https://doi.org/10.1016/j.fss.2021.02.018

An Improved Fuzzy Rule-based System using Evidential Reasoning and Subtractive Clustering for Environmental Investment Prediction

Long-Hao Yang [a, c, d], Fei-Fei Ye [a, c, d], Jun Liu [c], Ying-Ming Wang [a, b, *], Haibo Hu [d]

[a] Decision Sciences Institute, Fuzhou University, Fuzhou, 350116, PR China

[b] Key Laboratory of Spatial Data Mining & Information Sharing of Ministry of Education, Fuzhou University, Fuzhou, 350116, PR China

[c] School of Computing, Ulster University at Jordanstown Campus, Newtownabbey, Northern Ireland, BT37 0QB, UK

[d] Department of Electronic and Information Engineering, Hong Kong Ploytechnic University, Hong Kong, PR China

*The corresponding author. Email: msymwang@hotmail.com

Abstract: Environmental investment prediction has attracted much attention in the last few years. However, there are still great challenges in investment prediction modeling, *e.g.*, 1) effective environmental indicators must be accurately selected to avoid the curse of dimensionality; 2) effective environmental data must be reasonably selected to downsize the scale of historical data; 3), the higher interpretability and lower complexity of prediction models must be considered. To address the above three challenges, a new environmental investment prediction model using fuzzy rule-based system (FRBS), evidential reasoning (ER) approach, and subtractive clustering (SC) algorithm is proposed in the present work, called FRBS-ERSC. In this new model, the FRBS is the core component for the modeling of environmental investment prediction and therefore provides good interpretability and complexity to environmental managers. Meanwhile, the ER approach is used as an improvement technique of the FRBS to combine the strengths of different feature selection methods for better indicator selection, and the SC algorithm is used as another improvement technique of the FRBS to select effective environmental data. An empirical case of environmental investment prediction is studied based on data on 31 provinces in China ranged from 2005 to 2018. The experimental results show that the proposed FRBS-ERSC not only provides interpretable and scalable environmental investment prediction based on effective indicator selection and data selection, but also produces satisfactory accuracy compared to some existing models.

Keywords: Fuzzy rule-based system; Environmental investment prediction; Evidential reasoning; Subtractive clustering

1. Introduction

The modern era has considerably increased the pollutant emission and the speed of ecological damages. Maintaining or improving effectiveness of environmental management is therefore one of important goals of sustainable developments. The research community has shown growing concern over minimizing the cause of environmental hazard by scientific environmental investment schemes [7][43][47]. The ever-increasing expenditures on environmental pollutant control and ecological restoration have become a challenging issue to be addressed. To adapt to the above situation, many investment planning models were developed and devoted to accurate investment predictions according to pollutant emissions and government financial situations. However, reasonable modeling of investment prediction is still an urgent problem to be solved at present.

Various environmental indicators and historical data can be used to develop a prediction model for making suggestions on the investments of environmental protection and pollution control. Effective investment prediction modeling usually requires good indicator selection, data selection, and the consideration of high interpretability and low complexity. Hence, all of these aspects constitute three great challenges in environmental investment prediction modeling. In the past decade, many studies have been devoted to solving these challenges, *e.g.*, expert knowledge-based, indicator integration-based, and feature selection-based indicator selection, time series forecasting-based and input-output relationship-based investment

prediction modeling. The details of these studies can be found in ***Section 2***.

From previous studies above, on the one hand, there are many kinds of endeavors for selecting effective indicators, but all these endeavors only focus on the application of one feature selection method to select indicators. It is clear that any kind of methods inevitably has its strengths and weaknesses [46]. Hence, the existing endeavors are usually difficult to provide a desired solution to address the challenge of indicator selection. On the other hand, time series forecasting-based modeling is the most popular approach in previous studies on investment prediction, but this kind of modeling just need to adapt the change law of investments, ignoring the influence of economic development and environmental pollution emission on investment inputs. Hence, input-output relationship-based modeling is a better approach for investment prediction.

The principle of input-output relationship-based modeling is based on a data-driven model to approximate the hidden internal relationship among economic development, environmental pollution, and investment inputs [47]. This is the main reason that input-output relationship-based modeling can be better than time series forecasting-based modeling for environmental investment prediction. But the former one has a higher requirement in data selection, namely selecting a smaller part of historical data and using them for investment prediction modeling. This is because environmental investment prediction is a periodic task to formulate investment schemes. Additionally, the core subject in environment management is managers, who should have a full view of investment prediction process, so it is necessary to consider high interpretability and low complexity for input -output relationship-based modeling.

To overcome the above challenges in investment prediction modeling, the fuzzy rule-based system (FRBS), which is one of the most important fields of the applications about fuzzy sets and fuzzy logic, is used as the main methodology to construct a new environmental investment prediction model. Moreover, the evidential reasoning (ER) approach [42] and subtractive clustering (SC) algorithm [8] are also introduced to improve the deficiencies of traditional FRBS for predicting environmental investments. The improved FRBS is so called FRBS-ERSC. Owing to these components of FRBS-ERSC, the present study has the following contributions to solve the challenges:

For the challenge of indicator selection, various kinds of feature selection methods are used to independently identify the relative importance of different environmental indicators from different perspectives. In order to collaboratively select representative indicators for investment prediction modeling, the ER approach is further used to combine the relative importance of each environmental indicator, so that the selection of environmental indicators can benefit from the strengths of different existing feature selection methods and also contribute to the modeling of a FRBS for environmental investment prediction. It is worth noting that, in the past development, the ER approach was always used as an inference engine to improve FRBSs, *e.g.*, [10][11][20], but this is the first time used for indicator selection when modeling a FRBS.

For the challenge of data selection, the similarity between environmental data is considered to decrease the redundancy of the historical data, whose size will grow indefinitely with annual human productions and activities, resulting in too many data used in investment prediction modeling. Based on this point of view, the SC algorithm is introduced to cluster the data and thus select representative data. Owing to the SC algorithm, the improved FRBS has a downsized fuzzy rule base and a desired accuracy for predicting environmental investments. It is worth noting that although FRBS with the SC algorithm has been used in different applications [1][3][29], to best of our knowledge, they have never been used together for the field of environmental investment prediction.

For the challenge of high interpretability and low complexity in modeling, the FRBS, in a sense, is the extension of classical IF-THEN rule-based system because of the improvements that fuzzy sets and fuzzy logic are used as tools for

better representing different forms of linguistic knowledge, as well as for modeling the hidden internal relationship existing between its causal variables. Moreover, thanks to the advantages of fuzzy logic, the inference process of FRBS has become more robust and interpretable better than traditional machine-learning techniques such as support vector machine (SVM) [6], support vector regression (SVR) [16][33], artificial neural network (ANN) [22][27], and extreme learning machine [21][39]. All of them provide a satisfactory solution for investment prediction modeling.

To demonstrate the effectiveness of the proposed FRBS-ERSC, an empirical case regarding the actual environmental investment data derived from 2005 to 2018 on 31 provinces in China is used to illustrate the development procedure of the proposed FRBS-ERSC and also provide the comparative analysis of the FRBS with and without improvements, and some existing time series forecasting-based and input-output relationship-based investment prediction models.

The remainder of this work is as follows: ***Section 2*** is the literature review and outlines challenges of environmental investment prediction. ***Section 3*** introduces the traditional FRBS used in environmental investment prediction. ***Section 4*** proposes an improved FRBS for environmental investment prediction. ***Section 5*** provides a case study to perform model validation. ***Section 6*** concludes this study.

2. Literature Review on Environmental Investment Prediction and its Challenges

In this section, the previous studies of environmental indicator selection and investment prediction modeling are firstly reviewed, followed by the summarization of the challenges to propose a new model for investment prediction.

2.1. Review of previous studies on environmental indicator selection

Environmental indicator selection is an important procedure in investment prediction, because there are various kinds of environmental indicators but some of them are irrelevant indicators, which may have negative influences on the resulting prediction models. In the previous studies, environmental indicator selection is mainly based on the following ways:

The first way is expert knowledge and this way is very common in the previous studies of environmental indicator selection [13], *i.e.*, in the analysis of China's sustainable development's investments [7], environmental indicators were selected by prior knowledge from pollutant emissions, such as total volume of sulphur dioxide, chemical oxygen demand, carbon dioxide, and others. Analogously, the prior knowledge was also used as an effective way to select environmental indicators in the existing studies, like investigating the influence of knowledge trade on sustainable development and environmental biased technical progress [36], and analyzing the impact of environmental regulation and foreign investments on the amount and intensity of carbon emissions [50]. Recently, by considering the data availability and referring previous researches, fourteen environmental indicators were introduced to validate the model proposed for corporate environmental performance prediction in China [51]. Obviously, the above existing studies have a strong subjectivity in indicator selection [44], which will lead to some disputes on the objectivity and rationality of the selected indicators.

The second way is indicator integration and the core of this way is based on a combination model to integrate indicator information, so that the integrity of indicator information can be guaranteed in environmental investment prediction. The representative combination models used to integrate indicators include: 1) math function-based combination models, *e.g.*, the study of evaluating corporate sustainable performance [14], which applied an average function to integrate the corporate environmental, economic, and social performance scores into a comprehensive sustainable performance score; and the study of assessing the environmental impact of municipal solid waste management [9], which utilized a geometric mean function to integrate the normalized indicators of solid waste generation, carbon dioxide emission, energy consumption, and waste quality index into a new indicator; 2) ER-based combination models, such as the study of environmental governance cost

prediction [47], which used the classical ER approach to integrate the set of undesirable output indicators and desirable output indicators into two new indicators, respectively, and the study of efficiency evaluation for air pollution management [49], which utilized the interval ER approach to integrate multiple air pollution-related indicators with interval uncertainty into a new indicator. However, for these existing studies, it is inevitable that useless indicators would be combined because these studies may lack the selection of effective environmental indicators.

The third way is feature selection or extraction and this way is mainly based on classical feature selection or extraction methods to select representative environmental indicators, where principal component analysis (PCA) is one of commonly used feature extraction method in the previous studies on indicator selection, *i.e.*, PCA was introduced to extract the top four principal components of the original environmental indicators as new indicators to construct an investment prediction model [48]; Similarly, PCA was also used to select effective environmental indicators in [2][14][38] and demonstrated the effectiveness of PCA on environmental indicator selection. Among the existing representative studies based on feature selection methods, Salcedo-Sanz *et al*. [31] discussed the importance of different feature selection methods on renewable energy applications and found that feature selection is an important process in the existing prediction systems for renewable energy applications; Bui *et al*. [4] used a fuzzy rule-based algorithm FURIA as indicator evaluator and the generic algorithm as search strategy in order to select optimal set of the indicators used in flood susceptibility modeling assessment. Recently, Wang *et al*. [43] used the correlation-based feature selection (CFS) method to select representative indicators for modeling under the background of environmental investment prediction. Clearly, in the above existing studies, the reasonability of the selected indicators is based on the performance of the used feature selection methods.

2.2. Review of previous studies on investment prediction modeling

The theoretical analysis of investment prediction mainly originated in the researches on economic investments, and the associated mathematical models were therefore used to describe the structure of economic systems. Thereafter, these models were gradually applied to the environmental investment prediction. In previous studies, investment prediction modeling is mainly based on the following approaches:

The first approach is time series forecasting and this approach should work on the assumption that future trends of data will hold similar to the historical trends of data, so that the future investments can be accurately predicted by analyzing the trends of the past investments. In the previous studies on investment prediction based on time series forecasting, the auto regressive integrated moving average (ARIMA) model is one of commonly used models and the related works include: based on the ARIMA model, the coal consumption, price, and investment of China from 2016 to 2030 were predicted and the results showed that the coal investment has the similar result with coal consumption [15]; on the basis of ARIMA model and an improved support vector machine (SVM), a hybrid prediction was proposed in [17] to predict the investment of electricity generation and distribution. In addition to the ARIMA model, the grey model (GM) is another commonly used model for investment prediction and the related works include: an optimized hybrid GM model was utilized to predict future energy consumption and the results indicated that the proposed GM model outperforms other prediction models [45]; On the basis of combining GM and inverse data envelopment analysis (DEA) model, an integrated method was proposed to predict and analyze the investment problem of China's sustainable development during the 2015-2024 period [7]; Recently, time series forecasting was also used to estimate the target-availability of China's investments for green growth. However, a common shortcoming can be found from the above studies and it is that these studies ignored the input-output relationship between environmental pollution and environmental investment.

The second approach is input-output relationship-based modeling and this approach takes the indicators of economic development and environmental pollution as inputs and the indicators of environmental investments as outputs to construct prediction models. The representative studies include: in the study of environmental governance cost prediction [47], which introduced tree structure to represent the mathematical function of each output by combining possible operators in non-leaf nodes and all inputs in leaf nodes and the results showed that the predicted costs based on tree structure are closer to actual costs over time series forecasting-based models. Meanwhile, another concise but effective expression, namely IF-THEN rule, is also used to represent the input-output relationship between environmental pollution and environmental investment, in which extended belief rule is one of IF-THEN rules and has been successfully used in investment prediction modeling, *i.e.*, extended belief rule-based system (EBRBS)-based model was introduced to predict environmental governance costs [43] and transportation industry governance costs [40]. The results of both two studies demonstrated that the EBRBS-based model has a high accuracy in investment prediction for labor, capital, and energy better than time series forecasting-based models. Recently, as a classical IF-THEN rule, fuzzy rule was introduced in [48] to achieve investment prediction modeling as well, and the resulting FRBS-based model not only obtains desired prediction accuracy, but also has advantages in the terms of interpretability and model complexity. From the above studies, it can be found that the input-output relationship-based modeling is becoming a trend in the field of investment prediction modeling.

2.3. Challenges of proposing new model for investment prediction

According to the review of environmental indicator selection and investment prediction modeling, the following three challenges must be overcome for proposing a new environmental investment prediction model.

Challenge 1: Since environmental investment prediction involves various indicators, the first challenge is how to select effective environmental indicators.

From previous studies, a large number of environmental indicators were used for environment management, *e.g.*, gross domestic product, emission of carbon dioxide, emission of wastewater, and others. However, too many indicators used in investment prediction modeling would cause the problem of dimension curse. A smart way to solve this problem is the application of a method to select indicators, as shown in ***Section 2.1***, but any kind of methods inevitably have its strengths and weaknesses. Hence, it is necessary to propose a collaborative strategy for selecting indicators using different methods.

Challenge 2: Due to the continuing cycle of economic development and environmental pollution, the second challenge is how to select effective environmental data.

Environmental investment prediction is a periodic task to formulate investment schemes for environment management according to the historical data of economic development and environmental pollution. Clearly, the size of these historical data would grow indefinitely with annual human productions and activities, resulting in too many data used in investment prediction modeling. Hence, it is necessary to propose a filtering process for selecting a smaller part of historical data and using that subset for investment prediction modeling.

Challenge 3: High interpretability and low complexity should be considered in a prediction model for better assisting environmental managers to formulate investment schemes.

Environmental investment prediction not only needs a model to accurately predict investments, but also requires higher interpretability and lower complexity in prediction process. This is because environmental managers must have a full view of investment prediction, so that they can have enough confidence to design an investment plan based on the predicted investments. Hence, it is necessary to propose a new investment prediction model with consideration of high interpretability

and low complexity.

The three challenges above clearly indicate the necessary conditions of proposing a new environmental investment prediction model. From the existing models shown in ***Section 2.2***, FRBS has shown its ability in high interpretability and low complexity because of its inherent superiorities, but it is still lack of a feasible method to select effective environmental indicators and data. Hence, in the subsequent sections, the traditional FRBS for predicting environmental investments and its improvements are proposed to overcome the above challenges.

3. Traditional FRBS for Predicting Environmental Investments

FRBS is a well-known fuzzy system with the advantages of high interpretability, accuracy and low complexity. To date, it has been widely used in various fields, such as basic classification problems [12][18] , wind power-related application and prediction [24][25][26][28], and shear strength prediction of concrete beams [23][30][32][35][37]. According to the existing study [48], the generation and inference of a FRBS for environmental investment prediction are showed as follows:

Step 1: To divide the input and output spaces of environmental indicators into fuzzy regions

Suppose an environmental investment problem includes M input indicators x_i (i=1,…, M), *e.g.*, economic development (x_1) and environmental pollution (x_2), and one output indicator y, *e.g.*, environmental investment. The value ranges of these input and output indicators are $[x_i^-, x_i^+]$ and $[y^-, y^+]$, respectively. Thus, according to decision-makers' preference, N fuzzy regions, such as *Low*, *Middle*, and *High*, can be assigned to each input indicator and output indicator, in which the type of fuzzy regions can be triangular, trapezoidal, and others; each fuzzy region has a fuzzy membership function. **Fig. 1** shows an example when assuming M=2 and N=3 with triangular fuzzy number for each fuzzy region.

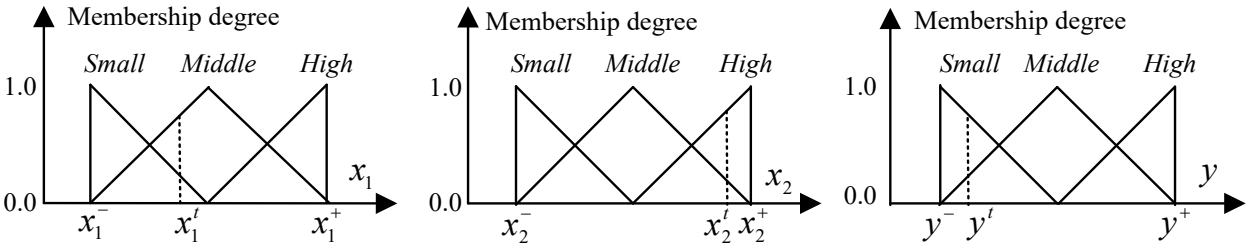

Fig. 1 Fuzzy regions and the corresponding membership functions

Step 2: To generate fuzzy rules from the given data pairs of environmental investments.

Suppose a set of input-output data pairs $(x_1^t,...,x_M^t, y^t)$ (t=1,…, T) is collected for environmental investment prediction. Hence, the membership degree of each input-output data pair can be calculated based on the corresponding fuzzy region of each indicator when determining the fuzzy region with maximum membership degree, and finally generate fuzzy rules with fuzzy regions. As shown in Fig. 1, when x_1^t is {(*Small*, 0.2), (*Middle*, 0.8), (*High*, 0.0)}, x_2^t is {(*Small*, 0.0), (*Middle*, 0.3), (*High*, 0.7)}, and y^t is {(*Small*, 0.9), (*Middle*, 0.1), (*High*, 0.0)}, the tth input-output data pair of environmental investment can generate one fuzzy rule showed as follows:

$$\begin{aligned}(x_1^t, x_2^t; y^t) \Rightarrow R_t : &\left[x_1, (0.8, Middle, \max); x_2, (0.7, High, \max); y, (0.9, Small, \max)\right] \\ \Rightarrow R_t : &\ IF\ x_1\ is\ Middle\ and\ x_2\ is\ High, THEN\ y\ is\ Small\end{aligned} \tag{1}$$

Step 3: To combine generated fuzzy rules according to their fuzzy region and importance degree.

Consider that some of generated fuzzy rules have the same fuzzy regions in IF part, all of these fuzzy rules should be combined using their importance degrees to generate the fuzzy rule with unique fuzzy regions, in which the importance degree is calculated as follows:

$$D(R_t) = m_{\max}(y^t) \prod_{i=1}^{M} m_{\max}(x_i^t) \tag{2}$$

where $m_{\max}(x_i^t)$ is the maximum membership degree of the ith (i=1,…, M) input indicator x_i in the tth (t=1,…, T) rule, $m_{\max}(y^t)$ is the maximum membership degree of the output indicator y in the tth rule.

Step 4: To predict environmental investments based on the given input data and fuzzy rules.

Suppose that there is a new input data $(x_1^{T+1},...,x_M^{T+1})$ which is required to predict its corresponding environmental investment. Hence, based on the membership functions of each input indicators, $T\times M$ membership degrees $m_t(x_m^{T+1})$ (t= 1,…, T; m= 1,…, M) can be calculated. Finally, the predicted environmental investment is obtained as follows:

$$z^{T+1}=\frac{\sum_{t=1}^{T} y^t \prod_{m=1}^{M} m_t(x_m^{T+1})}{\sum_{t=1}^{T} \prod_{m=1}^{M} m_t(x_m^{T+1})} \tag{3}$$

4. Improved FRBS for Predicting Environmental Investments

In this section, an ER-based indicator selection method and a SC-based data selection method are proposed to improve the traditional FRBS for environmental investment prediction, and they are introduced in Sections 4.1 and 4.2, respectively. Afterwards, the framework of the improved FRBS, called FRBS-ERSC, is showed in Section 4.3.

4.1. ER-based indicator selection method for improving FRBS

In order to address ***Challenge 1*** pointed out in ***Section 2.3***, an ER-based indicator selection method is proposed in this section, whose main principle is to utilize various feature selection methods for independently identifying the importance of environmental indicators, and then integrate all importance of each indicator by using the ER approach [42] for better indicator selection, in which the ER approach was developed on the basis of the decision theory and Dempster-Shafer theory of evidence and it is powerful in handling information fusion under belief structure. Owing to this principle, the proposed ER-based indicator selection method can combine the advantages of different feature selection methods and is able to overcome ***Challenge 1***. The steps of the ER-based indicator selection method are showed as follows:

Step 1**:** To identify the relative importance of environmental indicators by using existing feature selection methods. Without loss of generality, suppose that FS feature selection methods are used for the importance identification of NI indicators $\{U_1,\ldots, U_{NI}\}$. Thus, FS groups of importance for NI indicators can be obtained, where $\{r_{k,1},\ldots, r_{k,NI}\}$ is assumed to be the importance of the kth (k=1,…, FS) group for NI indicators.

Step 2**:** To combine the relative importance by using the ER approach. According to the FS groups of importance for NI indicators in ***Step 1***, a set of NI assessment grades, namely $H=\{H_1,\ldots, H_{NI}\}$, can be defined to describe the belief structure, which means the distributed assessment of NI indicators. Thereafter, the kth (k=1,…, FS) distributed assessment can be obtained from the kth (k=1,…, FS) group of importance and it is denoted by the following belief distribution:

$$\{(H_h,\gamma_h^k); h=1,...,NI\},\ \gamma_h^k=\frac{r_{k,h}}{\sum_{i=1}^{NI} r_{k,i}} \tag{4}$$

where γ_h^k denotes the belief degree assigned to the hth grade H_h (or indicator U_h) at the kth group of importance for NI indicators, all the belief degrees in the kth belief distribution satisfies $\sum_{h=1}^{NI}\gamma_h^k=1$.

Next, when the relative importance of FS feature selection methods are $\{\omega_1,\ldots, \omega_{FS}\}$, the basic probability assignment (BPA) for each belief distribution can be calculated as follows：

$$m_h^k=\omega_k\gamma_h^k \tag{5}$$

$$\overline{m}_H^k=1-\omega_k, \tag{6}$$

$$\tilde{m}_H^k=\omega_k(1-\sum_{h=1}^{NI}\gamma_h^k) \tag{7}$$

where m_h^k is the BPA of the kth feature selection method in the hth assessment grade; $\overline{m}_H^k$ is the uncertain BPA caused by

the weight of the kth feature selection method; $\widetilde{m}_H^k$ is the uncertain BPA caused by the belief distribution.

By using the analytical ER algorithm [42], the FS groups of BPAs can be combined for a new group of integrated BPAs. The corresponding integration formulas are as follows:

$$m_h = K[\prod_{k=1}^{FS}(m_h^k + \overline{m}_H^k + \widetilde{m}_H^k) - \prod_{k=1}^{FS}(\overline{m}_H^k + \widetilde{m}_H^k)] \tag{8}$$

$$\widetilde{m}_H = K[\prod_{k=1}^{FS}(\overline{m}_H^k + \widetilde{m}_H^k) - \prod_{k=1}^{FS}\overline{m}_H^k] \tag{9}$$

$$\overline{m}_H = K\prod_{k=1}^{FS}\overline{m}_H^k \tag{10}$$

$$K^{-1} = \sum_{h=1}^{NI}\prod_{k=1}^{FS}(m_h^k + \overline{m}_H^k + \widetilde{m}_H^k) - (NI-1)\prod_{k=1}^{FS}(\overline{m}_H^k + \widetilde{m}_H^k) \tag{11}$$

Finally, the integrated BPAs can be transformed into belief distribution by following formula:

$$\gamma_h = \frac{m_h}{1-\overline{m}_H}. \tag{12}$$

***Step 3*:** To select effective indicators by using the integrated belief distribution. From the integrated belief distribution $\{(H_h, \gamma_h); h=1,\ldots, NI\}$, the importance of NI environmental indicators can be ranked. For example, the relative importance of NI environmental indicators is $U_1<U_2<\cdots<U_{NI}$ when the integrated belief degree is $\gamma_1<\gamma_2<\cdots<\gamma_{NI}$. As a result, the effective indicators can be selected for investment prediction modeling according to the importance ranking of all indicators

4.2. SC-based data selection method for improving FRBS

In order to overcome ***Challenge 2*** pointed out in ***Section 2.3***, an SC-based data selection method is proposed in this section, whose main principle is to select representative environmental data using the SC algorithm [8], which is a fast clustering algorithm that treats each data as a potential central point of clustering. The complexity of the algorithm is independent of the dimension of problems and has a linear relationship with the number of data. Hence, the SC-based data selection method can decrease the computing efficiency of environmental investment prediction and overcome ***Challenge 2***. The steps of the SC-based data selection method are showed as follows:

***Step 1*:** To select environmental input data based on density values. Suppose that a set of T environmental input data $\boldsymbol{X}=\{\boldsymbol{x}^t; t=1,\ldots, T\}$ is collected for environmental investment prediction. Hence, for the ith input data $\boldsymbol{x}^i$ ($\boldsymbol{x}^i\in\boldsymbol{X}$), its density value can be calculated by following formula:

$$D_i = \sum_{\boldsymbol{x}^t\in\boldsymbol{X}}\exp(-\frac{4\,\|\,\boldsymbol{x}^t-\boldsymbol{x}^i\,\|^2}{r_a^{\,2}}) \tag{13}$$

where D_i is the density value of the ith input data; r_a is the neighborhood radius which can be calculated by:

$$r_a = \frac{1}{2}\min_{\boldsymbol{x}^t\in\boldsymbol{X}}\{\max_{\boldsymbol{x}^i\in\boldsymbol{X}}\{\|\,\boldsymbol{x}^t-\boldsymbol{x}^i\,\|\}\} \tag{14}$$

Thereafter, one input data with the biggest density value can be selected from the set of input data $\boldsymbol{X}$. Without loss of generality, the input data with the biggest density value is denoted as $\boldsymbol{x}^{ti}$ ($\boldsymbol{x}^{ti}\in\boldsymbol{X}$) and its density value is D_{ti}.

***Step 2*:** To update the density value of remaining input data. Based on the selected input data $\boldsymbol{x}^{ti}$, the remaining input data can be denoted as $\boldsymbol{X}=\boldsymbol{X}-\{\boldsymbol{x}^{ti}\}$.Hence, the density value of remaining input data, taking the ith input data $\boldsymbol{x}^i$ ($\boldsymbol{x}^i\in\boldsymbol{X}$) as an example, should be updated as follows:

$$D_i = D_i - D_{ti}\exp(-\frac{4\,\|\,x^i-x^{ti}\,\|^2}{r_b^{\,2}}) \tag{15}$$

where r_b is the neighborhood radius of remaining input data and can be set as $r_b=(1.2\sim1.5)\times r_a$ to avoid the situation that the

input data are repeatedly selected because of a smaller neighborhood.

Step 3*:** To select another input data based on density values. According to the updated density value of remaining input data, another input data with the biggest density values can be selected from the set of input data ***X. Similarly, the selected input data and the set of input data should be used to update the density value of remaining input data based on ***Step 2***, until the following condition satisfies:

$$\frac{D_{tL}}{D_{t1}} < \delta \tag{16}$$

where δ $(0 \le \delta \le 1)$ is the threshold used to judge if the process of data selection should be terminated; D_{t1} denotes the density value of the first selected input data; D_{tk} denotes the density value of the kth selected input data.

For the above SC-based data selection method, the following remarks can be given:

(1) The bigger the value of δ, r_a and r_b are, the smaller the number of selected data obtain, and the resulting FBRS has few fuzzy rules. Additionally, when $\delta = 0$, the proposed method will select data as many as it could.

(2) After repeating ***Step 2*** and ***Step 3*** until the condition in Eq. (16) is met, a total of L input data can be selected from T input data. Hence, these L input data and corresponding output data should be used for investment prediction modeling.

4.3. Framework of FRBS-ERSC for environmental investment prediction

In this section, a generic framework of the improved FRBS, called FRBS-ERSC, is introduced to illustrate how to build a new model for environmental investment prediction. As shown in **Fig. 2**, the framework includes three components: ER-based indicator selection, SC-based data selection, and FRBS-based modeling. It is worth noting that the integration of three components is able to overcome the three challenges shown in ***Section 2.3***.

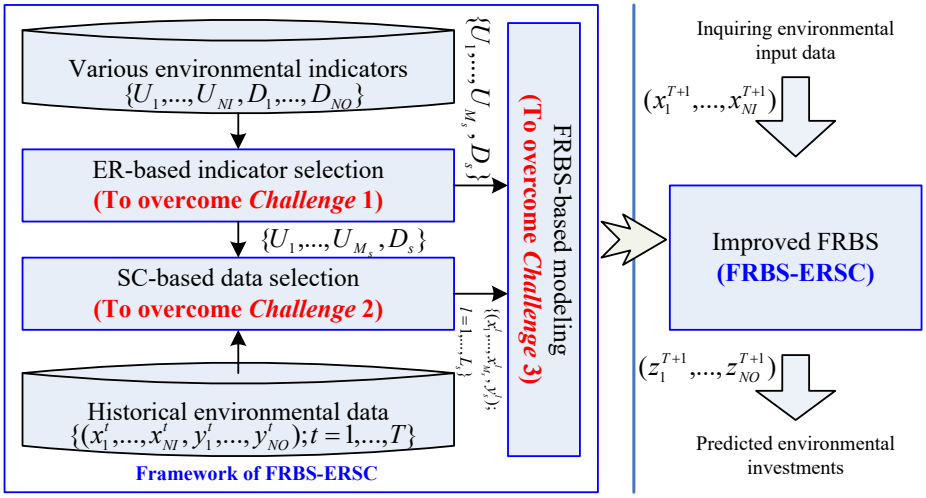

Fig. 2 Generic framework of FRBS-ERSC for environmental investment prediction

From **Fig. 2**, the following steps are provided to introduce the framework for environmental investment prediction.

Step 1: ER-based indicator selection. Suppose there are original NI environmental input indicators $\{U_1,\ldots,U_{NI}\}$ and NO environmental output indicators $\{D_1,\ldots, D_{NO}\}$. For each output indicator D_s (s=1,…, NO), M_s key input indicators, denoted as $\{U_1,\ldots, U_{Ms}\}$, can be selected according to the steps of the ER-based indicator selection method shown in ***Section 4.1***.

Step 2: SC-based data selection. Suppose there are T historical environmental data $\{(x_1^t,...,x_{NI}^t,y_1^t,...,y_{NO}^t);t=1,...,T\}$. For each set of key input and output indicators, *e.g.*, $\{U_1,\ldots, U_{Ms}, D_s\}$, the corresponding L_s key input-output data pairs $\{(x_1^l,...,x_{M_s}^l,y_s^l);l=1,...,L_s\}$ can be selected based on the steps of the SC-based data selection method shown in ***Section 4.2***.

Step 3: FRBS-based modeling. Based on the selected key indicators and key data, L_s fuzzy rules can be generated from

the selected L_s key input-output data pairs according to ***Step 1*** to ***Step 2*** of traditional FRBS shown in ***Section 3***. Here, it is unnecessary to combine the generated fuzzy rules, as shown in ***Step 3*** of traditional FRBS, because all these fuzzy rules are generated from key environmental data. Finally, a total of *NO* FRBS-ERSCs can be constructed.

Step 4: To predict environmental investments. When *NO* FRBS-ERSCs are constructed, the predicted environmental investments z_s^{T+1} (s=1,…, *NO*) can be obtained for the inquiring environmental input data $(x_1^T,...,x_{NI}^T)$ by using ***Step 4*** of traditional FRBS shown in ***Section 3***.

5. Case study

In order to verify the effectiveness of the proposed FRBS-ERSC on environmental investment prediction, the regional data of environmental management of 31 provinces in China are utilized to carry out experimental study in this section.

5.1. Data source and indicator determination

The historical data related with 31 provinces in the mainland of China are used to verify the effectiveness of the FRBS-ERSC for environmental investment prediction, in which these data ranges from 2005 to 2018 *China Statistical Yearbook* and *China Environmental Statistical Yearbook*, respectively, which are the most commonly used and reliable public database for the study of environmental management in China [5][19][44].

According to the previous studies on the environmental management and investment prediction of China [7][43][47] [48][49], ten environmental input indicators, which are adopted from previous studies, and the commonly used three environmental output indicators are applied for environmental investment prediction modeling. The specification of these input and output indicators are shown in **Table 1**. From **Table 1,** it has significant regional differences in input and output indicators for 31 provinces in China in terms of the maximum and minimum values. For example, the minimum value of SO_2 is only 1000, while the maximum value of SO_2 is 2002000, which is more than 2000 times of the minimum value. The maximum value of EC is 5959, while the minimum value is 9. What's more, it is obvious that there has significantly regional difference in economic development of different provinces in China, for example, the minimum value of TP is -91.89, while the maximum value of TP is 10574.

Table 1 Introduction of input and output indicators in investment prediction

No.	Indicator name	Abbr.	Specific interpretation of indicators	Symbol	Min	Average	Max
1	gross domestic product	GDP	Value of gross domestic product	U_1	220.34	15626.84	89705.23
2	Total profit	TP	Total profit of Enterprises above Designated Size	U_2	-91.89	1530.90	10574.40
3	Garbage clean-up	GCU	Garbage removal and transportation volume	U_3	16.30	547.74	2644.50
4	Sulfur dioxide	SO_2	Emission of sulfur dioxide	U_4	1000	662937	2002000
5	Smoke and dust	SM	Emission of smoke and dust	U_5	1000	364270	1797683
6	Carbon dioxide	CO_2	Emission of carbon dioxide	U_6	7.07	1096.90	4677.79
7	Waste water	WW	Total emission of waste water	U_7	2685	201796	938261
8	Chemical oxygen demand	COD	Emission of chemical oxygen demand	U_8	1.38	53.60	198.25
9	Lead emission	LE	Lead emission in waste water	U_9	0.002	1328.511	42466.480
10	Petroleum emissions	PE	Petroleum emissions in waste water	U_{10}	0.03	497.75	2937.40
11	Energy consumption	EC	Total electricity consumption	D_1	9.00	1397.33	5958.97
12	Capital investment	CI	Fixed assets investment	D_2	162.36	10306.17	55202.72
13	Labor investment	LI	Total number of employees	D_3	15.00	469.31	1973.28

Additionally, to validate the proposed FRBS-ERSC, the leave-one-out cross-validation, namely the data of each year as testing data in turn and the data of remaining years as training data, is used to construct environmental investment models. The following three criteria including mean absolute error (MAE), mean absolute percentage error (MAPE), and correlation coefficient (R) are applied to evaluate the performance of the models

$$MAE = \frac{\sum_{t=1}^{T} | z^t - y^t |}{T} \tag{17}$$

$$MAPE = \frac{1}{T} \sum_{t=1}^{T} \left| \frac{z^t - y^t}{y^t} \right| \times 100\% \tag{18}$$

$$R = \frac{\sum_{t=1}^{T} (z^t - \bar{z})(y^t - \bar{y})}{\sqrt{\sum_{t=1}^{T} (z^t - \bar{z})^2 \cdot \sum_{t=1}^{T} (y^t - \bar{y})^2}} \tag{19}$$

where

$$\bar{z} = \frac{1}{T} \sum_{t=1}^{T} z^t, \ \bar{y} = \frac{1}{T} \sum_{t=1}^{T} y^t \tag{20}$$

where z^t and y^t denote the tth (t=1,…, T) the predicted and real values of environmental investments; T is the total number of data. Here, the larger R and the smaller MAPE and MAE are considered to be a better performance for an investment prediction model.

5.2. Development procedure of FRBS-ERSC for investment prediction

In this section, the procedure of developing a FRBS-ERSC for environmental investment prediction is analyzed via three steps, including ER-based indicator selection, SC-based data selection, and FRBS-based modeling.

5.2.1 Development procedure of ER-based indicator selection

To effectively select key input indicators for each output indicator in environmental investment prediction, five kinds of indicator selection methods are introduced for identifying the relative importance of environmental input indicators, namely Pearson's Correlation (PC)-based, ReliefF Algorithm (RA)-based, Simple Linear Regression (SLR)-based, Entropy Algorithm (EA)-based, and Correlation Coefficient Standard Deviation (CCSD)-based indicator selection methods. Note that the former three methods are implemented by WEKA software and the latter two methods are obtained from [34][41].

According to ***Step 1*** in ***Section 4.3***, the relative importance of ten input indicators should be identified in terms of each output indicator using five indicator selection methods. Taking the prediction of CI as example, the relative importance of ten input indicators obtained from five methods are showed in **Table 2**. In the view of indicators' relative importance, it can be found that GDP and TP are the most important indicators based on PC-based and SLR-based methods, SO_2 and COD are the most important indicators based on RA-based and EA-based methods, COD and GDP are the most important indicator based on CCSD-based methods. This is because any kind of method has its strengths to select key input indicators so that the most important indicators are not exactly the same for the five methods. Additionally, the relative importance of LE and PE are lower than other indicators, indicating that the emission of LE and PE are lower than that of other pollutants.

Table 2 Relative importance of ten input indicators for CI prediction by five indicator selection methods

Methods	Input indicators									
	GDP	TP	GCU	SO_2	SM	CO_2	WW	COD	LE	PE
PC	0.8920	0.8520	0.6130	0.1770	0.3650	0.8230	0.6740	0.5490	0.1930	0.1430
CCSD	0.0293	0.0276	0.0227	0.0202	0.0199	0.0196	0.0207	0.0296	0.0108	0.0118
RA	0.0912	0.0987	0.0809	0.1368	0.0727	0.1077	0.0930	0.1153	0.0326	0.0715
EA	0.0760	0.0827	0.0691	0.1845	0.0683	0.1172	0.0938	0.1342	0.0200	0.0623
SLR	5524.9994	4812.7241	2084.0329	131.1909	634.8454	4342.1303	2599.1132	1647.5836	142.9804	48.5799

From **Table 2** and the calculation of belief degrees shown in ***Step 2*** of ***Section 4.1***, the relative importance of ten input indicators obtained from five indicator selection methods can be transformed into belief degrees, when the set of assessment grades corresponds to the set of indicators, namely $H=\{H_1,\ldots, H_{10}\}=$ {GDP, TP, GCU, SO_2, SM, CO_2, WW, COD, LE, PE}. As a result, the belief degrees of each grade and its integrated belief degrees are shown in **Table 3**.

Table 3 Belief degree (rankings) of ten assessment grades for five indicator selection methods

Methods	Input indicators (Assessment grades)									
	GDP (H_1)	TP (H_2)	GCU (H_3)	SO_2 (H_4)	SM (H_5)	CO_2 (H_6)	WW (H_7)	COD (H_8)	LE (H_9)	PE (H_{10})
PC	0.1689 (1)	0.1613 (2)	0.1161 (5)	0.0335 (9)	0.0691 (7)	0.1558 (3)	0.1276 (4)	0.1040 (6)	0.0365 (8)	0.0271 (10)
CCSD	0.1381 (2)	0.1301 (3)	0.1070 (4)	0.0952 (6)	0.0938 (7)	0.0924 (8)	0.0975 (5)	0.1395 (1)	0.0509 (10)	0.0556 (9)
RA	0.1013 (6)	0.1096 (4)	0.0898 (7)	0.1519 (1)	0.0808 (8)	0.1196 (3)	0.1033 (5)	0.1281 (2)	0.0362 (10)	0.0794 (9)
EA	0.0837 (6)	0.0911 (5)	0.0761 (7)	0.2031 (1)	0.0752 (8)	0.1291 (3)	0.1033 (4)	0.1477 (2)	0.0221 (10)	0.0686 (9)
SLR	0.2515(1)	0.2191 (2)	0.0949 (5)	0.0060 (9)	0.0289 (7)	0.1977 (3)	0.1183 (4)	0.0750 (6)	0.0065 (8)	0.0022 (10)
Integrated	0.1510 (1)	0.1442 (2)	0.0961 (7)	0.0966 (6)	0.0681 (8)	0.1407 (3)	0.1099 (5)	0.1192 (4)	0.0292 (10)	0.0450 (9)

According to the integrated belief degrees shown in **Table 3**, the importance of ten input indicators can be ranked and they are GDP (0.1510) > TP (0.1442) > CO_2 (0.1407)> WW (0.1099) > SO_2 (0.0966) > GCU (0.0961) > SM (0.0681) > PE (0.0450) > LE (0.0292). By comparing the rankings and belief degrees shown in **Table 3**, it can be found that the integrated belief degree of input indicators is the comprehensive results of individual belief degrees. For example, GDP is the most important indicator for CI prediction according to PC-based and SLR-based methods, and the 2nd important indicator for CI prediction according to CCSD-based method, thus the integrated belief degree of GDP is bigger than other input indicators. Accordingly, SM, PE, and LE are relatively insignificant input indicators for CI prediction according to the integrated belief degrees and rankings, and this relationship can be also found in five indicator selection methods.

5.2.2 Development procedure of SC-based data selection

To effectively select key historical data for environmental investment prediction, the proposed SC-based data selection method should be performed. Without loss of generality, the thresholds of SC-based data selection method shown in ***Section 4.2*** are set as the number of selected indicators $M = 4$ (namely the top four input indicators GDP, TP, CO_2, and WW), $r_b = 1.2\times r_a$, and $\delta = 0$, respectively. For the sake of convenience description, the CI prediction of Fujian at 2018 (namely the historical data derived from 2005-2017 as the training data) is taken as an example to illustrate the development procedure of the SC-based data selection method. **Table 4** shows a total 9 iterations of data selection using the calculated density values and the selected historical data for environmental investment prediction.

Table 4 Density value of the data derived from 2005 to 2017 for Fujian's CI prediction

Year (t)	Density values (r_a = 0.5664; r_b = 0.6797)									Selection
	1st	2nd	3th	4th	5th	6th	7th	8th	9th	
2017	1.0011 (13)	1.0011 (8)	1.0001 (4)	1.0001 (2)	1.0001 (1)	-	-	-	-	Yes
2016	1.9872 (11)	1.9872 (5)	0.9686 (5)	0.9686 (3)	0.9495 (3)	0.9417 (2)	0.9416 (1)	-	-	Yes
2015	2.5235 (8)	2.5235 (2)	0.5530 (9)	0.5530 (6)	0.4631 (4)	0.4612 (3)	0.4612 (2)	-0.2806(4)	-0.2806(1)	No
2014	2.7140 (6)	2.7140 (1)	-	-	-	-	-	-	-	Yes
2013	2.5174 (9)	2.5174 (3)	0.5958 (8)	0.5958 (5)	-0.2129(8)	-0.2129(7)	-0.2131(5)	-0.2955(5)	-0.2955(3)	No
2012	1.9915 (10)	1.9915 (4)	1.0098 (3)	1.0098 (1)	-	-	-	-	-	Yes
2011	1.5170 (12)	0.9660 (9)	0.9656 (6)	0.9612 (4)	0.9611 (2)	0.9610 (1)	-	-	-	Yes
2010	3.1119 (5)	0.3014 (11)	0.3014 (10)	0.1819 (7)	0.1819 (5)	0.1819 (4)	-0.2599(6)	-0.2599(3)	-0.2855(2)	No
2009	3.6349 (3)	0.0580 (12)	0.0580 (11)	-0.2270(10)	-0.2270(9)	-0.2270(8)	-0.4512(7)	-0.4512(6)	-0.5068(5)	No
2008	3.8451 (1)	-	-	-	-	-	-	-	-	Yes
2007	3.7480 (2)	0.7989 (10)	0.7989 (7)	-0.1307 (9)	-0.1307(7)	-0.1307(6)	-0.1565(4)	-0.1565(2)	-0.3088(4)	No
2006	3.2685 (4)	1.3707 (7)	1.3707 (2)	0.1771 (8)	0.1771 (6)	0.1771 (5)	0.1703 (3)	0.1703 (1)	-	Yes
2005	2.6725 (7)	1.4060 (6)	1.4060 (1)	-	-	-	-	-	-	Yes

From **Table 4**, it is clear that the data in 2008 is the first key data selected for Fujian's CI prediction because its density value is bigger than the density value of other data. Similarly, the data with the biggest density value is selected as the key data of each iteration according to the SC-based data selection methods. After 9 iterations, the data derived from 2005, 2006, 2008, 2011, 2012, 2014, 2016, and 2017 are selected to perform the FRBS-based modeling of Fujian's EC prediction.

5.2.3 Development procedure of FRBS-based modeling

Continuing with the case of Fujian's CI prediction in 2018, the selected input indicators, namely GDP, TP, CO_2, and WW, and the selected environmental data are used to illustrate the development procedure of the FRBS-based modeling. Firstly, each input indicator and output indicator are all set as two triangular fuzzy regions, *e.g.* {*Low*, *High*}. Based on **Table 1**, their fuzzy numbers can be defined as *Low*(GDP)=(-89264.60, 220.34, 89705.23), *High*(GDP)=(220.34, 89705.23, 179190.10), *Low*(TP)=(-10758.20, -91.89, 10574.40), *High*(TP)=(-91.89, 10574.40, 21240.69), *Low*(CO_2)=(-4663.66, 7.07, 4677.79), *High*(CO_2)=(7.07, 4677.79, 9348.52), *Low*(WW)=(-932891.03, 2685.00, 938261.03), *High*(WW)=(2685.00, 938261.03, 1873837.06), *Low*(CI) =(-54878.00, 162.36, 55202.72), and *High*(CI)=(162.36, 55202.72, 110243.08). Consequently, **Table 5** shows the fuzzy membership degrees of the historical data selected from ***Section 5.2.2***.

Table 5 Fuzzy membership degree of the selected data for Fujian's CI prediction

Year (t)	GDP (x_1^t)		TP (x_2^t)		CO_2 (x_3^t)		WW (x_4^t)		CI (y_2^t)		y_t
	Low	*High*	*Low*	*High*	*Low*	*High*	*Low*	*High*	*Low*	*High*	
2017	0.1276	0.8724	0.1171	0.8829	0.0995	0.9005	0.8969	0.1032	0.1296	0.8704	23237.35
2016	0.2348	0.7652	0.3035	0.6965	0.1801	0.8199	0.2181	0.7819	0.2086	0.7914	21301.38
2014	0.3904	0.6096	0.3502	0.6499	0.2845	0.7155	0.1259	0.8741	0.4522	0.5478	15327.44
2012	0.5535	0.4465	0.3899	0.6101	0.4121	0.5879	0.0000	1.0000	0.6731	0.3270	9910.89
2011	0.6603	0.3397	0.5168	0.4832	0.5507	0.4493	0.9548	0.0452	0.7429	0.2572	8199.12
2008	0.8681	0.1319	0.8195	0.1805	0.7680	0.2320	0.9237	0.0763	0.9023	0.0977	4287.75
2006	0.9701	0.0300	0.9910	0.0090	0.9363	0.0637	0.8894	0.1106	0.9827	0.0173	2316.72
2005	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000	1892.92

From **Table 5**, the fuzzy regions with the biggest fuzzy membership degree in each indicator can constitute a fuzzy rule. Taking the data in 2017 as an example, the indicators GDP, TP, CO_2, and CI have the biggest fuzzy membership degree in fuzzy region *High*, and the indicator WW has the biggest fuzzy membership degree in fuzzy region *Low*, thus the fuzzy rule can be written as "IF GDP is *High* and TP is *High* and CO_2 is *High* and WW is *Low*, THEN CI is *High*". Similarly, the resulting fuzzy rules are all shown in **Table 6**. It is worth noting that all fuzzy rules of FRBS-ERSC are generated from the data selected by the SC-based data selection method, so it is unnecessary to combine fuzzy rules based on the same fuzzy regions in IF part. All fuzzy rules shown in **Table 6** can be used to predict the CI of Fujian in 2018.

Table 6 Fuzzy rules for Fujian's CI prediction

Rule no. (t)	GDP (x_1^t)	TP (x_2^t)	CO_2 (x_3^t)	WW (x_4^t)	CI (y^t)	y_t
2017	*High*	*High*	*High*	*Low*	*High*	23237.35
2016	*High*	*High*	*High*	*High*	*High*	21301.38
2014	*High*	*High*	*High*	*High*	*High*	15327.44
2012	*Low*	*High*	*High*	*High*	*Low*	9910.89
2011	*Low*	*Low*	*Low*	*Low*	*Low*	8199.12
2008	*Low*	*Low*	*Low*	*Low*	*Low*	4287.75
2006	*Low*	*Low*	*Low*	*Low*	*Low*	2316.72
2005	*Low*	*Low*	*Low*	*Low*	*Low*	1892.92

5.3. Investment prediction for different environmental output indicators

In this section, the FRBS-ERSC is applied to predict the investment of three output indicators, namely EC, CI and LI, according to leave-one-out cross-validation. Through the same development procedure shown in ***Section 5.2***, 31×13×3 FRBS-ERSCs can be constructed for 31 provinces, 13 years, and 3 output indicators, respectively. On basis of these FRBS-ERSCs, **Fig. 2** to **Fig. 4** show the predicted and the real values of summational investments from 2005 to 2018 for each province and output indicator.

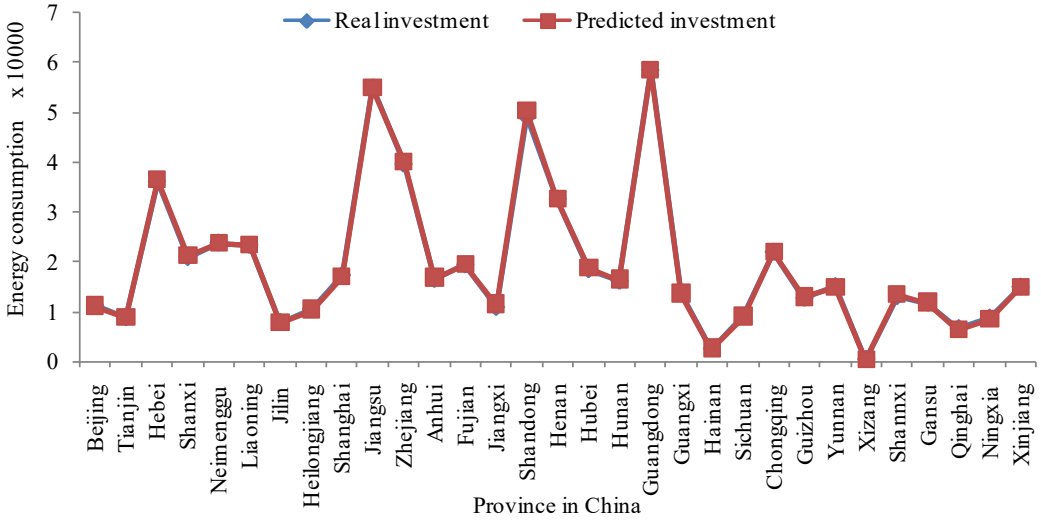

Fig. 2 Predicted and real summational ECs of 31 provinces under leave-one-out cross-validation

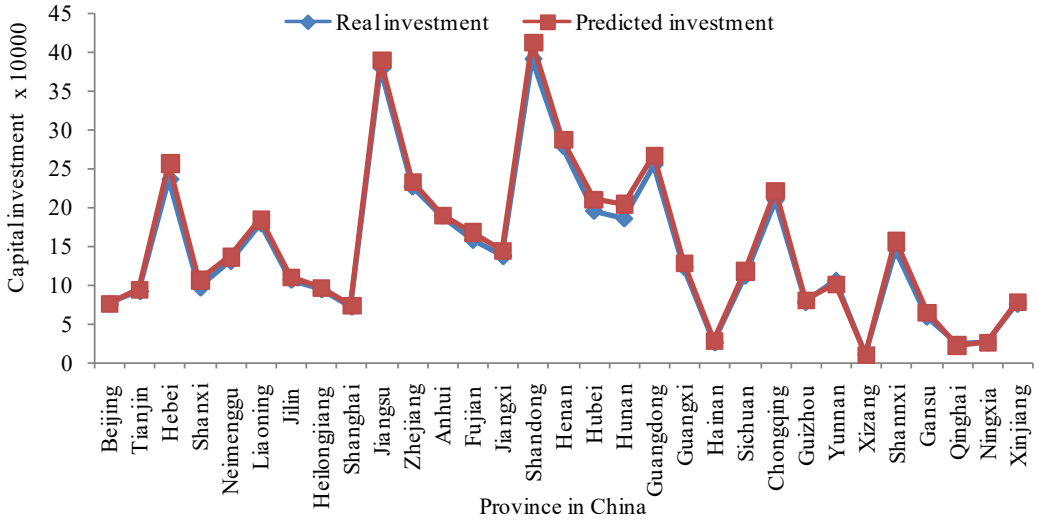

Fig. 3 Predicted and real summational CIs of 31 provinces under leave-one-out cross-validation

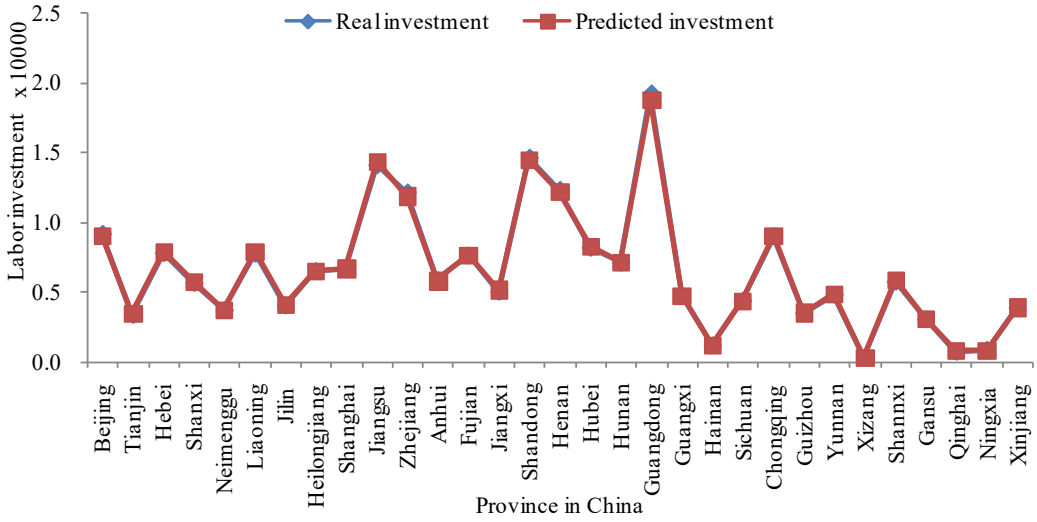

Fig. 4 Predicted and real summational LIs of 31 provinces under leave-one-out cross-validation

From **Fig. 2** to **Fig. 4**, it is obvious that three kinds of predicted summational investments basically match their real summational investments in most provinces. Although the predicted CIs in Hubei and Hunan have a little difference to the real CIs, the predicted CIs of the remaining provinces matches the real CIs accurately. From the view of regional difference, the investments of EC in Shandong, Guangdong, and Jiangsu are obviously higher than other provinces, and the investments of EC in the most western regions are lower than the eastern region in China, the reason is that the most eastern provinces have large population and industry, especially for the population and industry development in Shandong, Guangdong, and Jiangsu. Additionally, the predicted CIs and LIs shown in **Fig. 3** and **Fig.4** indicate that most eastern provinces in China have higher investments than the most provinces located in Western China, which reveal that environmental investments have a great relationship with population, economy, and government's emphasis on environmental management in China.

5.4. Comparative analysis of existing investment prediction models

In order to demonstrate the function of the ER-based indicator selection and SC-based data selection to improve a FRBS, the original FRBS, the FRBS considering either these two parts (denoted as FRBS-ER and FRBS-SC) are introduced to compare the accuracy of a FRBS-ERSC. The comparative results are shown in **Table 7**. It can be found that the FRBS-ERSC has higher accuracy than other models in terms of MAE, in which the MAE of the FRBS-ERSC for EC, CI and LI is 118.5628, 1931.4011, and 28.3688, respectively. For the comparison of MAPE and R, although the MAPE of the FRBS and FRBS-SC for CI and the R of the FRBS-ER for EC and CI are slightly better that of the FRBS-ERSC, the performance of

the FRBS-ERSC outperforms the FRBS, FRBS-SC, and FRBS-ER in the remaining items of MAPE and R. Additionally, by comparing the FRBS with the FRBS-SC and FRBS-ER, it can be also found that the FRBS usually obtains the worst prediction results. Therefore, the comparative analysis in **Table 7** indicates that either the ER-based indicator selection or the SC-based data selection can improve the accuracy of the original FRBS. Furthermore, both of them can be used together to achieve the maximum improvement of the original FRBS.

Table 7 Comparison of different FRBSs with considering indicator selection and/or data selection

	Investment	FRBS	FRBS-SC	FRBS-ER	FRBS-ERSC
MAE	EC	178.3189	176.2234	133.2424	118.5628
	CI	2199.8769	2153.0272	2039.5363	1931.4011
	LI	42.4041	41.9176	28.5968	28.3688
MAPE	EC	14.5886%	14.3545%	11.6235%	10.3938%
	CI	29.9668%	29.3920%	31.3238%	30.8341%
	LI	8.5090%	8.4321%	5.8671%	5.7530%
R	EC	0.9447	0.9435	0.9959	0.9955
	CI	0.9418	0.9339	0.9800	0.9730
	LI	0.9308	0.9310	0.9917	0.9918

In order to compare the prediction accuracy of the FRBS-ERSC with some existing investment prediction models, including ARIMA-based model [15], GM-based model [7], traditional FRBS-based model [48], and EBRBS-based model [43], in which the former two models are the typical representative based on time series forecasting and the latter two models are the typical representative based on input-output relationship. Additionally, to ensure the fairness of each investment prediction model, none of learning mechanism is applied to train the parameters of the models. **Table 8** shows the comparison of different models in investment prediction based on MAE, MAPE, and R. It is clear from **Table 8** that the predicted investments of the FRBS- ERSC are closer to the three real investments than other models, and the models based on input- output relationship are better than the models based on time series forecasting. In summary, the whole prediction accuracy of the FRBS-ERSC can outperform all listed studies and has higher ability than the models proposed in previous studies to accurately predict environmental investments

Table 8 Comparison of different investment prediction models

	Investment	ARIMA[15]	GM[7]	FRBS[48]	EBRBS[43]	FRBS-ERSC
MAE	EC	1351.6211	539.8448	145.7149	150.6405	118.5628
	CI	29306.9363	17094.4526	1982.5927	2112.296	1931.4011
	LI	265.3371	591.9844	30.3818	31.8017	28.3688
MAPE	EC	123.3895%	61.3806%	11.5171%	14.7302%	10.3938%
	CI	480.7537%	311.9239%	24.1034%	27.1555%	30.8341%
	LI	50.8216%	83.2448%	6.0614%	6.3515%	5.7530%
R	EC	0.9581	0.8335	0.9947	0.9765	0.9955
	CI	0.8635	0.5261	0.9742	0.9172	0.9730
	LI	0.8821	0.6944	0.9901	0.9691	0.9918

6. Conclusions

In this study, a new environmental investment prediction model, called FRBS-ERSC, was proposed to predict different kinds of environmental investments based on enhanced indicator selection and data selection. A case study of 31 provinces in China using the real data illustrated the process of investment prediction using FRBS-ERSC and the comparative analysis to verify its effectiveness. The main conclusions of this study are further summarized as three aspects:

(1) Previous investment prediction modeling focused on single feature selection method in representative indicator selection, the proposed FRBS-ERSC provided a new strategy to collaboratively select indicators by using various kinds of existing feature selection methods and utilized the ER approach as a new ranking combination method, which avoided the redundant indicator information and irrelevant indicator selection in investment prediction. Based on the above contribution, the FRBS-ERSC has an effective indicator selection for environmental investment prediction.

(2) The FRBS-ERSC is generated from the input-output sample data of environmental management, in which the generation of fuzzy rules is under the consideration of representative environmental data selection using the SC algorithm. This reduced the complexity in data selection and decreased the complexity of rule generation. Moreover, the prediction process of environmental investment using the FRBS-ERSC is interpretable, so that it is convenient for decision makers to make a clear investment prediction scheme according to the investment predicted by the FRBS-ERSC.

(3) The case study on the data for 31 provinces in China from 2005 to 2018 verified several advantages of the proposed model: 1) from the comparative analysis, the FRBS-ERSC is more reasonable and effective than the FRBS which did consider indicator selection and/or data selection; 2) the FRBS-ERSC has a higher accuracy than those prediction models proposed in previous studies for environmental investment prediction; 3) the investments predicted by the FRBS-ERSC is much more reasonable and suitable for the decision making of actual investment planning better than some existing models.

For the future research initiatives, it can devote on the combination of efficiency evaluation and investment prediction to improve the effectiveness of environmental management. Meanwhile, few studies focused on the investment prediction in the field of economic development and industrial investment on environmental protection, future research can also devote in environmental investment prediction of different industries.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Nos. 72001043, 61773123, and 72001042), the Natural Science Foundation of Fujian Province of China (No. 2020J05122), the Humanities and Social Science Foundation of the Ministry of Education of China (No. 20YJC630188), the Social Science Foundation of Fujian Province of China (No. FJ2019C032), the Chengdu International Science Cooperation Project (No. 2020-GH02-00064-HZ), and the Research Grants Council, Hong Kong SAR, China (No. 15218919)

Reference:

[1] Alam S. M. M., Ali M. H., 2020, A New Subtractive Clustering Based ANFIS System for Residential Load Forecasting, IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Washington, DC, USA,

[2] Bautista S., Enjolras M., Narvaez P., Camargo M., Morel L., 2016, Biodiesel-triple bottom line (TBL): A new hierarchical sustainability assessment framework of principles criteria & indicators (PC&I) for biodiesel production. Part II-validation, *Ecological Indicators*, 68: 803-817.

[3] Benmouiza K., Cheknane A., 2019, Clustered ANFIS network using fuzzy c-means, subtractive clustering, and grid partitioning for hourly solar radiation forecasting. *Theoretical and Applied Climatology*, 137: 31-43.

[4] Bui D. T., Tsangaratos P., Ngo P. T. T., Pham T. D., Pham B. T., 2019, Flash flood susceptibility modeling using an optimized fuzzy rule based feature selection techniques and tree based ensemble methods. *Science of the Total Environment*, 668: 1038-1054.

[5] Cao H. J., Fujii H., Managi S., 2015. A productivity analysis considering environmental pollution and diseases in China. *Journal of Economic Structures*, 4(6):11-25.

[6] Chahnasir E. S., Zandi Y., Dehghani E., Mohamd E. T., Shariati A., Safa M., Wakil K., Khorami M., 2018, Application of support vector machine with firefly algorithm for investigation of the factors affecting the shear strength of angle shear connectors, *Smart Structures and Systems*, 22(4): 413-424.

[7] Chen L., Wang Y. M., Lai F. J., Feng F., 2017. An investment analysis for China's sustainable development based on inverse data envelopment analysis. *Journal of Cleaner Production*, 142, 1638-1649.

[8] Chiu S. L., 1994. Fuzzy Model Identification Based on Cluster Estimation, *Journal of Intelligent and Fuzzy Systems*, 2: 267-278.

[9] Deus R. M., Mele F. D., Bezerra B. S., Battistelle R. A. G., 2020, A municipal solid waste indicator for environmental impact assessment and identification of best management practices, *Journal of Cleaner Production*, 242: 1-14.

[10] Dymova L., Sevastjanov P., Kaczmarek K., 2016, A Forex trading expert system based on a new approach to the rule-base evidential reasoning, *Expert Systems with Applications*, 51: 1-13.

[11] Dymova L., Sevastjanov P., Kaczmarek K., 2012, A stock trading expert system based on the rule-base evidential reasoning using Level 2 Quotes, *Expert Systems with Applications*, 39(8): 7150-7157.

[12] Elkano M., Galar M., Sanz J., Bustince H., 2018, CHI-BD: A fuzzy rule-based classification system for Big Data Classification problems, *Fuzzy Sets and Systems*, 348: 75-101.

[13] Heslouina C., Perrot-Bernardet V., Cornier A., Perry N., 2017. A user oriented framework to support environmental performance indicators selection. *Procedia CIRP*, 61,709-714.

[14] Jiang Q., Liu Z., Liu W., Li T., Cong W., Zhang H., Shi J., 2018, A principal component analysis based three-dimensional sustainability assessment model to evaluate corporate sustainable performance, *Journal of Cleaner Production*, 187: 625-637.

[15] Jiang S., Yang C., Guo J., Ding Z., ARIMA forecasting of China's coal consumption, price and investment by 2030, Energy Sources, Part B: Economics, Planning, and Policy, 2018, 13(3): 190-195.

[16] Jović S., Danesh A. S., Younesi E., Anicic O., Pavlović D., Shamshirband S., 2016, Forecasting of Underactuated Robotic Finger Contact Forces by Support Vector Regression Methodology, *International Journal of Pattern Recognition and Artificial Intelligence*, 30(7): 1-11.

[17] Kaytez F., 2020, A hybrid approach based on autoregressive integrated moving average and least-square support vector machine for long-term forecasting of net electricity consumption, *Energy*, 197: 1-12.

[18] Kerr-Wilson J., Pedrycz W., 2020. Generating a hierarchical fuzzy rule-based model, *Fuzzy Sets and Systems*, 381: 124-139.

[19] Lee M., Zhang N., 2012. Technical efficiency, shadow price of carbon dioxide emissions, and substitutability for energy in the Chinese manufacturing industries. *Energy Economics*, 34 (5), 1492–1497.

[20] Liu J., Yang J. B., Wang J., Sii H. S., Wang Y. M., 2004, Fuzzy Rule-Based Evidential Reasoning Approach for Safety Analysis, *International Journal of General Systems*, 33(2-3): 183-204.

[21] Mehrabi P., Nguyen-Thoi T., Wakil K., Khorami M., Shariati M., Safa M., 2019, Moment-rotation estimation of steel rack connection using extreme learning machine, *Steel and Composite Structures*, 31(5): 427-435.

[22] Mohammadhassani M., Nezamabadi-pour H., Suhatril M., Shariati M., 2013, Identification of a suitable ANN architecture in predicting strain in tie section of concrete deep beams, *Structural Engineering and Mechanics*, 46(6): 853-868.

[23] Mohammadhassani M., Nezamabadi-pour H., Suhatril M., Shariati M., 2014, An evolutionary fuzzy modelling approach and comparison

of different methods for shear strength prediction of high-strength concrete beams without stirrups, *Smart Structures and Systems*, 14(5): 785-809.

[24] Nikolić V., Mitić V. V., Kocić L., Petković D., 2017, Wind speed parameters sensitivity analysis based on fractals and neuro-fuzzy selection technique, *Knowledge and Information Systems*, 52(1): 255-265.

[25] Petković D., Ćojbašić Ž., Nikolić V., 2013, Adaptive neuro-fuzzy approach for wind turbine power coefficient estimation, *Renewable and Sustainable Energy Reviews*, 28: 191-195.

[26] Petković D., Ćojbašić Ž., Nikolić V., Shamshirband S., Mat Kiah M. L., Anuar N. B., Abdul Wahab A. W., 2014, Adaptive neuro-fuzzy maximal power extraction of wind turbine with continuously variable transmission, *Energy*, 64: 868-874.

[27] Petković D., Nikolić V., Mitić V. V., Kocić L., 2017, Estimation of fractal representation of wind speed fluctuation by artificial neural network with different training algorothms, *Flow Measurement and Instrumentation*, 54: 172-176.

[28] Petković D., Pavlović N. T., Ćojbašić Ž., 2016, Wind farm efficiency by adaptive neuro-fuzzy strategy, *International Journal of Electrical Power & Energy Systems*, 81: 215-221.

[29] Razani M., Yazdani-Chamzini A., Yakhchali S. H., 2013, A novel fuzzy inference system for predicting roof fall rate in underground coal mines, *Safety Science*, 55: 26-33.

[30] Safa M., Shariati M., Ibrahim Z., Toghroli A., Baharom S. B., Nor N. M., Petković D., 2016, Potential of adaptive neuro fuzzy inference system for evaluating the factors affecting steel-concrete composite beam's shear strength, *Steel and Composite Structures*, 21(3): 679-688.

[31] Salcedo-Sanz S., Cornejo-Bueno L., Prieto L., Paredes D., Garcia-Herrera R., 2018, Feature selection in machine learning prediction systems for renewable energy applications, *Renewable and Sustainable Energy Reviews*, 90: 728-741.

[32] Sedghi Y., Zandi Y., Shariati M., Ahmadi E., Azar V. M., Toghroli A., Safa M., Mohamd E. T., Khorami M., Wakil K., 2018, Application of ANFIS technique on performance of C and L shaped angle shear connectors, *Smart Structures and Systems*, 22(3): 335-340.

[33] Shamshirband S., Petković D., Amini A., Anuar N. B., Nikolić V., Ćojbašić Ž., Mat Kiah M. L., Gani A., 2014, Support vector regression methodology for wind turbine reaction torque prediction with power-split hydrostatic continuous variable transmission, *Energy*, 67: 623-630.

[34] Shannon C. E., 1948. A mathematical theory of communication, *The Bell System Technical Journal*, 27(3): 379-423.

[35] Shariati M., Ramli Sulong N. H., Shariati A., Kueh A. B. H., 2016, Comparative performance of channel and angle shear connectors in high strength concrete composites: An experimental study, *Construction and Building Materials*, 120: 382-392.

[36] Song M. L., Cao S. P., Wang S. H., 2019. The impact of knowledge trade on sustainable development and environment-biased technical progress. *Technological Forecasting & Social Change*, 144, 512-523.

[37] Toghroli A., Mohammadhassani M., Shariati M., Suhatril M., Lbrahim Z., Sulong N. H. R., 2014, Prediction of shear capacity of channel shear connectors using the ANFIS model, *Steel and Composite Structures*, 17(5): 1-16.

[38] Tong C., Ding S., Wang B., Yang S. L., 2020, Assessing the target-availability of China's investments for green growth using time series prediction, *Physica A*, 537: 1-11.

[39] Trung N. T., Shahgoli A. F., Zandi Y., Shariati M., Wakil K., Safa M., Khorami M., 2019, Moment-rotation Prediction of Precast Beam-to-column Connections using Extreme Learning Machine, *Structural Engineering and Mechanics*, 70(5): 639-647.

[40] Wang S., Ye F. F., 2020, Environmental Governance Cost Prediction of Transportation Industry by Considering the Technological Constraints, *Symmetry-BASEL*, 12: 1-15.

[41] Wang Y. M., Luo Y., 2010. Integration of correlations with standard deviations for determining attribute weights in multiple attribute decision making, *Mathematical and Computer Modelling*, 51(1-2): 1-12.

[42] Wang Y. M., Yang J. B., Xu D. L., 2006. Environmental impact assessment using the evidential reasoning approach, *European Journal of Operational Research*, 174(3): 1885-1913.

[43] Wang Y. M., Ye F. F., Yang L. H., 2020. Extended belief rule based system with joint learning for environmental governance cost prediction. *Ecological Indicators*, 111, 1-14.

[44] Wu J., Li M. J., Zhu Q. Y., Zhou Z. X., Liang L., 2019. Energy and environmental efficiency measurement of China's industrial sectors: A DEA model with non-homogeneous inputs and outputs. *Energy Economics*, 78, 468-480.

[45] Xu N., Dang Y. G., Gong Y. D., 2017. Novel grey prediction model with nonlinear optimized time response method for forecasting of electricity consumption in China. *Energy*, 118, 473-480.

[46] Yang L. H., Ye F. F., Wang Y. M., 2020, Ensemble belief rule base modeling with diverse attribute selection and cautious conjunctive rule for classification problems, *Expert Systems with Applications*, 146: 1-14.

[47] Ye F. F., Yang L. H., Wang Y. M., 2019. A new environmental governance cost prediction method based on indicator synthesis and different risk coefficients. *Journal of Cleaner Production*, 212, 548-566.

[48] Ye F. F., Yang L. H., Wang Y. M., 2019. Fuzzy rule based system with feature extraction for environmental governance cost prediction. J*ournal of Intelligent & Fuzzy Systems*, 37, 2337-2349.

[49] Ye F. F., Yang L. H., Wang Y. M., 2020, An interval efficiency evaluation model for air pollution management based on indicators integration and different perspectives, *Journal of Cleaner Production*, 245: 1-15.

[50] Zhang W., Li G. X., Uddin M. K., Guo S. C., 2020. Environmental Regulation, Foreign Investment Behavior, and Carbon Emissions for 30 provinces in China. *Journal of Cleaner Production*, 248, 1-11.

[51] Zheng S., He C., Hsu S. C., Sarkis J., Chen J. H., 2020, Corporate environmental performance prediction in China: An empirical study of energy service companies. *Journal of Cleaner Production*, 266: 1-16.