,\ldots,Q\}, i\neq q}{\arg\min} \left\|x^q - x^i\right\|_{L_p} \tag{3}$$
$$d^q = \frac{1}{Q-1}\sum_{i=1, i\neq q}^{Q} \left\|y^q - y^i\right\|_{L_p}$$

where $0 \leq SmI \leq 1$. Based on the previously mentioned assumption, exponential SmI is defined in Equation (4):

$$SmI = \frac{1}{Q}\sum_{Q}^{q=1} \exp\left(-\frac{\left\|y^q - y^{q^*}\right\|_{L_p}}{d^q}\right) \tag{4}$$

where $0 \leq SmI \leq 1$.

SmI is important for a prediction problem. Inputs with higher SmI are more suitable and lead to a better feature selection. This is due to more possible and predictable interpolation (provided that data diversity is high enough) when SmI is higher. More smoothness and better interpolation result in more prediction power. One of the advantages of SmI is that it does not

force the map with a linear or Gaussian distribution; it just needs the map to be continuous and smooth. Functions are called smooth if they are derivable in their domain. Smoothness indicates to what extent the closeness of two inputs results in their outputs' closeness. Limited distance of inputs leads to limited distance of outputs. This keeps the continuity of output. Furthermore, SmI can be indirectly obtained by loss functions, including *MSE* and entropy, and directly by finding a proper combination of inputs so that the smoothness grows.

SmI is used for feature selection in this study, and it is a forward selection with the initial number of features equal to n. At first, each feature separately is examined with SmI, and a feature with the highest smoothness on output is selected. Then it is studied which feature, along with selected features, has the most smoothness with the output. A handcrafted illustrative example of the selection of related and unrelated features from the perspective of smoothness is provided. In Figure 2, 1,000 input data points with X_1 and X_2 features were randomly generated. Also, the y output is defined by the two-dimensional synchronous functions introduced in Equation (5). Denote white-noise variables with X_3, and X_4 features have a uniform distribution function. All the features are considered in the range of 0 to 5.

$$y = \frac{\sin(x_1)\sin(x_2)}{x_1 x_2}, \quad 0 < |x_1| \leq 5, 0 < |x_2| \leq 5, 0 < |x_3| \leq 5, 0 < |x_4| \leq 5 \tag{5}$$

Six different subsets are analyzed for selection of features. As is shown in Table 4, an appropriate feature space is created out of the primary input space (X_1 and X_2) by designing a feature selection section using the smoothness indicator.

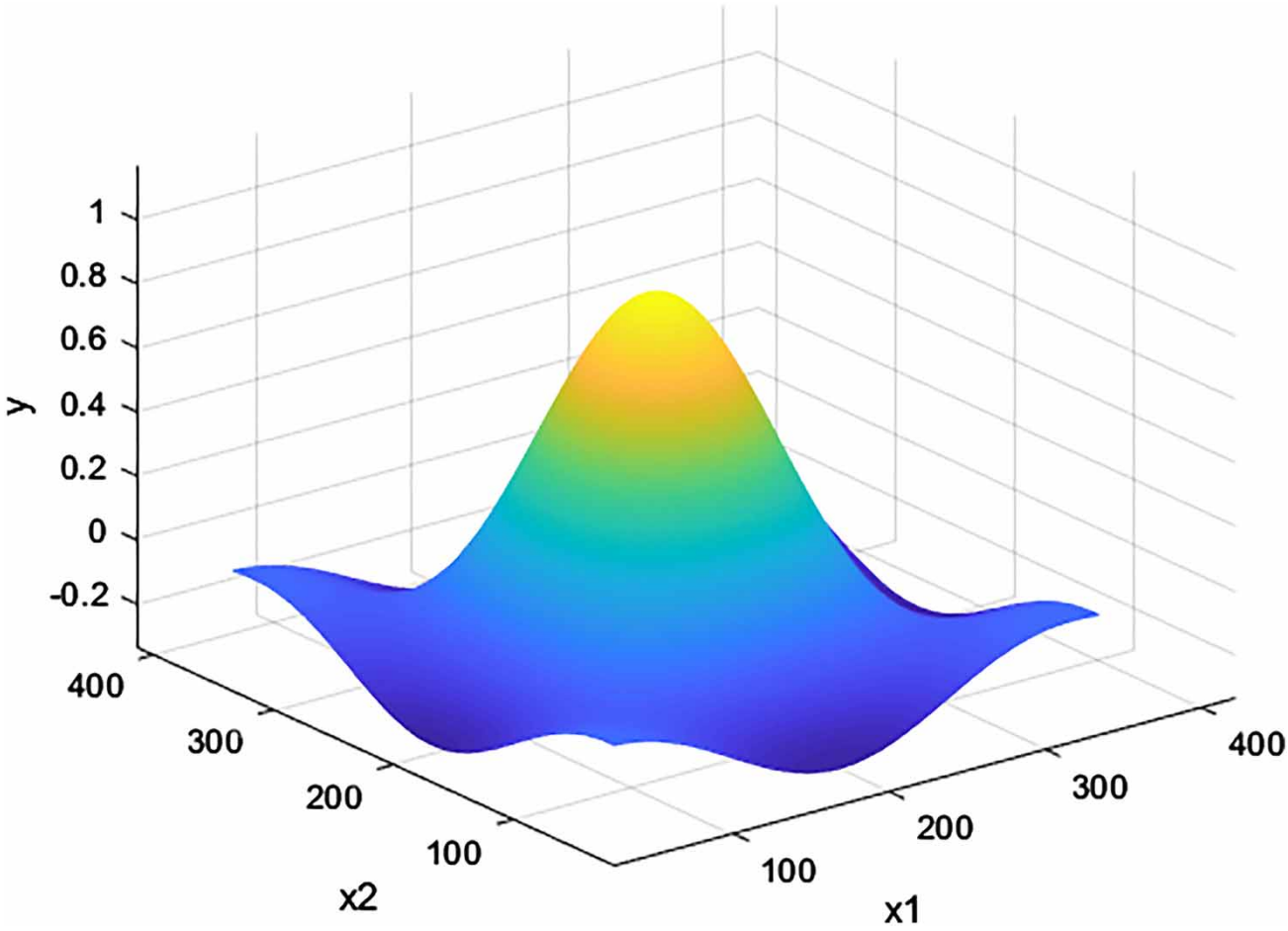


Figure 2 | Two-dimensional synchronous function of 1,000 randomly generated data points.

Table 4 | SmI comparison for different subsets of handmade data

Different subsets of two main inputs and two non-related inputs					Feature smoothness index	
Subsets/Inputs	X_1	X_2	X_3	X_4	Linear	Exponential
1	×	×			0.9783	0.9788
2	×	×	×		0.9159	0.9249
3	×	×	×	×	0.8314	0.8615
4		×	×		0.4781	0.5972
5		×	×	×	0.4711	0.5888
6				×	0.3464	0.4929

This happens because of the topmost quantity of linear and exponential SmI. In such an example, it is expected that for each input data point, the data points in the closest vicinity produce an output that is very close to the output of the given data point and their difference is a minor error. The reason is that using smoothness provides us with the ability of prediction. On the contrary, when one or two primary inputs are removed and replaced with unrelated or noise input, SmI gets worse and is reduced. This means that the mapping of the data becomes unsmooth and jaggy. Because of the high sensitivity of the output to the quantitative alterations in feature space, it is expected that the generalizations of such a function decrease.

Feature selection with other methods

Wrapper methods select or eliminate a subset of features based on the prediction model's performance, i.e., they use an evaluation metric dependent on the prediction algorithm in a recursive way to determine the usefulness of selected features. The methods in this category, due to high computational cost, are not very popular. The RFE from this category is used in this study.

Embedded methods, like wrapper methods, work in interaction with features and models, but models in this category embed the feature selection process in themselves. Lasso, Ridge, and Elastic are ML models with this category of feature selection.

PCC is used to summarize the strength of the linear relationship between two data variables. Its value can be from −1 to 1: −1 for negative correlation, 1 for positive correlation, and 0 for no linear correlation. Equation (6) calculates the PCC for variables x and y (Weaver *et al.* 2017):

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{6}$$

where $\bar{x}$ is the mean of the x variable, $\bar{y}$ is the mean of the y variable, n is the sample size, and x_i and y_i are single samples indexed with i.

A high correlation of a variable and target is usually useful because one can predict the other. However, a high correlation between two data variables is interpreted as redundant information relating to the target, and removing one of them reduces the problem dimensions.

MI is a method to evaluate the relevance of a subset of features in predicting the target variables, while considering redundancy relating to other features. It quantifies the amount of information obtained about random variable Y through random variable X, shown in Equation (7) (Brown *et al.* 2012):

$$\mathrm{MI}(X;Y) = \int_X \int_Y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy \tag{7}$$

where $p(x,y)$ is the joint probability density function of X and Y, and $p(x)$ and $p(y)$ are the marginal density functions.

For feature selection, the MI between the subset of selected features X_S and the target variable y should be maximized, as shown in Equation (8) (Brown *et al.* 2012):

$$\tilde{S} = \arg\max_{S} \mathrm{MI}(X_S; y) \quad s.t. |S| = k \tag{8}$$

where $\tilde{S}$ is joint mutual information, and k is the number of features that it is intended to select. Although calculating the quantity $\tilde{S}$ is an NP-complete problem, there are some solutions to estimate the result.

Other feature selection techniques are as follows: RFE that recursively fits a model and removes the weakest feature until reaching a specified number of features, VT that removes features with variation below a specific cutoff, KBS that selects features based on the k highest scores, and GUS that is a univariate selector with a configurable strategy. A univariate selector examines statistical tests on features and targets and keeps high-scored features, i.e., it works by selecting the best features based on univariate statistical tests. All feature selection methods are implemented using Python and the ML library Scikit-learn (Buitinck *et al.* 2013).

ML methods

The potable WDF can be considered as a regression problem, and several ML methods have been studied to solve it (Antunes *et al.* 2018). This paper explores seven ML methods and will demonstrate the results. These methods are k nearest neighbors (KNN), RFR, SVR, MLR, least absolute shrinkage and selection operator (Lasso), Ridge, and Elastic regression.

KNN is a type of instance-based learning that predicts the target by values of the k nearest neighbors in the available data, measured with a weight function. There are several weight functions available for this purpose. RFR is an ensemble learning method that aggregates many decision trees; each is trained on a different data sample where sampling is done with replacement. SVR's objective function is to minimize the L2 norm of the coefficient vector while ensuring the absolute error is less than or equal to ϵ, the defined maximum error. MLR is the extension of the ordinary least-squares regression with several explanatory variables to predict the target by fitting a linear equation to the data.

The other three ML methods embed feature selection procedures in them. Lasso regression performs L1 regularization while Ridge regression makes use of L2 regularization. Elastic regression combines L1 and L2 priors as its regularizer.

Model performance evaluation

The model validation is obtained by a k-fold cross-validation approach (Stone 1974) with $k = 5$. The randomized samples are split into five equal-size folds. For each fold, a model is trained by other than that fold's data and validated by the fold's data. Averaging errors of these five validations results in the final model score. After evaluation of different values for k, it was concluded that based on the minimum error and prediction performance methods, cross-validation would be suitable for $k = 5$.

There are several standard statistical metrics for performance assessment; amongst them this paper uses root mean square error (*RMSE*), *MSE*, coefficient of determination (R^2), and mean absolute error (*MAE*). *RMSE*, *MSE*, R^2, and *MAE* equations are represented by Equations (9), (10), (11), and (12), respectively:

$$RMSE = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(W_{\text{predicted}}(l_i) - W_{\text{measured}}(l_i))^2} \tag{9}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(W_{\text{predicted}}(l_i) - W_{\text{measured}}(l_i))^2 \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(W_{\text{measured}}(l_i) - W_{\text{predicted}}(l_i))^2}{\sum_{i=1}^{n}(W_{\text{measured}}(l_i) - \bar{w})^2} \tag{11}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|W_{\text{measured}}(l_i) - W_{\text{predicted}}(l_i)| \tag{12}$$

where $W(l_i)$ indicates WD for the i-th location, $\bar{w}$ is the mean measured WD, and n is the sample size. *RMSE* and *MSE* values close to 0 are the best possible outcome. R^2 values lie between 0 and 1, and closeness to 1 indicates the regression predictions nearly fit the data. *MAE* is a parameter for measuring the closeness of predictions to outcomes, and a smaller value of *MAE* means a more accurate model (Jang & Choi 2017). Using these performance assessment measures, along with a five-fold cross-validation approach on every ML model and feature selection method, provides a thorough comparison.

RESULTS AND DISCUSSION

In this section, besides previous regression methods and feature selection methods, the proposed new FSSmI method is tested on a newly collected rural water consumption dataset. After evaluating the models' performance using the evaluation metrics mentioned in Section 'Model performance evaluation', the models have been improved by setting the appropriate parameters. The results of the performance of the models have been analyzed and compared.

Selected features

Feature selection is an essential part of an ML method that can be defined as determining relevant features and eliminating irrelevant or redundant features to reach a subset of features that well define the problem.

ML methods, in conjunction with the feature selection methods studied in this paper, result in 34 different models. This section explores features selected for each model using the feature selection methods. Each feature selection method

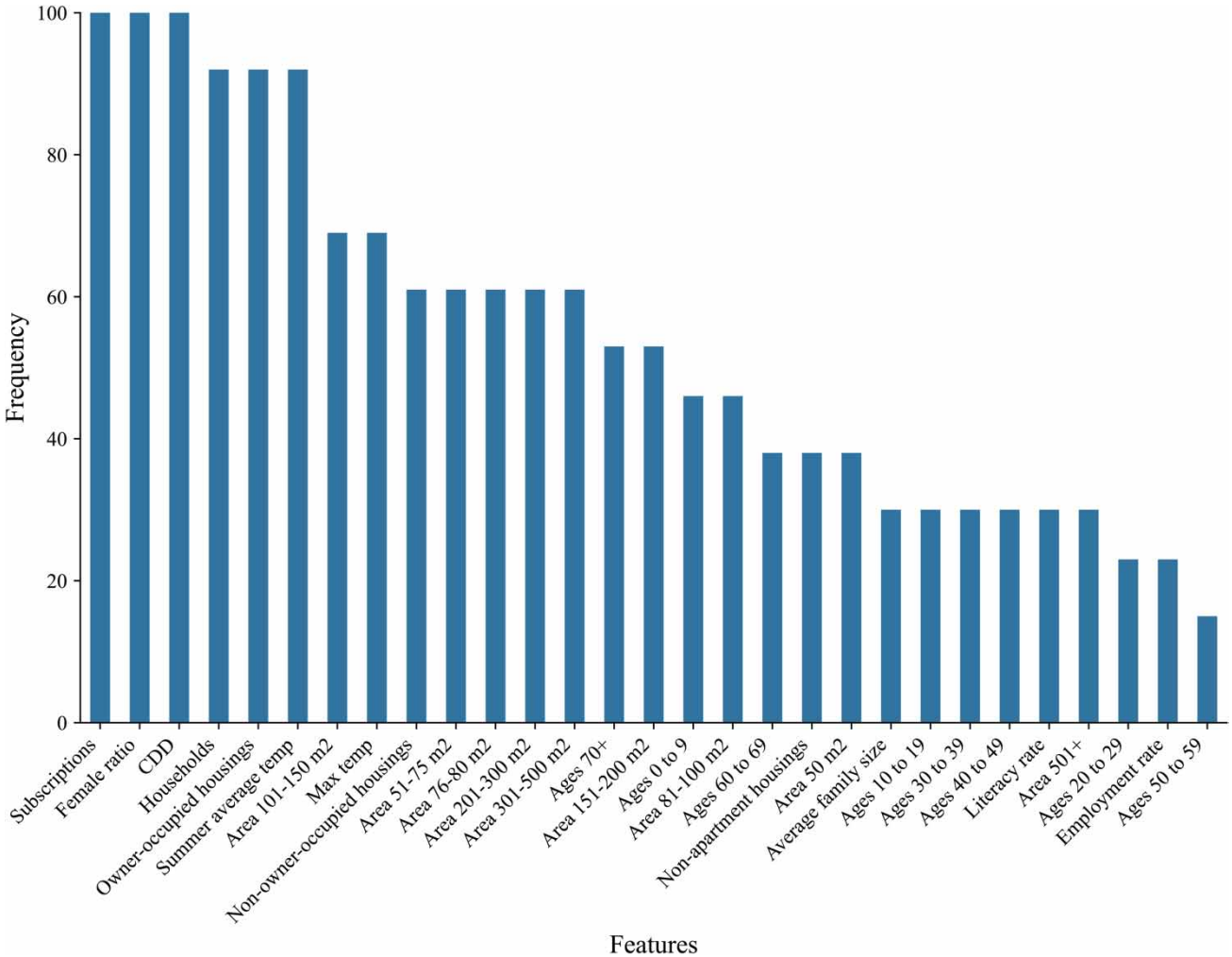


Figure 3 | Feature appearance frequency in the feature selection methods' outputs. It shows the importance of subscriptions, housing ownership status, ambient temperature, and housing area distribution.

shares the same features among ML methods, except RFE, which selects different features for each model. The most anticipated observation is that subscriptions (which is tightly in proportion to population) appears in all models. Other important features that are selected by more than 60 percent of models include temperature features (CDD, max, and average summer temperature), and demographic features (households, non-owner/owner-occupied housings, female ratio, and housing area distribution features including 51–75m^2, 76–80m^2, 101–150m^2, 201–300m^2, and 301–500m^2). Surprisingly, each of the age distribution groups, except for age 70+, appears in less than 50 percent of models. Figure 3 illustrates the features' appearance frequency in the outputs of the feature selection methods.

It is inferable from Figure 3 that population features, especially subscriptions, show a strong relation to water consumption. On the other hand, age distribution features do not affect much water consumption behavior. Based on the combination of MLR, RFR, and SVR with GUS, RFE, and FSSmI, the 70+ years old feature has a stronger relationship with WD than other age distribution features. The six most important features that appear in more than 90% of models are subscriptions, female ratio, CDD, households, owner occupied housings, and summer average temperature, respectively. It can be inferred from

Table 5 | Selected features

Feature	KBS	VT	PCC	MI	GUS	FSSmI	RFE on MLR	RFE on RFR	RFE on SVR	Lasso	Ridge	Elastic
Subscriptions	×	×	×	×	×	×	×	×	×	×	×	×
Households	×	×	×	×	×	×	×	×	×		×	×
Average family size						×		×		×		
Female ratio	×	×	×	×	×	×	×	×	×	×	×	×
Age 0 to 9					×	×	×		×	×		
Age 10 to 19					×		×		×			
Age 20 to 29								×	×			
Age 30 to 39							×	×	×			
Age 40 to 49							×	×	×			
Age 50 to 59							×					
Age 60 to 69					×		×		×	×		
Age 70+					×	×	×	×	×	×		
Literacy rate						×					×	×
Employment rate					×	×						
Owner-occupied housings	×	×	×	×		×	×	×	×	×	×	×
Non-owner-occupied housings	×	×	×	×	×						×	×
Non-apartment housings						×		×			×	×
Area −50 m^2		×						×			×	×
Area 51 to 75 m^2	×	×	×	×					×		×	×
Area 76 to 80 m^2	×	×	×	×				×			×	×
Area 81 to 100 m^2	×	×	×	×								×
Area 101 to 150 m^2	×	×	×	×			×	×			×	×
Area 151 to 200 m^2	×	×	×	×							×	×
Area 201 to 300 m^2	×	×	×	×	×					×		×
Area 301 to 500 m^2		×	×		×	×				×	×	×
Area 501 + m^2		×			×	×						
Max temperature		×	×	×		×	×		×		×	×
Summer temperature		×	×	×	×	×	×	×	×	×	×	×
CDD	×	×	×	×	×	×	×	×	×	×	×	×
Number of features	12	17	15	14	14	15	15	15	15	11	16	18

The features, households, subscriptions, and female ratio are selected by all the feature selection methods that show their influences on regression.

Figure 3 that the temperature has a significant impact on water consumption pattern. However, the literacy rate, family size, and employment rate do not show a meaningful effect.

PCA and filter methods are combined with ML methods, creating 28 models. Selected features of each filter method are the same among all ML methods. The wrapper method is used with MLR, RFR, and SVR; it selects 15, but not necessarily the same, features for each of them. ML methods perform feature selection embeddedly. Table 5 shows the selected features for all models.

Figure 4 illustrates feature importance based on the mutual information feature selection method. Accordingly, the most important features are subscriptions, female ratio, households, and owner-occupied-housings, respectively.

The SmI presented in Figure 5 is based on the dataset dimensions. The point with the maximum smoothness indicates an optimum number of features, where adding other features probably does not end up with more information. It is evident from Figure 5 that the optimum number of features is 15. This method selects subscriptions, owner-occupied housings, average family size, area 301 to 500, summer temperature, CDD, female ratio, households, age 70+, non-apartment housings, area 501+, employment rate, literacy rate, max temperature, and age 0 to 9.

Model settings

Parameter settings for feature selection methods

The number of components (components with the highest variance) is selected from two to ten for the PCA. The calculation of how many components are required to explain the data is a critical part of using PCA in practice. This number can be

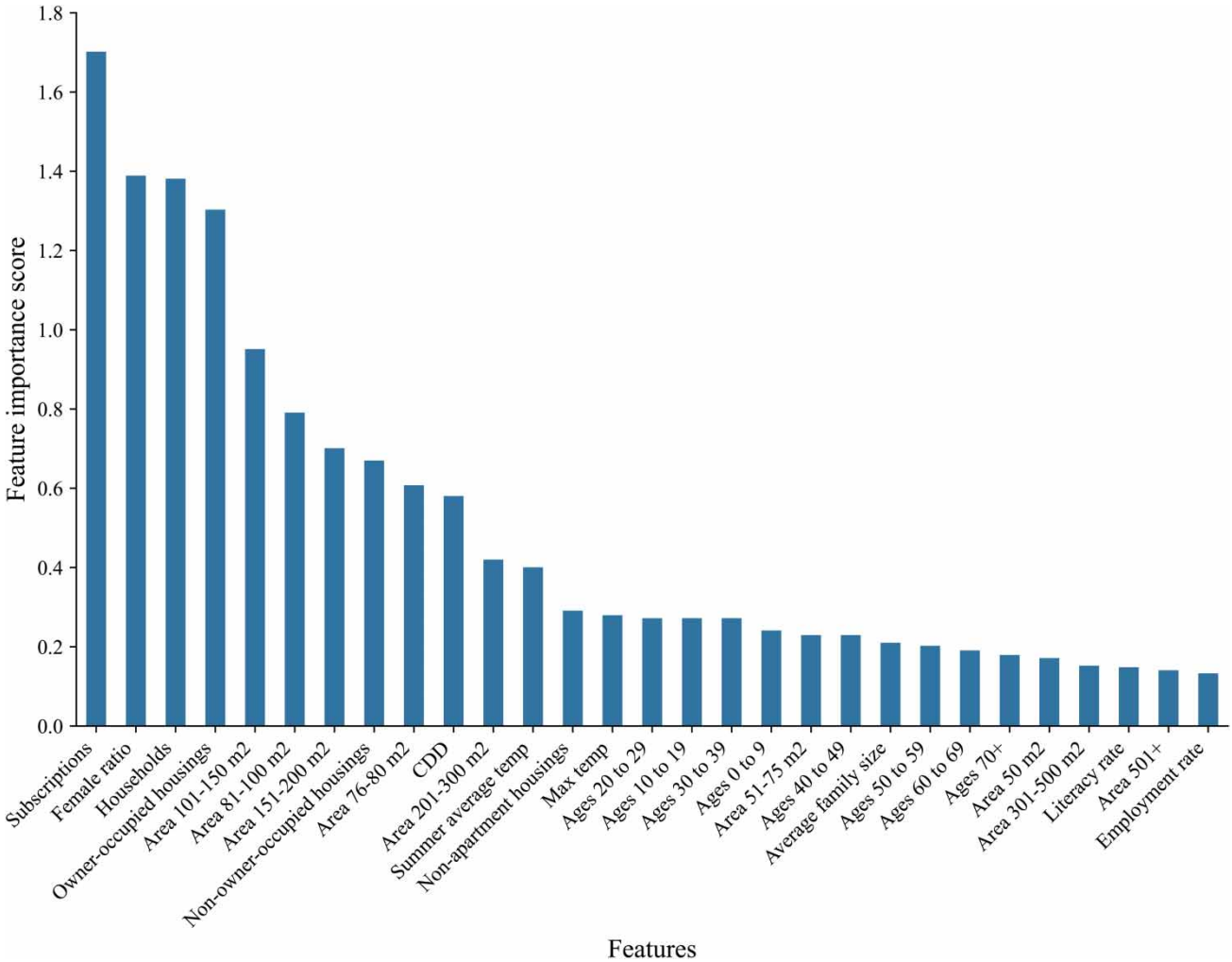


Figure 4 | Feature importance based on mutual information. The scores of features such as subscriptions, households, female ratio, and owner-occupied housings are greater than others, i.e., they provide more information for models.

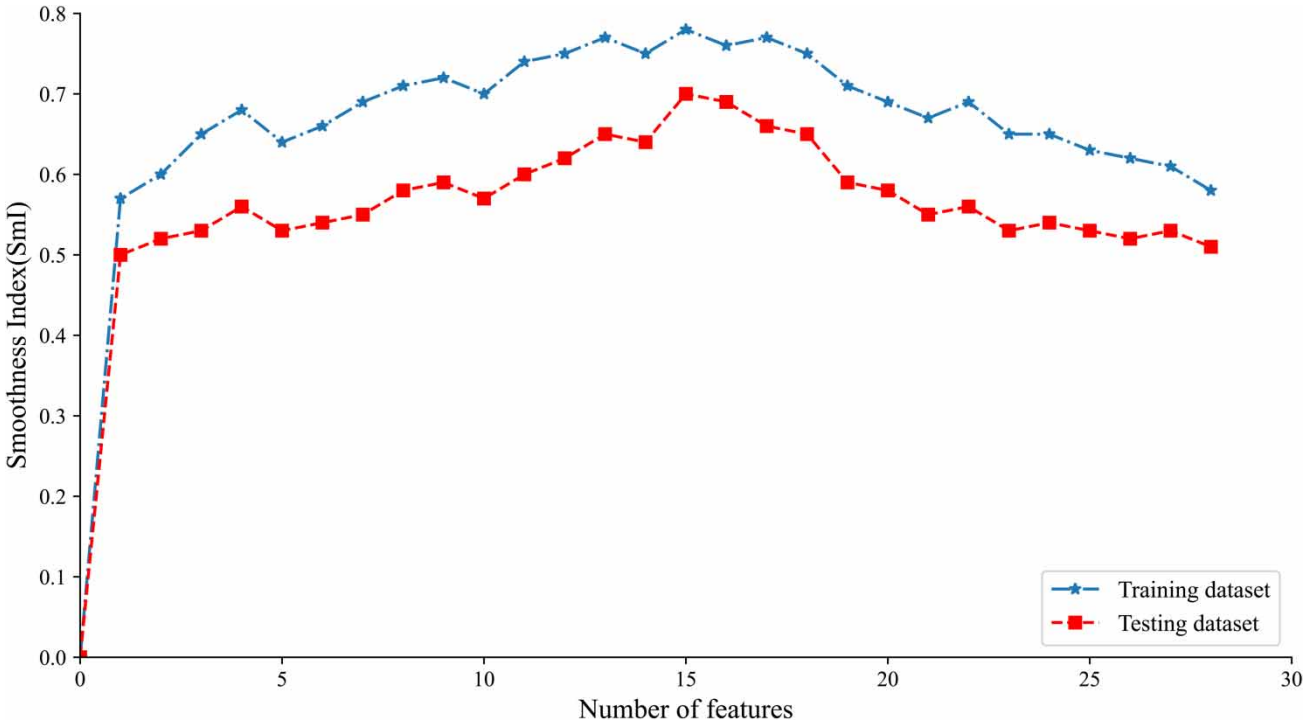


Figure 5 | Smoothness Index (SmI) based on number of features. There is a good correlation between the smoothness charts of the training and test datasets, i.e., the selected features based on the absent data are the same as those that give the highest SmI in the training dataset.

estimated by looking at the cumulative explained variance ratio as a function of the components' number. The explained variance (the amount of variance explained by each of the selected components) says how much data (variance) can be applied to each of the principal components. For instance, Figure 6 reveals that the first two components constitute approximately

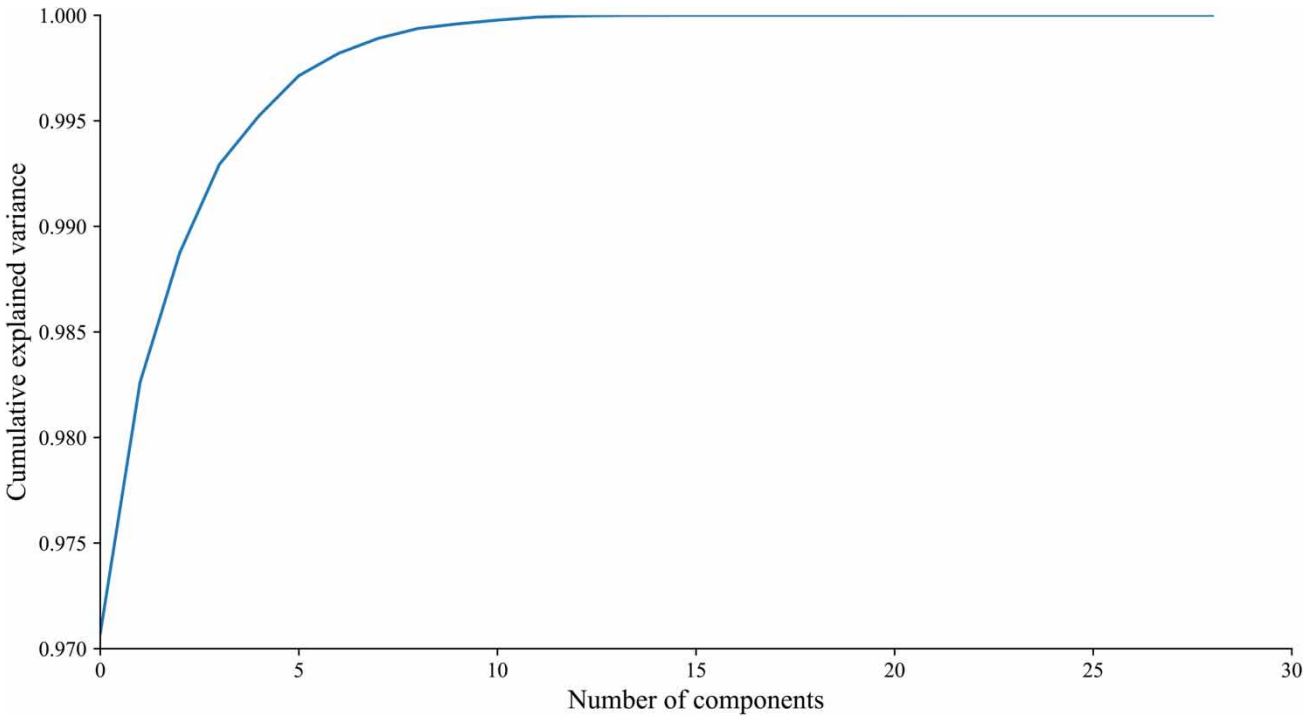


Figure 6 | Cumulative proportion of explained variance. Nearly 99% of variance can be explained with two components.

99% of the variance while describing approximately 100% of the variance requires about ten components. For GUS, the feature selection mode is 'Percentile', and the corresponding mode parameter is 20.

For RFE (recursive feature elimination ranking), the number of selected features is 15, and the number of features to be removed for each iteration, termed 'step', is 1.0. RFE is a feature selection method that fits a model and, until a specified number of features is achieved, eliminates the weakest feature(s). Features are ranked by the model's attributes of coef_ or feature_importances_, then RFE aims to remove dependencies and collinearity that may exist in the model by recursively removing some features per loop. The threshold is set to 0.16 for VT, which excludes features with a training set variance lower than this threshold. PCC is adjusted to select only features correlated with water consumption higher than 0.5, excluding one from each pair of features that correlate higher than 0.9. For MI, discrete_features is set to false for dense inputs and true for sparse inputs, and it selects features based on the 14 highest scores measured with KBest. MI is a non-negative value that scores the dependency between two random variables, that if and only if variables are independent, it is equal to 0; otherwise, higher values stand for higher dependencies. For FSSmI, $p = 2$ is used in L_p, which is Euclidean distance.

According to Figure 6, the number of principal components (components with maximum variance) of PCA are taken from the range two to ten, which produces eigenvalues above 97%. The central part of PCA is estimating the number of components for well-describing data. It can be determined by considering the cumulative explained variance ratio as a function of the number of components. The explained variance states how much data (variance) can be related to each of the principal components. For instance, Figure 6 shows the first two components encompass roughly 97% of the variance while extending the percentage to 100% needs around ten components.

The Pearson correlation matrix shows that the highest correlation coefficient, which is 0.98, holds between subscriptions and water consumption. All of the correlation results are found in Figure 7. The presence of a high correlation among features (independent variables) shows multicollinearity, which probably makes bias in regression analysis results. For the Pearson correlation, only features with more than a 0.5 factor of correlation with water consumption are chosen. If there are features (other than water consumption) with a correlation higher than 0.9 in between, one of them is eliminated. Figure 7 is the heatmap of features correlations.

Parameter settings for forecasting methods

For MLR, the ordinary least-squares method is adopted in this paper, which minimizes the sum of squared errors (SSE) to obtain the optimal regression function. In the case of RFR models, the expansion of the number of trees in the forest and the performance evaluation metrics positively affect the predictions' quality. In this paper, the number of trees in the forest was set to be 20, and the maximum tree depth was open, i.e., the nodes expand as long as all the leaves are pure. MSE was selected as the performance evaluation metric. The choice of SVR core function is one of the most critical parameters that affect its performance and is more important than tolerance in predictions. In this paper, a linear function was selected as the core function, the regularization parameter C was set to 15, and the stop criterion's tolerance was 0.001. Also, shrinking heuristics are not used in the training process. The performance of KNN models has a strong relationship with the number of neighbors, so the number of neighbors is optimally obtained in this paper. The results show no significant difference in weight functions, although the result is slightly better when using the uniform weight function. Ridge regression solves a regression model whose loss function is a linear least-squares function, and its regularization is done with the L2 norm, which improves the conditioning of the problem and reduces the variance of the estimates. In this paper, alpha or regularization strength is set to 1, and the solver, which is used in computational procedures, is set to auto (based on data type). For ElasticNet regression, in this paper, the alpha value is selected as 1, and the L1 ratio is selected as 0.5. In this study, because of five-fold cross-validation, there is no need to use LassoCV anymore, Since part of the LassoCV method is selecting cross-validation, its reselection will run a loop. Also, for Lasso regression, alpha is set to 1.

Model performance

The important point of every forecasting method is its performance evaluation. After normalizing data and determining essential features, a prediction model (MLR, RFR, SVR, KNN, Lasso, Ridge, and Elastic) is run on them based on k-fold cross-validation with $k = 5$ and the metrics introduced in Section 'Model performance evaluation' (R^2, *MAE*, *MSE*, and *RMSE*). The collected data are used to compare models' performance for WDF in this section. R^2 shows how predicted values are close to the observed data. The 0 value of R^2 indicates that no prediction is close to the observed data, and 1 is for exact prediction. *MSE* shows the mean of the error squared, which is the difference between predicted and observed values. For *MSE*,

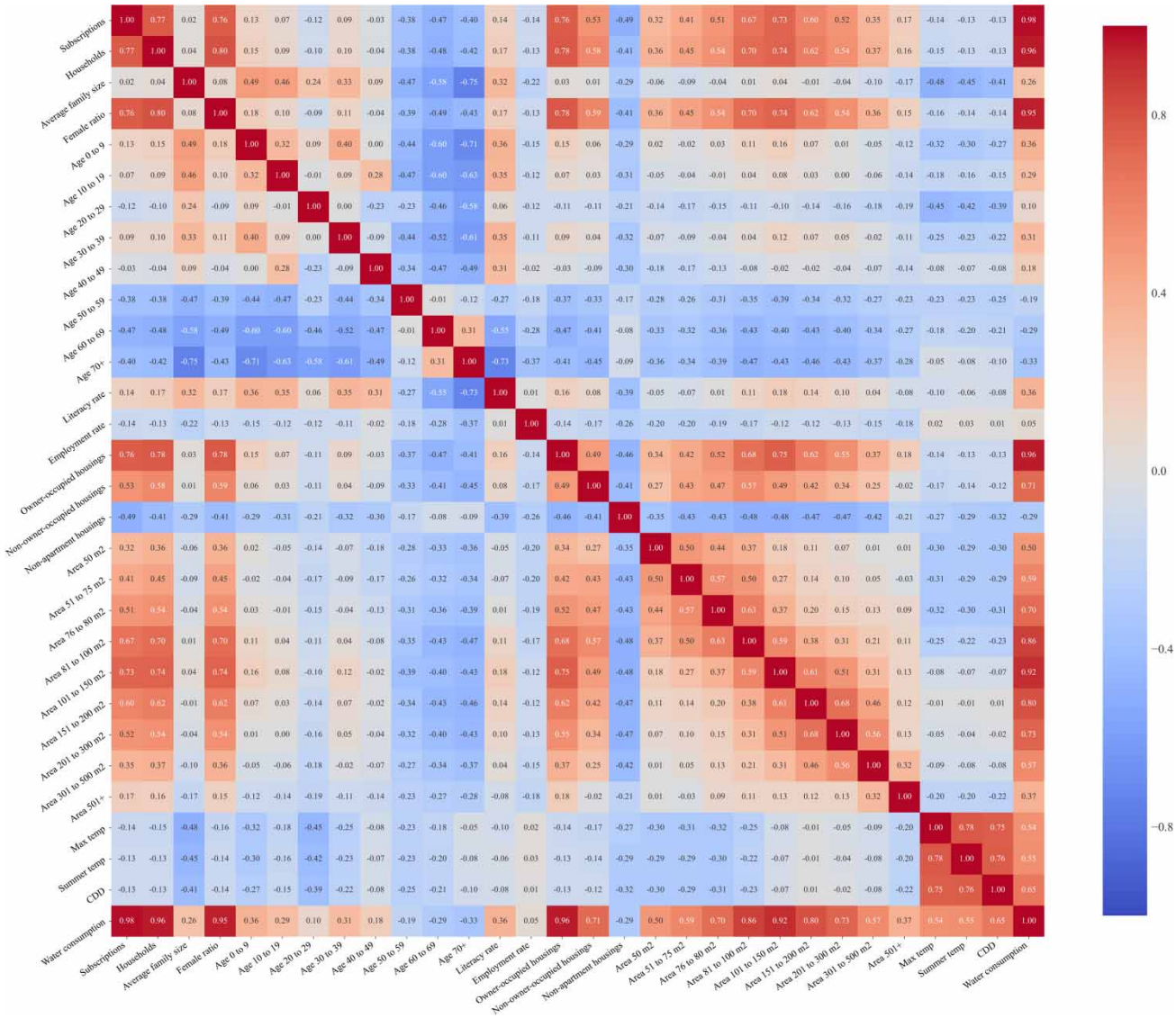


Figure 7 | Heatmap of Pearson correlations. Colors indicate the degree of correlation of the data. Dark red and dark blue represent perfect positive and negative correlations, respectively, and light colors stand for weak correlations.

RMSE, and *MAE*, values close to 0 indicate more accurate forecasting. The performance comparison of forecasting methods is represented in Tables 6–9. The MLR forecasting model shows the best performance based on the four metrics, which means it has the least forecasting error and most accuracy for WD prediction. So, the MLR model has the best forecasting performance for residential potable WDF of rural Isfahan. SVR has the second-best performance, while KNN shows the least accuracy and most error among all, and acts poorly in WDF.

The results represented in Tables 6–9 can be interpreted in two different aspects: (a) the best model-prediction performance of each feature selection method, independently for each evaluation criterion and (b) the best feature selection methods having the best model-prediction performance for different evaluation criteria.

(a) MLR prediction has the best result for PCA feature selection for all evaluation criteria. RFR prediction for the all evaluation criteria has the best result for RFE feature selection. MLR prediction for the R^2 evaluation criterion and RFR prediction for the other evaluation criteria had the best result for GUS feature selection. SVR prediction for all evaluation criteria has the best result for KBS feature selection. MLR prediction for the R^2 evaluation criterion and SVR prediction for the other evaluation criteria had the best results for VT, PCC, and MI feature selections. RFR prediction for all evaluation criteria had the best result for FSSmI feature selection. Ridge had the best result for the R^2 and *MAE* evaluation

Table 6 | Performance evaluation using R^2 for all models

	PCA	GUS	RFE	KBS	VT	PCC	MI	FSSmI	Embedded
MLR	**0.9657**	**0.9645**	0.9658	0.9653	***0.9684***	**0.9659**	**0.9670**	0.9669	*
RFR	0.9387	0.9637	**0.9669**	0.9550	0.9605	0.9564	0.9568	**0.9682**	*
SVR	0.9523	0.9549	0.9556	**0.9661**	0.9674	0.9652	0.9658	0.9671	*
KNN	0.9179	0.9383	0.9545	0.9547	0.9524	0.958	0.9541	0.9604	*
Lasso	*	*	*	*	*	*	*	*	0.9658
Ridge	*	*	*	*	*	*	*	*	**0.9675**
Elastic	*	*	*	*	*	*	*	*	0.9654

*The methods are not comparable.

Table 7 | Performance evaluation using *MAE* for all models ($\times 10^4$)

	PCA	GUS	RFE	KBS	VT	PCC	MI	FSSmI	Embedded
MLR	**0.5232**	0.5499	0.5220	0.5219	0.4969	0.5179	0.5105	0.5081	*
RFR	0.6965	**0.4816**	**0.4796**	0.5203	0.5014	0.5495	0.5351	**0.4986**	*
SVR	0.5386	0.5066	0.4975	**0.4857**	***0.4781***	**0.4874**	**0.4868**	0.5162	*
KNN	0.8168	0.7123	0.5381	0.5727	0.5707	0.5555	0.5728	0.5167	*
Lasso	*	*	*	*	*	*	*	*	0.5290
Ridge	*	*	*	*	*	*	*	*	**0.4914**
Elastic	*	*	*	*	*	*	*	*	0.4919

*The methods are not comparable.

Table 8 | Performance evaluation using *MSE* for all models ($\times 10^6$)

	PCA	GUS	RFE	KBS	VT	PCC	MI	FSSmI	Embedded
MLR	**0.2786**	0.3031	0.2764	0.2726	0.2470	0.2687	0.2655	0.2495	*
RFR	0.4912	**0.2347**	**0.2306**	0.2714	0.2520	0.3034	0.2901	***0.2301***	*
SVR	0.2916	0.2600	0.2501	**0.2365**	**0.2301**	**0.2400**	**0.2391**	0.2555	*
KNN	0.6696	0.5080	0.2576	0.3290	0.3264	0.3100	0.3337	0.2581	*
Lasso	*	*	*	*	*	*	*	*	0.2815
Ridge	*	*	*	*	*	*	*	*	0.2457
Elastic	*	*	*	*	*	*	*	*	**0.2447**

*The methods are not comparable.

criteria and Elastic for the *MSE* and *RMSE* evaluation criteria for embedded feature selection. KNN and Lasso predictions in combination with various feature selection methods were not among the best for any evaluation criteria.

(b) VT and FSSmI feature selections had the best results for the four evaluation criteria presented in this paper (which are marked in italic bold in the tables). The VT feature selection had the best results for the R^2 and *MAE* evaluation criteria and the FSSmI feature selection had the best results for the *MSE* and *RMSE* criteria.

As previous studies suggest, after feature selection, the prediction model improves in related metrics, prediction performance and accuracy, and computation time; thus, feature selection is employed in this paper. The results indicate that the best

Table 9 | Performance evaluation using *RMSE* for all models ($\times 10^4$)

	PCA	GUS	RFE	KBS	VT	PCC	MI	FSSmI	Embedded
MLR	**0.5279**	0.5506	0.5257	0.5221	0.4970	0.5184	0.5153	0.4994	*
RFR	0.7008	**0.4845**	**0.4802**	0.5210	0.5020	0.5508	0.5386	***0.4796***	*
SVR	0.5400	0.5099	0.5001	**0.4864**	**0.4797**	**0.4899**	**0.4890**	0.5054	*
KNN	0.8183	0.7127	0.5075	0.5736	0.5713	0.5567	0.5777	0.5080	*
Lasso	*	*	*	*	*	*	*	*	0.5305
Ridge	*	*	*	*	*	*	*	*	0.4957
Elastic	*	*	*	*	*	*	*	*	**0.4947**

*The methods are not comparable.

feature selection method is the proposed method, FSSmI. Based on the results of all performance evaluation metrics on all ML methods, FSSmI leads to better outcomes, and thus better predictions.

In all cases, ML methods can produce accurate predictions, which is a standard method and usually is used for forecasting, although for different datasets, different models make the best prediction. Thus, testing various methods and models while analyzing new datasets is necessary to find the best outcome.

CONCLUSIONS

As one of the most vital resources for fulfilling life needs, water plays a critical role in daily lives. WDF of the household sector and recognizing water-use factors are the main steps of water crisis management and control. The residential WDF and identifying its influential factors are among the most vital steps of water crisis management, because supplying more water is not always a practical solution; it is necessary to adopt new policies and approaches based on water consumption pattern and use them for planning and decision-making for foreseen WD changes in upcoming years. The novel method developed in this paper provides a practical tool for policymakers to achieve sustainable management of potable WD. The importance of water consumption pattern makes this paper adopt a new feature selection approach for residential WD prediction. Several prediction methods such as SVR, MLR, and RFR, combined with various feature selection methods including RFE, GUS, PCA, and MI, were tuned, implemented, analyzed, and compared in predicting potable residential WD. Seven feature selection methods result in different subsets of predicting parameters. Although the resulting regression coefficients are related to a specific province's rural areas, this method can be employed to analyze residential water consumption of different locations with various climatic and socioeconomic conditions. The outcomes of this study are:

1. Water consumption is considered as a long-term yearly scale and a function of climatic and socioeconomic parameters. Various ML-based forecasting and feature selection methods are implemented in this study and their performance is assessed for a specific problem.
2. Recent developments in feature selection have resolved the challenge of multivariate input spaces' performance so that datasets with many features are released from the curse of dimensionality in which the data may be overfitted. The outputs of this study indicate that many combinations of features do not necessarily improve the models' performance.
3. A novel feature selection framework is proposed to improve forecasting models and indicate the most appropriate input combination, which is a mixture of the forward selection and SmI. The RFR regression method, along with feature selection with SmI, improves WDF performance. In general terms, greater smoothness and better interpolation make better predictions.
4. The presented method has a significant role in identifying features effective in modeling WDF, compared with other methods. The feature selection methods are used to reduce feature dimensions by eliminating redundant features. The reduced subset of features results in more accurate predictions. The FSSmI method selects effective features, including CDD, average family size, and age.
5. The performance evaluation metrics (*RMSE*, R^2, and *MSE*) are improved in this study. The FSSmI method shows an improvement of about 1.5%, 9%, and 8% based on previously mentioned measures, respectively, comparing to previous methods. The results confirm the accuracy and advantage of this method.

6. The results of feature selection using different methods show a close relationship between water consumption and some characteristics such as subscriptions, housing ownership status, ambient temperature, and housing area distribution. On the other hand, some other features such as literacy rate, housing type, unemployment rate, and age distribution will impact water consumption to a lesser extent. Among the features related to temperature, CDD, and among the features related to the housing area, housings with an area between 101 and 150 square metres have the strongest relationship with water consumption.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Antunes, A., Andrade-Campos, A., Sardinha-Lourenço, A. & Oliveira, M. S. 2018 Short-term water demand forecasting using machine learning techniques. *Journal of Hydroinformatics* **20** (6), 1343–1366. doi:10.2166/hydro.2018.163.

Billings, R. B. & Jones, C. V. 2008 *Forecasting Urban Water Demand*. American Water Works Association, Denver, CO, USA.

Bishop, C. M. 2006 *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, USA.

Brown, G., Pocock, A., Zhao, M. J. & Luján, M. 2012 Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of Machine Learning Research* **13**, 27–66. Available from: http://www.jmlr.org/papers/volume13/brown12a/brown12a.pdf.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Müller, A. C., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B. & Varoquaux, G. 2013 API design for machine learning software: experiences from the scikit-learn project. In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122. Available from: https://hal.inria.fr/hal-00856511.

Cano, J.-R. 2013 Analysis of data complexity measures for classification. *Expert Systems with Applications* **40** (12), 4820–4831. https://doi.org/10.1016/j.eswa.2013.02.025.

Cleveland, W. S., Devlin, S. J. & Grosse, E. 1988 Regression by local fitting: methods, properties, and computational algorithms. *Journal of Econometrics* **37** (1), 87–114. https://doi.org/10.1016/0304-4076(88)90077-2.

Donkor, E. A., Mazzuchi, T. A., Soyer, R. & Roberson, J. A. 2014 Urban water demand forecasting: review of methods and models. *Journal of Water Resources Planning and Management* **140** (2), 146–159. doi:10.1061/(ASCE)WR.1943-5452.0000314.

Eslamian, S. A., Li, S. S. & Haghighat, F. 2016 A new multiple regression model for predictions of urban water use. *Sustainable Cities and Society* **27**, 419–429. doi:10.1016/j.scs.2016.08.003.

Fullerton, T. M. & Molina, A. L. 2010 Municipal water consumption forecast accuracy. *Water Resources Research* **46** (6), W06515. doi:10.1029/2009WR008450.

Glantz, M. & Mun, J. 2011 Projections and risk assessment. In: *Credit Engineering for Bankers*, 2nd edn (Glantz, M. & Mun, J.), Academic Press, Boston MA, USA, pp. 185–236. https://doi.org/10.1016/B978-0-12-378585-5.10008-9.

Iran Meteorological Organization 2016 Available from: https://www.irimo.ir/eng/.

Isfahan Province Rural Water and Wastewater Co 2016 Available from: https://abfar-isfahan.ir/Index.aspx?tempname=ENMain{Ŭ&}lang=2{Ŭ&}sub=0.

Jang, D. & Choi, G. 2017 Estimation of non-revenue water ratio for sustainable management using artificial neural network and Z-score in Incheon, Republic of Korea. *Sustainability* **9** (11), 1933. doi:10.3390/su9111933.

Jang, D., Park, H. & Choi, G. 2018 Estimation of leakage ratio using principal component analysis and artificial neural network in water distribution systems. *Sustainability* **10** (3), 750. ISSN 20711050. doi:10.3390/su10030750.

Kalhor, A., Saffar, M., Kheirieh, M., Hoseinipoor, S. & Araabi, B. N. 2019 Evaluation of dataflow through layers of deep neural networks in classification and regression problems. *arXiv* 1096.05156.

Kindler, J. & Russell, C. S. 1984 *Modeling Water Demands*. Academic Press, London, UK. Available from: http://pure.iiasa.ac.at/2392.

Kreyszig, E. 2010 *Advanced Engineering Mathematics*. John Wiley & Sons, Chichester, UK.

Lee, S.-J., Wentz, E. A. & Gober, P. 2010 Space–time forecasting using soft geostatistics: a case study in forecasting municipal water demand for Phoenix, Arizona. *Stochastic Environmental Research and Risk Assessment* **24** (2), 283–295. doi:10.1007/s00477-009-0317-z.

Levin, E. R., Maddaus, W. O., Sandkulla, N. M. & Pohl, H. 2006 Forecasting wholesale demand and conservation savings. *Journal/American Water Works Association* **98** (2), 102–111. doi:10.1002/j.1551-8833.2006.tb07592.x.

Makki, A. A., Stewart, R. A., Beal, C. D. & Panuwatwanich, K. 2015 Novel bottom-up urban water demand forecasting model: revealing the determinants, drivers and predictors of residential indoor end-use consumption. *Resources, Conservation and Recycling* **95**, 15–37. doi:10.1016/j.resconrec.2014.11.009.

Nasseri, M., Moeini, A. & Tabesh, M. 2011 Forecasting monthly urban water demand using Extended Kalman Filter and Genetic Programming. *Expert Systems with Applications* **38** (6), 7387–7395. https://doi.org/10.1016/j.eswa.2010.12.087.

Schleich, J. & Hillenbrand, T. 2009 Determinants of residential water demand in Germany. *Ecological Economics* **68** (6), 1756–1769. doi:10.1016/j.ecolecon.2008.11.012.

Statistical Centre of Iran 2016 *Metadata. Statistical Centre of Iran*. Available from: https://www.amar.org.ir/english/Metadata/Classifications.

Stone, M. 1974 Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)* **36** (2), 111–147. doi:10.1111/j.2517-6161.1976.tb01573.x. Available from: https://www.jstor.org/stable/2984809.

Villarin, M. C. & Rodriguez-Galiano, V. F. 2019 Machine learning for modeling water demand. *Journal of Water Resources Planning and Management* **145** (5), 04019017. doi:10.1061/(ASCE)WR.1943-5452.0001067.

Weaver, K. F., Morales, V., Dunn, S. L., Godde, K. & Weaver, P. F. 2017 Pearson's and Spearman's correlation. In: *An Introduction to Statistical Analysis in Research*, John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 435–471. doi:10.1002/9781119454205.ch10.

First received 22 April 2022; accepted in revised form 14 June 2022. Available online 27 June 2022