A Method for Prediction of Waterlogging Economic Losses in a Subway Station Project

Han Wu and Junwu Wang *

School of Civil Engineering and Architecture, Wuhan University of Technology, Wuhan 430070, China; wu.han@whut.edu.cn
* Correspondence: junwuwang@whut.edu.cn; Tel.: +86-189-7165-0318

Abstract: In order to effectively solve the problems of low prediction accuracy and calculation efficiency of existing methods for estimating economic loss in a subway station engineering project due to rainstorm flooding, a new intelligent prediction model is developed using the sparrow search algorithm (SSA), the least-squares support vector machine (LSSVM) and the mean impact value (MIV) method. First, in this study, 11 input variables are determined from the disaster loss rate and asset value, and a complete method is provided for acquiring and processing data of all variables. Then, the SSA method, with strong optimization ability, fast convergence and few parameters, is used to optimize the kernel function and the penalty factor parameters of the LSSVM. Finally, the MIV is used to identify the important input variables, so as to reduce the predicted input variables and achieve higher calculation accuracy. In addition, 45 station projects in China were selected for empirical analysis. The empirical results revealed that the linear correlation between the 11 input variables and output variables was weak, which demonstrated the necessity of adopting nonlinear analysis methods such as the LSSVM. Compared with other forecasting methods, such as the multiple regression analysis, the backpropagation neural network (BPNN), the BPNN optimized by the particle swarm optimization, the BPNN optimized by the SSA, the LSSVM, the LSSVM optimized by the genetic algorithm, the PSO-LSSVM and the LSSVM optimized by the Grey Wolf Optimizer, the model proposed in this paper had higher accuracy and stability and was effectively used for forecasting economic loss in subway station engineering projects due to rainstorms.

Keywords: subway station project; waterlogging economic losses; sparrow search algorithm; least-squares support vector machine; mean impact value

1. Introduction

In recent years, extreme weather events have frequently occurred in the world, and natural disasters, such as waterlogging, have posed a great threat to public safety [1]. A subway station project is more vulnerable to waterlogging due to its low terrain elevation. Accurate and rapid prediction of flood disaster losses in a subway station project not only is helpful for their quick recovery, but also is of utmost importance to ensure smooth implementation and sustainable development of the project.

The losses caused by flood disasters are mainly divided into two categories, namely economic losses and life losses [2]. In the practice of engineering, it is easy for project managers to count the loss of life within the construction scope of a subway station project after the flood disaster. However, the calculation and statistics of economic losses lag behind and can only be obtained after the postdisaster recovery period. Therefore, this study examined the economic losses caused by flood disasters.

After sudden flood disasters, the economic losses of subway stations and other construction projects are extremely complicated. According to the causes of disaster losses, disaster losses are related not only to the characteristics of flood disasters, but also to the engineering and management characteristics of subway station projects [3]. From the
perspective of accounting and finance, disaster losses are related to the disaster loss rate and the value of assets. As a result, there are many influencing factors that affect the disaster losses caused by storm waterlogging in the subway station projects, and there is a nonlinear mapping relationship between these influencing factors and disaster losses. Currently, the valuation methods based on bill of quantities and index estimation methods are commonly used to calculate disaster losses in an engineering project. However, these two methods only consider the influence of a small number of influencing factors on disaster losses, and thus they have some shortcomings, such as low calculation accuracy and efficiency [4]. Principal component analysis, multiple linear regression and factor analysis have disadvantages of poor interpretability and applicability of screening results.

The least-squares support vector machine (LSSVM) not only can solve the problems of few samples, nonlinearity and many input variables, but also has the advantages of simple operation, rapid convergence and high prediction accuracy, which is more suitable for the prediction of recovery cost after storm waterlogging in open-cut subway station engineering. The sparrow search algorithm (SSA) has the advantages of good stability, strong global search ability and few parameters, and it can be used to optimize two key parameters of LSSVM. Using the mean impact value (MIV) method to screen input variables can not only improve the calculation accuracy of a prediction model, but also have the advantage of strong interpretability.

In order to effectively solve the problems of low prediction accuracy and calculation efficiency of existing methods for estimating economic loss in a subway station engineering project due to rainstorm flooding, a new intelligent prediction model is developed using the SSA, the LSSVM and the MIV method.

The sections of this paper are arranged as follows: In Section 2, the related work is summarized. In Section 3, the system of input variables and the forecasting model using the MIV, SSA and LSSVM methods are studied in detail. In Section 4, the prediction model is applied to 45 stations of lines 11 and 27 of the Chengdu Metro system and line 8 of the Wuhan Metro system. Additionally, the nonlinear relationship among the variables is also analyzed in this section to highlight the necessity of using nonlinear modeling methods. Section 5 compares the accuracy and stability of the calculation results of different prediction models and compares the calculation results of different key index extraction methods. Section 6 summarizes the full text and gives directions worthy of further study.

2. Related Work

At present, some research achievements have been made in the prediction of disaster losses, but they have often been made with large-scale areas as research objects, and there are almost no research achievements in typical disaster-bearing bodies in small-scale areas. For example, Zhao (2014) used the triangular grey relational theory to study the risk level of flood and disaster loss in China [5]. Guo (2020) constructed an economic evaluation system for urban rainstorm disasters [6]. Tan (2019) used the input–output model to predict the rainstorm economic loss in Beijing, China [7]. Using the region of Zhu Jiang River in China as a research object, Kang (2015) constructed a linear calculation model between natural disaster loss in cultivated land and cultivated land area, submerged height and disaster loss rate [8]. Yang (2019) pointed out that research on disaster management in recent years has been gradually shifting from the large-scale areas to the small-scale areas, as disaster management in small-scale areas was found to have better practical guiding significance [9]. To the best of our knowledge, there is no published research on forecasting waterlogging economic loss in a subway station project.

Many forecasting methods have been proposed to solve the problem of natural disaster loss prediction, but the theoretical basis of most of them is to assume that there is a simple linear relationship between the influencing factors and natural disaster losses. A prediction model of natural disaster losses based on a linear regression method was developed [10] by analyzing the relationship between the gross domestic product (GDP), the population,
the land area and natural disaster losses. Cao (2020) used the improved grey model (1, 1) to predict the direct economic losses caused by marine disaster damage to coastal cities in China [11]. Since nonlinear data samples are difficult to handle with the grey model (1, 1), this study still holds that there is a linear relationship between input and output variables, which are actual marine disaster losses. Yin (2017) used the dispersion of disaster data and the grey relational analysis (GRA) to construct a prediction model of storm tide disaster loss [12]. The basic idea of the GRA is to compare whether the trends of two factors are consistent, but the complex nonlinear relationship between input and output variables is ignored. Due to the inability to reveal the complex nonlinear relationship between the influencing factors and natural disaster losses, it is difficult to achieve accurate prediction of natural disaster losses with these methods.

With the development of artificial intelligence, machine learning algorithms represented by the backpropagation neural network (BPNN) have been applied in the engineering field [13]. He (2014) used a dynamic recurrent neural network to predict regional flood disaster losses [14]. Wang (2018) constructed a BPNN optimized by the beetle antennae search (BAS) and applied it to the preassessment of storm surge economic losses [15]. Zhao (2019) used a BPNN optimized by the adaptive lifting algorithm to predict the direct economic losses due to marine disasters [16]. However, the BPNN has some shortcomings, such as slow convergence speed and poor generalization ability. Even if the metaheuristic optimization algorithms are used to optimize the initial weights and thresholds of the BPNN, its calculation is still complicated and its operation efficiency is low. The support vector machine (SVM) has strong local and global optimization ability and can solve the problems of few samples, nonlinearity and numerous input variables [17]. Based on the traditional SVM, the least-squares support vector machine (LSSVM) combines the regularization theory to transform the inequality constraints into equality constraints, thereby avoiding the complex quadratic programming problem of the SVM. As a result, the LSSVM not only has the strong local and global optimization ability of the SVM, but also has better calculation accuracy and prediction speed than the SVM [18].

Based on the above analysis, this study intended to introduce the LSSVM, which has the advantages of simple operation, rapid convergence and high prediction accuracy, for the prediction of waterlogging economic losses in a subway station project. However, the kernel function parameter and the penalty factor parameter of the LSSVM directly affect its prediction accuracy [17]. At present, the genetic algorithm (GA) [18], the particle swarm optimization (PSO) [19], the grey wolf optimizer (GWO) [20] and other metaheuristic algorithms are often used to optimize two key parameters of the LSSVM. However, the GA has some limitations, such as complex programming, complex parameter setting and slow convergence speed. The PSO and GWO algorithms often converge to a local optimal solution. Recently, Xue and Shen (2020) proposed the sparrow search algorithm (SSA) inspired by the biological phenomenon of sparrow population foraging [21]. The SSA can achieve higher global exploration ability and local development ability with the help of flexible foraging and antipredation mechanism in a sparrow population.

As the SSA is brand new, there are few published studies on the engineering application or theoretical analysis of the SSA. Recently, Xue and Shen (2020) analyzed the computational performance of the SSA in detail [21]. Compared with other classical metaheuristic optimization algorithms, the SSA was found to have the advantages of good stability, strong global search ability and few parameters. In order to solve the problem of insufficient image clarity caused by the traditional image segmentation method, Lv (2020) used the SSA to improve the retrieval accuracy and retrieval range and effectively improved the image clarity after segmentation [22]. In order to effectively solve the problems of the large amount of computation and difficulty in the convergence of the unmanned aerial vehicle (UAV) path planning, Tang (2020) used the improved SSA to optimize the UAV path planning, which was a typical high-dimensional function optimization problem [23]. Benchmark tests and case studies showed that the improved SSA outperformed the PSO, the BAS, the whale optimization algorithm (WOA) and the GWO. To the best of our knowl-
edge, there is no published research on the LSSVM optimized by the SSA. Besides, the computational performance of the same optimization algorithm in different optimization problems may be quite different [24]. Therefore, this study used the SSA to optimize the LSSVM. In addition, it assessed the optimization performance of the SSA in the LSSVM. Furthermore, it undertook to effectively optimize two key parameters of the LSSVM with good stability, strong global search ability and few parameters of the SSA.

When an artificial intelligence algorithm is applied to engineering, it is difficult to determine the input variables of the neural network due to the lack of a clear theoretical basis. When some unimportant independent variables are introduced into an artificial intelligence algorithm, it will not only reduce the model accuracy, but also increase the model operation time [25]. Therefore, the selection of key input variables also determines the prediction accuracy of the LSSVM. Currently, principal component analysis (PCA) [26], the multiple linear regression method (MLRM) [27] and the factor analysis method (FAM) [28] are often used to screen for the key input variables based on the correlation between variables. However, the interpretability of the variable screening results of these methods is poor. Since their basic idea was to analyze the data relationship among multiple input variables, the analysis results can explain well the linear or nonlinear relationship among the variables but cannot clearly determine the influence of the linear or nonlinear relationship between the input variables on the prediction results.

In recent years, the mean impact value (MIV) has been widely used in the screening of input variables. Through sensitivity analysis of input and output variables, the MIV can retain the input variables that have a significant influence on output variables [29]. Recently, Yan (2019) effectively reduced the input variables in laser-induced breakdown spectroscopy (LIBS) [29]. The case study showed that the calculation time and performance of the model were significantly improved after adopting the MIV method. In order to quickly and accurately predict the air passenger flow, Sun (2019) successfully extracted the key input variables of a nonlinear vector autoregression neural network (NVARNN) using the MIV [30]. Considering many factors that affect lithology prediction, Gu (2020) used the MIV to screen indicators to reduce the number of input variables of the PSO-BPNN, and the experiments showed that this method achieved better computational performance than other methods [31]. In addition, according to the calculation principle of the MIV, the input variables with a greater influence on the prediction results are selected as key variables, so the screening results are well interpretable.

Based on the above analysis, in this study, the MIV, SSA and LSSVM methods are used to construct the forecasting model of economic loss in a subway station engineering project due to flooding, and a case analysis is performed.

3. Materials and Methods
3.1. The System of Input Variables
3.1.1. The Analysis of Input Variables

The study of economic losses due to waterlogging disasters in subway station projects has typical interdisciplinary characteristics. From the financial accounting perspective, the economic losses due to waterlogging disaster in a subway station project are related to the disaster loss rate and the asset value. From the disaster science perspective, the economic losses due to disaster are related to the disaster-causing factors, the disaster-bearing environment, the vulnerability of disaster-bearing bodies and the ability of disaster prevention and mitigation. Considering the interpretability of loss calculation results, this study analyzed the factors related to the disaster-causing factors, the disaster-bearing environment, the vulnerability and the ability of prevention and mitigation as input variables from the disaster loss rate and original asset value. Except for some indicators related to disaster-bearing environment and vulnerability, all other indicators are input variables related to disaster loss rate.

The greater the disaster loss rate, the greater the disaster loss caused by waterlogging. (1) Rainfall is an important disaster-causing factor of waterlogging disaster, which can
effectively represent the waterlogging pressure caused by short-term heavy rainfall. The more intense the rainfall, the greater the loss rate of the waterlogging disaster. (2) The catchment area is the catchment area along the ridge line of the high slope around the subway station project, which is an important disaster-causing factor of waterlogging disaster. The larger the catchment area of a low-lying subway station project, the more prone it is to waterlogging disaster, and the greater the loss rate of loss disaster. (3) When there are geomorphic elements, such as ponds, high slopes, irrigation canals and rivers, around the subway station under construction, the possibility of waterlogging disasters increases sharply [3]. (4) The height of the retaining wall directly determines whether waterlogging water enters the subway station project. The higher the retaining wall is, the less likely the accumulated water is to flow backward into the subway station under construction, and the smaller the loss rate of the waterlogging disaster. (5) The management ability of the participating units is related to many factors, such as construction production recovery plan, construction production condition recovery ability, construction production order recovery ability and postdisaster psychological intervention plan. It is generally believed that the better the management ability of participating units, the smaller the loss rate of the waterlogging disaster in the subway station engineering. (6) When a disaster occurs, the participating units often implement emergency plans to reduce disaster losses. Generally speaking, the more science-based the preparation, decision-making and implementation of the emergency plan, the more effective it is in reducing the waterlogging disaster, and the smaller the disaster loss rate is [32].

The greater the value of the original assets, the greater the loss caused by waterlogging. (1) The greater the total investment in the subway station project, the greater the value of assets within the construction scope when the waterlogging disaster occurs, and the more likely the loss of the waterlogging disaster is. (2) The construction contents are different in different construction stages, and the statuses of workers, materials and machines on the site are also quite different. The running statuses, self-toughness and sensitivity of subway projects under construction are quite different. (3) In a waterlogging disaster, it is difficult to transfer large machinery on the construction site, so large machinery is more vulnerable to disaster losses. Moreover, the value of large-scale machinery is high, and the disaster losses are often great. (4) The greater accumulation of raw materials and semifinished products in the construction site, the greater the loss caused by waterlogging. (5) When the waterlogging disaster occurs in a subway station engineering project, the flood may enter shield engineering and cause more disaster losses.

3.1.2. The Methods of Selection and Quantification of Input Variables

According to the research results described in Section 3.1.1, on the basis of following the principles of scientificity, representativeness, independence, availability and operability, 11 variables, as shown in Table 1, were selected as the input variables for the forecast of economic losses due to waterlogging in a subway station project.

Maximum rainfall in 24 h (F1) can be calculated by referring to local meteorological data and determining the data of this index according to the actual rainfall in the area where the station project is located. The unit is mm/24 h.

Catchment area (F2) can be calculated according to the planning and design drawings or measured on-site. The unit is km².

Unfavorable geomorphological features (F3) is an index obtained by the expert scoring method, considering the current lack of research on the formation mechanism of waterlogging due to environmental characteristics and geomorphological characteristics around the project. The scoring basis is shown in Table 2.

Height of retaining wall (F4) can be obtained through design data and field investigation, and the unit is m.

Management ability of participating units (F5) is a qualitative index, which comprehensively considers the disaster management capabilities of the owner, construction unit and
design unit. In this study, the expert scoring method is used to obtain the data of this variable, and the scoring basis is shown in Table 3.

Table 1. System of input variables.

| Input Variable                                     | No.  | Unit          | Data Acquisition Method                                      |
|---------------------------------------------------|------|---------------|--------------------------------------------------------------|
| Maximum rainfall in 24 h                          | F1   | mm/24 h       | Access to local meteorological data                          |
| Catchment area                                     | F2   | km\(^2\)      | Refer to planning and design documents or on-site investigation |
| Unfavorable geomorphological features              | F3   | -             | Questionnaire survey                                         |
| Height of retaining wall                          | F4   | M             | Refer to planning and design documents or on-site investigation |
| Management ability of participating units          | F5   | -             | Questionnaire survey                                         |
| Implementation of emergency plan                   | F6   | -             | Questionnaire survey                                         |
| Total project investment                           | F7   | Million       | Access to design information or government notices           |
| Proportion of accumulated completed output value   | F8   | %             | On-site investigation or consulting the construction log      |
| Quantity of large machinery on construction site   | F9   | -             | Site survey                                                 |
| Value of raw materials and semifinished products stacked on the construction site | F10  | Million       | On-site investigation or consulting the construction log      |
| Number of associated shield sections               | F11  | -             | Refer to planning and design documents or on-site investigation |

Table 2. Score basis of unfavorable geomorphological features.

| Score     | Explanation                                                                 |
|-----------|-----------------------------------------------------------------------------|
| [0, 20]   | There is no adverse surrounding environment, such as ponds, high slopes, irrigation canals or rivers. |
| [20, 40]  | There are 1–2 kinds of adverse surrounding environments, but the distance is far away.          |
| [40, 60]  | There are 1–2 kinds of adverse surrounding environments, and the distance is close.          |
| [60, 80]  | There are many kinds of adverse surrounding environments, but they have not caused waterlogging disasters in the region. |
| [80, 100] | There are many kinds of adverse surrounding environments, which have caused waterlogging in the region. |

Table 3. Score basis of the management ability of participating units.

| Score     | Explanation                                                                 |
|-----------|-----------------------------------------------------------------------------|
| [0, 20]   | There is a scientific and feasible construction order recovery plan, professional construction condition recovery ability, good construction order recovery ability and good crisis learning ability. |
| [20, 40]  | There is a good construction order recovery plan, good construction condition recovery ability, good construction order recovery ability and good crisis learning ability. |
| [40, 60]  | The construction order restoration plan is essentially complete, basically has the ability to restore construction conditions, basically has the ability to restore construction order, and basically has the ability to learn crisis. |
| [60, 80]  | The construction order restoration plan is incomplete, and has the ability to restore construction conditions, construction order and crisis-free learning, or is weak. |
| [80, 100] | There is no construction order recovery plan, weak construction condition recovery ability, poor construction order recovery ability and crisis-free learning ability. |

Implementation of emergency plan (F6), is a qualitative indicator due to the complexity of the implementation of the emergency plan. In this study, the expert scoring method is used to obtain the data for this variable, and the scoring basis is shown in Table 4.
Table 4. Score basis for the implementation of emergency plan.

| Score   | Explanation                                                                                                                                 |
|---------|---------------------------------------------------------------------------------------------------------------------------------------------|
| [0, 20) | Emergency plans are well prepared, with special emergency plans related to flood control and waterlogging prevention. In the emergency rescue, the participating units effectively and flexibly implement the emergency plan, which effectively reduces the disaster losses caused by waterlogging. |
| [20, 40) | The emergency plan is well prepared, and there are special emergency plans related to flood control and waterlogging prevention. During the emergency rescue process, the participating construction units implement the emergency plan, which reduces the disaster losses caused by waterlogging. |
| [40, 60) | The emergency plan is basically complete, and there are emergency plans related to flood control and waterlogging prevention. During the emergency rescue process, the participating construction units implement the basic emergency plan, which initially reduces the disaster losses caused by waterlogging. |
| [60, 80) | The preparation of the emergency plan is incomplete, and there is no emergency plan related to flood control and waterlogging prevention. In the emergency rescue process, the participating units basically do not implement the emergency plan, and basically do not reduce the disaster losses caused by waterlogging. |
| [80, 100) | No emergency plan is prepared. In the emergency rescue process, the participating units have no emergency plans that can be executed. |

*Total project investment* (F7) is a quantitative input variable that can be obtained by consulting the design data or government announcement, and the unit is million of CNY (Chinese Yuan).

*Proportion of accumulated output value* (F8) is a quantitative input variable that can be obtained by field investigation or consulting the construction log, and it is expressed in percentage.

*Number of large machines in the construction site* (F9) is a quantitative input variable and can be obtained through field investigation. In this study, a piece of mechanical equipment with a value of more than RMB 500,000 or USD 100,000 is considered as a large-scale machine, such as a gantry crane, bridge erecting machine, tower crane, crawler crane or construction elevator. This variable has no units.

*Value of raw materials and semifinished products stacked on the construction site* (F10) is a quantitative input variable that can be obtained by field investigation or consulting the construction log, and the unit is million.

*The number of associated shield sections* (F11) is a quantitative input variable and can be obtained from the design data, from the construction organization design and through field investigation, and it has no units. At present, in engineering practice, the number of connected shield sections is often 0, 1 or 2.

3.1.3. Data Processing Methods for the Input and Output Variables

Before forecasting, we need to deal with the data in two ways: (1) eliminate the influence of the time value of funds and convert the funds occurring at different time points to the same time point; (2) normalize the input variables to avoid the LSSVM falling prematurely into a local minimum.

(1) Eliminating the influence of the time value of funds

Total project investment (F7), value of raw materials and semifinished products piled up in the construction site (F10) and disaster loss (output variable) are all variables related to the time value of funds. In this study, the cost index method is used to deal with the data of these three variables, and the calculation equation is as follows:

\[ v_1^* = \frac{B_2}{B_1} v_i , \]  

where \( v_1^* \) is the capital value adjusted to the base period, \( v_i \) is the capital value before adjustment, \( B_1 \) is the cost index of the base period and \( B_2 \) is the cost index of the reporting period. \( B_1 \) and \( B_2 \) can refer to the data issued by the cost association.

(2) Normalizing the input and output variables
In order to avoid the LSSVM model falling prematurely into a local minimum as a result of the large difference between the absolute values of the input and output data, it is necessary to normalize in advance the input and output vectors of the training data sample set [19].

In this study, the input and output data are transformed by linear normalization, and the transformation equation is as follows [20]:

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}},$$  (2)

where $x_i^*$ represents the normalized data, $x_i$ represents the collected data, $x_{\min}$ is the minimum value and $x_{\max}$ is the maximum value.

3.2. Prediction Model of Waterlogging Economic Loss in Subway Station Project

3.2.1. Introduction to the LSSVM

The LSSVM is a new machine learning algorithm that extends and improves the standard support vector machine (SVM). It replaces the previous inequality constraint problem with equality constraint, simplifies the Lagrange multiplier solving process and greatly reduces the computational complexity of the SVM [17].

With training samples $\{x_i, y_i\}_{i=1}^l$, $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$, where $x_i$ and $y_i$ represent the input vector and output vector of LSSVM, respectively, the following prediction function can be developed [19]:

$$f(x) = w^T \varphi(x) + b,$$  (3)

where $w^T$ represents the function weight variable, $\varphi(x)$ is the nonlinear mapping function that maps the input samples to the high-dimensional space and $b$ represents the deviation coefficient.

Combining Equation (3) with a structural risk function, it can be transformed into the following optimization problem [19]:

$$\begin{align*}
\min_{w, b, \delta} J(w, \delta) &= \frac{1}{2} \|w^2\| + \frac{C}{2} \sum_{i=1}^l \delta_i^2 \\
\text{s.t. } y_i = w^T \varphi(x) + b + \delta_i^2, i = 1, 2, \ldots, l
\end{align*},$$  (4)

where $C$ represents the penalty factor parameter and $C > 0$, $\delta_i$ represents the relaxation variable.

Equation (5) can be obtained by introducing the Lagrange multiplier $\alpha_i$ into Equation (4):

$$L(w, b, \delta, \alpha) = 1/2 \|w^2\| + C/2 \sum_{i=1}^l \delta_i^2 - 1/2 \sum_{i=1}^l \alpha_i^2 \left( w^T \varphi(x) + b + \delta_i - y_i \right)$$  (5)

According to the Karush–Kuhn–Tucker (KKT) optimization conditions, the partial derivatives of $w$, $b$, $\delta$ and $\alpha$ in the above Equation (5) can be obtained [20]:

$$\begin{align*}
    w &= \sum_{i=1}^l \alpha_i \varphi(x) \\
    \sum_{i=1}^l \alpha_i &= 0 \\
    \alpha_i &= C \delta_i \\
    w^T \varphi(x) + b + \delta_i &= y_i
\end{align*},$$  (6)

The two variables $w$ and $\delta$ in Equation (6) are eliminated, and the nonlinear equation solving problem is transformed into a linear solving problem [20]:

$$\begin{bmatrix}
0 \\
y
\end{bmatrix} = \begin{bmatrix}
0 & I_l^T \\
I_l & K + C^{-1} I_l
\end{bmatrix} \begin{bmatrix}
b \\
\alpha
\end{bmatrix}$$  (7)

where $\alpha$ is the Lagrange multiplier $I_l = [1, 2, \ldots, l]^T$ and $K$ is the kernel function.
Then, the final optimization function of the LSSVM is [19,20]

\[ y(x) = \sum_{i=1}^{l} \alpha_i K(x, x_i) + b \]  

(8)

where \( K(x, x_i) \) is the kernel function.

The essence of a kernel function is to realize a feature transformation through the inner product of sample data set in the feature space. Introducing a kernel function can effectively and conveniently solve the linear indivisibility problem. At present, there are four commonly used kernel functions, the linear kernel function, the polynomial kernel function, the radial basis function (RBF) kernel function and the perceptron kernel function [17–19]. Among them, the most widely used RBF kernel function has the following three advantages: (1) The representation of RBF kernel function is simple, and it does not increase complexity too much even for multivariate input. (2) The RBF kernel function has radial symmetry and good smoothness, and it can be used with derivatives of any order. (3) Because the RBF kernel function is simple and analytical, it is convenient for theoretical analysis.

Based on these advantages of RBF kernel function, this paper adopts RBF kernel function as the kernel function in the LSSVM model, as shown in Equation (9) [19,20]:

\[ K(x, x_j) = \exp \left( \frac{\| x - x_j \|^2}{2\sigma^2} \right) \]  

(9)

where \( \sigma \) represents the width parameter of the RBF.

3.2.2. Introduction to the SSA Method

The SSA is a new swarm intelligence optimization algorithm, which was put forward in 2020 and is mainly inspired by the foraging behavior and antipredation behavior of sparrows.

When using virtual sparrows to search for food, the population composed of \( n \) sparrows can be expressed as follows [21]:

\[ X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix} \]  

(10)

where \( d \) denotes the dimension of the problem variable to be optimized and \( n \) is the number of sparrows.

The fitness values of all sparrows can be expressed as follows [21]:

\[ F_x = \begin{bmatrix} f(x_{11}, x_{12}, \cdots, x_{1d}) \\ f(x_{21}, x_{22}, \cdots, x_{2d}) \\ \vdots \\ f(x_{n1}, x_{n2}, \cdots, x_{nd}) \end{bmatrix} \]  

(11)

where \( f \) denotes the fitness value.

In the SSA, the discoverer with better fitness value will find food first in the search process. In addition, the discoverer is responsible for finding food for the whole sparrow population and providing foraging directions for all participants. Therefore, the discoverer can cover a larger foraging search range than the entrant. During each iteration, the location update of the discoverer is described as follows [21]:

\[ x_{i,j}^{t+1} = \begin{cases} x_{i,j}^t \cdot \exp \left( \frac{-i}{\text{iter}_{\text{max}}} \right) & \text{if } R_x < ST \\ x_{i,j}^t + Q \cdot L & \text{if } R_x \geq ST \end{cases} \]  

(12)
where \( t \) denotes the current iteration number; \( \text{iter}_{\text{max}} \) is a constant, indicating the maximum iteration number; \( X_{i,j} \) represents the position information of the \( i \)-th sparrow in the \( j \)-th dimension; \( \alpha \) is a random number greater than 0 and less than or equal to 1; \( R_2 \) \((R_2 \in [0, 1])\) and \( ST \) \((ST \in [0, 1])\) represent the warning value and the safety value, respectively; \( Q \) is a random number that obeys normal distribution; and \( L \) represents a \( 1 \times d \) matrix, in which all elements in the matrix are 1.

When \( R_2 < ST \), it means that there are no predators around the foraging environment, and the discoverer can perform extensive search operations. When \( R_2 \geq ST \), it means that some sparrows in the population have found predators and sent an alarm to other sparrows in the population. At this time, all sparrows need to fly quickly to other safe places for foraging.

In the SSA, the position update of the enrollee (follower) is described as follows [21]:

\[
X^{t+1}_{i,j} = \begin{cases} 
Q \cdot \exp \left( \frac{X^{t}_{\text{worst}} - X^{t}_{i,j}}{\nu} \right) & \text{if } i > n/2 \\
X^{t}_{\hat{P}} + \left| X^{t}_{i,j} - X^{t}_{\hat{P}} \right| \cdot A^+ \cdot L & \text{otherwise}
\end{cases}
\]

(13)

where \( X_{\hat{P}} \) is the best position occupied by the discoverer at present; \( X_{\text{worst}} \) is the worst position in the world at present; and \( A \) represents a \( 1 \times d \) matrix, in which each element is randomly assigned 1 or \(-1\), and \( A^+ = A^T (AA^T)^{-1} \).

When \( i > n/2 \), it indicates that the \( i \)-th participant with low fitness value has not obtained food and is in a very hungry state. At this time, it must fly to other places to forage for more energy.

When aware of the danger, the sparrow population will display antipredation behavior, and its mathematical expression is as follows [21]:

\[
X^{t+1}_{i,j} = \begin{cases} 
X^{t}_{\text{best}} + \beta \left| X^{t}_{i,j} - X^{t}_{\text{best}} \right| & \text{if } f_i > f_g \\
X^{t}_{i,j} + K \frac{x^{t}_{i,j} - x^{t}_{\text{best}}}{f_i - f_w} & \text{if } f_i = f_g
\end{cases}
\]

(14)

where \( X^{t}_{\text{best}} \) is the current global optimal position; \( \beta \), as a step control parameter, is a random number with normal distribution with mean value of 0 and variance of 1; \( K \in [-1, 1] \) is a random number; \( f_i \) is the fitness value of current sparrows; the \( f_g \) and \( f_w \) are the best and worst fitness values in the world, so there will be no case where \( f_i < f_g \); and \( \epsilon \) is the smallest constant to avoid zero denominator.

For simplicity, \( f_i > f_g \) indicates that sparrows are at the edge of the population at this time and are extremely vulnerable to predators; \( f_i = f_g \) implies that the sparrows in the middle of the population are aware of the danger and need to approach other sparrows to minimize the risk of predation; \( K \) represents the moving direction of the sparrows and is also a step control parameter.

### 3.2.3. Introduction to the MIV Algorithm

The MIV is one of the best algorithms to evaluate the relevance of each index in artificial intelligence algorithms.

It is assumed that the initially selected input variable set is denoted as \( X(t) = \{x_1(t), x_2(t), \cdots, x_n(t)\} \), which contains \( n \) input variables, with \( t \) representing the \( t \)-th sample and \( t = 1, 2, \cdots, m \). The sequence of output variables is \( Y = \{y(1), y(2), \cdots, y(m)\} \), and \( m \) is the number of samples.

The specific implementation steps of the MIV algorithm are as follows:

- **Step 1. Establishing a preliminary model.** With \( X(t) \) as the input and \( Y \) as the output, the data set is standardized and preprocessed, and the simulation model is preliminarily constructed. The model adopts the SSA-LSSVM method, which makes the prediction effect reach a satisfactory state.
Step 2. Constructing new training samples. The value of the \( i \)-th input variable in \( X(t) \) is correspondingly increased by 10% to form a new training sample \( P_i = \{ P_i(1), P_i(2), \cdots, P_i(m) \} \), where \( P_i(t) \) represents the value of the \( i \)-th factor in the original input variable set increased by 10%. Similarly, the value of each input variable in \( X(t) \) is reduced by 10%, and a new training sample \( Q_i = \{ Q_i(1), Q_i(2), \cdots, Q_i(m) \} \) is formed, where \( Q_i(t) \) represents that the value of the \( i \)-th factor in the original input variable set is reduced by 10%.

Step 3. Using the standardization rules of the original sample data, the newly constructed training sample set \( \{ P_1; P_2; \cdots; P_n \} \) is standardized, and the newly constructed data sets are simulated by simulation models, and the simulation results are \( M = \{ M_1; M_2; \cdots; M_n \} \) and \( N = \{ N_1; N_2; \cdots; N_n \} \). \( M_i \) and \( N_i \) are the prediction results of newly constructed training sample sets \( P_i \) and \( Q_i \), respectively.

Step 4. Calculating the average influence change value (MIV). The equation for calculating the average influence value of the \( i \)-th input variable is as follows:

\[
MIV_i = \frac{1}{m} \sum_{t=1}^{m} (M_i(t) - N_i(t)).
\]

Step 5. Determining the influence degree of characteristic factors on output variables. Calculate the relative contribution rate of each variable \( \alpha_i \) [29]:

\[
\alpha_i = \frac{|MIV_i|}{\sum_{i=1}^{n} |MIV_i|}.
\]

Input variables with relative contribution rate greater than 10% are selected as key input variables [29–31].

3.2.4. The Prediction Method of Economic Losses Due to Waterlogging in a Subway Station Project

The flow chart of the economic loss prediction model developed in this study is shown in Figure 1. The forecasting model proposed in this paper is mainly divided into three parts: (1) putting all input variables into the prediction model based on the SSA-LSSVEM method, (2) using the MIV method to obtain key input variables and (3) putting the key variables into SSA-LSSVEM algorithm to repredict the economic losses due to rainstorm.

In this study, LSSVM is used to build the prediction model, SSA is used to find the optimal parameters of LSSVM and MIV is used to find the optimal input variable system of the prediction model. The calculation steps of the prediction model proposed in this paper are as follows:

Step 1. Obtain data according to the procedure described in Table 1, and build training set and test set. The input data and output data are preprocessed using Equations (1) and (2).

Step 2. Determine the learning parameters of the SSA and LSSVM algorithms according to the size and characteristics of the learning samples. Use Equation (10) to initialize the sparrow population. Use Equations (3)–(9) to initialize the LSSVM model.

Step 3: Train the parameters of the LSSVM model with the combination of each generation \( \mathcal{C} \) and \( \sigma \), and calculate the fitness values of all sparrows using the fitness function. In this study, the root-mean-square error (RMSE) is selected as a function to evaluate particle fitness [33]:

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2},
\]

where \( n \) is the number of samples in the test set, \( y_i \) is the predicted value and \( \hat{y}_i \) is the true value.

The fitness value of each virtual sparrow is calculated by Equation (17), and the fitness values of each virtual sparrow are sorted. Update the discoverer’s position of the discoverer
with Equation (12), the joiner’s position of the joiner with Equation (13) and the watcher’s position of the watcher with Equation (14).

Step 4: Update the fitness value and position of the sparrow.
Step 5: Determine whether the stop condition is met; if yes, exit and output the result. Otherwise, repeat Steps 3–4.

Figure 1. Flow chart of the prediction model proposed in this paper.
Step 6. Using the MIV method described in Section 3.2.3, calculate the relative contribution rate of each variable, and select the input variable whose relative contribution rate is greater than 10% as the key input variable.

Step 7: Re-execute Steps 2–5, and bring the obtained key input variables into SSA-LSSVM model to obtain the final output result.

4. Case Study

4.1. Engineering Background and Data Sources

The Chengdu Metro line 11 is 22 km long and passes through Chengdu High-tech Zone, Tianfu New District and Shuangliu District. There are 16 projects for newly built stations in this line. In this study, the economic losses of all station projects due to the waterlogging disaster that occurred 2–5 July 2018 are selected as a case study. The Wuhan Metro line 8 has a total length of 17.6 km, passing through Wuchang District and Hongshan District of Wuhan City. There are 12 station projects in this line. This study takes the economic losses of all 12 station projects due to the waterlogging disaster that occurred 20–25 June 2019 as a case study. The Chengdu Metro line 27 is 24.9 km long and passes through Xindu District, Jinniu District, Chenghua District and Qingyang District of Chengdu. There are 23 subway stations in this project, 17 of which are underground stations. In this study, the economic losses of these 17 underground stations due to the rainstorm disaster that occurred 15–17 August 2020 are selected as a case study. Therefore, in this study, economic loss data collected from 45 subway station engineering projects during a waterlogging disaster were used for case analysis.

Using the index acquisition method of each input variable described in Table 1, the engineering data of the 45 subway station projects were obtained as shown in Table 5. In order to facilitate the subsequent preprocessing calculation, the last two columns in Table 5 are the maximum value and minimum value of the input variable.

| No. | F1  | F2  | F3  | F4  | F5  | F6  | F7  | F8  | F9  | F10 | F11 | Y   |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1   | 159 | 2.43| 17  | 0.6 | 20  | 11  | 135.6| 45  | 5   | 3.56| 2   | 0.05|
| 2   | 159 | 0.38| 84.5| 0.6 | 69  | 88  | 164.0| 60  | 6   | 2.87| 2   | 0.15|
| 3   | 178 | 4.47| 63  | 0.3 | 33.5| 27.5| 114.9| 35  | 8   | 5.18| 2   | 21.34|
| 4   | 178 | 0.76| 73  | 0.9 | 67  | 30.5| 79.5 | 50  | 5   | 3.49| 2   | 0.52|
| 5   | 178 | 1.59| 88  | 0.6 | 86.5| 12.5| 150.0| 35  | 3   | 2.75| 1   | 0.73|
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 41  | 203 | 2.57| 33.5| 0.9 | 27  | 29.5| 135.5| 80  | 2   | 0.65| 2   | 0.04|
| 42  | 127 | 0.42| 13  | 1.2 | 38  | 20.5| 94.5 | 75  | 7   | 4.57| 1   | 0.09|
| 43  | 127 | 0.73| 60.7| 0.9 | 27  | 46.5| 174.3| 60  | 9   | 8.94| 2   | 0.25|
| 44  | 127 | 1.42| 57  | 1.2 | 27.5| 72  | 130.2| 70  | 3   | 3.50| 2   | 0.54|
| 45  | 127 | 1.75| 67.5| 0.6 | 45.5| 65.5| 158.7| 55  | 7   | 5.62| 2   | 1.54|
| Max | 247 | 3.42| 92  | 1.8 | 88.5| 92.5| 195.5| 85  | 9   | 9.42| 3   | 21.34|
| Min | 127 | 0.38| 10  | 0.3 | 20  | 5.5 | 79.5 | 30  | 1   | 0.65| 0   | 0.04|

The three subway lines in the case study all span multiple areas, so different station projects on the same line have different 24 h maximum rainfall during the same rainstorm. The index scores for F3, F5 and F6 were obtained by questionnaire survey, and the average score of 10 experts was taken as the final index score. In order to ensure the validity of expert scoring data, 10 experts participated in the construction of case projects, all of whom are senior engineers with more than 15 years of working experience. In addition, the Cronbach’s α coefficients were 0.712, 0.794 and 0.803, respectively, which indicated the good reliability of the results of this questionnaire survey [34].

4.2. Correlation Analysis of Various Variables

In order to determine the linear correlation between input variables and output variables, Pearson correlation analysis is used to quantitatively measure the linear correlation
degree of each parameter [35]. The detailed calculation principle and method were as previously reported [35]. A Pearson correlation coefficient ($r$) greater than 0.8 indicates a strong linear correlation between the two indicators, and a Pearson correlation coefficient greater than 0.7 indicates a certain linear correlation between them; otherwise, there is almost no obvious linear relationship between them. When $r$ is positive, the indicators are positively correlated. When $r$ is negative, they are negatively correlated.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}},$$  

(18)

where $x_i$ is the value of an input variable, $\bar{x}$ is the average value of an input variable, $y_i$ is the value of another input variable and $\bar{y}$ is the average value of another input variable.

The calculation results obtained by substituting the data in Table 2 into Equation (17) are shown in Table 6.

Table 6. Pearson correlation coefficient among indexes.

| Factor | F1   | F2   | F3   | F4   | F5   | F6   | F7   | F8   | F9   | F10  | F11  |
|--------|------|------|------|------|------|------|------|------|------|------|------|
| Y      | 0.513| 0.842| 0.311| -0.494| 0.320| 0.227| 0.562| 0.733| 0.228| 0.162| 0.207|

According to the calculation results shown in Table 6, among the 11 input variables, only the correlation coefficient between F2 and Y is greater than 0.8, which indicates that only F2 and Y have a strong positive linear correlation. Only the correlation coefficient between F8 and y is greater than 0.7, which indicates that there is a certain positive linear correlation between them. The absolute values of the correlation coefficients between the other nine input variables and the output variable y are all less than 0.7, which indicates that the linear correlation between these input variables and the output variables is weak. Therefore, when using the input variables constructed in Section 3.1 above to predict the economic loss due to the waterlogging disaster in the subway station engineering project, the nonlinear modeling method should be preferred [29,30].

4.3. Forecast of Economic Loss Due to Waterlogging

In this study, 45 subway station projects selected by the prediction model described in Section 3.2.4 are used for the prediction of economic loss.

Step 1. In the modeling process by the SSA-LSSVM model, data sets can be divided into training and test sets. Among them, the data of the training set are used to train the model, while the data of the test set are used to verify the model. The ratio of the common training set to test set is 90%:10%, 80%:20% or 70%:30% [36]. Considering that there are 45 sets of data in this study, the first 40 samples are used as training sets and the last 5 samples are used as test sets. Specifically, the ratio of the training set to test set in this study was 88.89%:11.11%.

According to the data shown in Table 5, the input and output data are preprocessed by using Equations (1) and (2), and the results are shown in Table 7. It should be noted that this study selected the cost index issued by the cost association where the subway station project is denoted as the cost index in Equation (2).

Step 2. In the SSA, the population size is set to $n = 50$, the maximum number of iterations is 200, the safety threshold $ST = 0.8$, the discoverers account for 20% of the population size and the number of sparrows aware of danger is $SD = 5$. The simulation experiment was conducted in MATLAB 2016a on a 3.40 GHz Intel i7 processor and a computer with 16 GB of memory. In the LSSVM, the regularization parameter $\mathcal{C} \in [0.1, 100]$ and the kernel function width coefficient $\sigma \in [0.01, 1000]$. According to the data shown in Table 7, the sparrow population is initialized by Equation (10). Equations (3)–(9) are used to initialize the LSSVM model.
Table 7. Engineering data after pretreatment.

| No. | F1  | F2  | F3  | F4  | F5  | F6  | F7  | F8  | F9  | F10 | F11 | Y   |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| (1) | 0.267| 0.674| 0.085| 0.200| 0.000| 0.063| 0.484| 0.273| 0.500| 0.332| 0.667| 0.000|
| (2) | 0.267| 0.000| 0.909| 0.200| 0.715| 0.948| 0.728| 0.545| 0.625| 0.253| 0.667| 0.005|
| (3) | 0.425| 1.345| 0.646| 0.000| 0.109| 0.764| 0.437| 0.727| 0.250| 0.325| 0.667| 0.023|
| (4) | 0.425| 0.125| 0.768| 0.400| 0.200| 0.372| 0.690| 0.683| 0.455| 0.750| 0.364| 0.500|
| (5) | 0.425| 0.398| 0.951| 0.200| 0.000| 0.013| 0.037| 0.600| 0.263| 0.172| 0.129| 0.818|

Steps 3–5. The curve of the adaptation function after calculation is shown in Figure 2.

![Figure 2. Curve of the adaptation function.](image)

The calculation results in Figure 2 reveal that the fitness of the SSA rapidly decreases in the initial stage (about 10–30 generations). With the increase in the number of iterations, the fitness eventually converges to about 3.55% in generations 70–90.

The optimization calculation process of the SSA was tracked in detail, as shown in Table 8.

Table 8. Detailed optimization calculation process.

| Iteration (n) | Fitness (n – 1) | Fitness (n) | Fitness (n) – Fitness (n – 1) | Result     |
|--------------|----------------|-------------|--------------------------------|------------|
| 77           | 3.54036%       | 3.54036%    | 0 < 0.00001                    | Continue   |
| 78           | 3.54036%       | 3.54034%    | 0.00002 > 0.00001             | Continue   |
| 79           | 3.54034%       | 3.54034%    | 0 < 0.00001                    | Continue   |
| 200          | 3.54034%       | 3.54034%    | 0 < 0.00001                    | Stop       |

According to the calculation results shown in Table 8, the SSA reached the convergence condition in the 79th generation and converged to 3.54%. Comprehensively, Figure 2 and Table 8 show that the SSA effectively optimizes the LSSVM model. A comparison of the
convergence curves of GA, PSO and GWO revealed that the SSA has better convergence speed. The analysis of the calculation results of different algorithms is discussed in detail in Section 4. The best parameter combination of the LSSVM after optimization is shown in Table 9.

Table 9. Optimal parameter combination of the LSSVM.

| Parameter | +10% F1 | −10% F1 | +10% F2 | −10% F2 | +10% F3 | −10% F3 |
|-----------|---------|---------|---------|---------|---------|---------|
| σ         | 21.583  | 23.412  | 24.172  | 22.473  | 24.327  | 25.323  |
| C         | 8.361   | 7.311   | 6.384   | 7.472   | 6.836   | 5.295   |
| Parameter | +10% F4 | −10% F4 | +10% F5 | −10% F5 | +10% F6 | −10% F6 |
| σ         | 23.481  | 28.384  | 23.429  | 27.384  | 20.672  | 25.482  |
| C         | 8.345   | 5.741   | 7.842   | 8.482   | 6.312   | 4.942   |
| Parameter | +10% F7 | −10% F7 | +10% F8 | −10% F8 | +10% F9 | −10% F9 |
| σ         | 28.391  | 30.136  | 21.468  | 22.950  | 27.643  | 26.495  |
| C         | 6.328   | 9.034   | 7.006   | 8.314   | 7.381   | 6.070   |
| Parameter | +10% F10| −10% F10| +10% F11| −10% F11| Original input system | Filtered input system |
| σ         | 19.371  | 22.341  | 23.470  | 22.642  | 29.473  | 25.423  |
| C         | 5.398   | 11.785  | 8.382   | 5.361   | 8.392   | 5.986   |

The calculation results of the test set are shown in Table 10. By bringing the predicted and preprocessed values calculated by this model into Equations (1) and (2), the usable and preprocessed predicted value of disaster loss can be obtained in reverse, as shown in Table 10.

Table 10. Calculation results of the test set.

| No. | Actual Value | Predicted Value of Original Input System | Input Value of Index System after MIV Processing |
|-----|--------------|------------------------------------------|-----------------------------------------------|
|     | Pretreated | Un-Pretreated | Pretreated | Un-Pretreated | Pretreated | Un-Pretreated |
| (41)| 0.000      | 0.04          | 0.0002     | 0.045         | 0.0001     | 0.043         |
| (42)| 0.002      | 0.09          | 0.0018     | 0.087         | 0.0017     | 0.086         |
| (43)| 0.011      | 0.25          | 0.0107     | 0.242         | 0.0111     | 0.251         |
| (44)| 0.023      | 0.54          | 0.0227     | 0.536         | 0.0229     | 0.539         |
| (45)| 0.069      | 1.54          | 0.0700     | 1.548         | 0.0693     | 1.540         |

The data shown in Table 10 reveal that the predicted value of the original input system is quite close to the actual value. To further analyze the calculation error of the prediction results, this study analyzes the calculation accuracy of the prediction results by using the mean absolute percentage error MAPE commonly used in machine learning research. The average relative error is the average value of relative errors expressed by the MAPE. Thus, the smaller the value of the MAPE, the higher the accuracy of the prediction results. When the value of the MAPE is zero, it means that the prediction result is the same as the real value. The calculation equation of the MAPE is as follows [37]

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{\hat{y}_i},$$

(19)

where $n$ is the number of samples in the test set, $y_i$ is the predicted value and $\hat{y}_i$ is the true value.

When bringing the data shown in Table 10 into Equation (19), the value of the MAPE is 2.03%. To sum up, when the original input variable system is adopted, the RMSE
is 3.54% and the MAPE is 2.03%. This shows that the prediction results without MIV processing have higher overall accuracy. However, the relative error of the test set (41) is as high as 12.5%, which does not meet the 10% prediction accuracy often required in engineering practice.

Step 6. Considering that the input variables may affect the prediction accuracy of the SSA-LSSVM, this study uses the MIV method to screen out the key variables. The value of each input variable is increased or decreased by 10%, and Steps 3–5 are re-executed. The calculation results with the best parameter combination of the LSSVM are shown in Table 9. The calculation results of the cumulative contribution rate of each input variable are shown in Table 11.

### Table 11. Cumulative contribution rate of input variables.

| Variables | MIV  | Order | Relative Contribution Rate | Cumulative Contribution Rate |
|-----------|------|-------|-----------------------------|------------------------------|
| F2        | 0.0184 | 1     | 17.78%                      | 17.78%                       |
| F4        | -0.0153 | 2     | 14.78%                      | 32.56%                       |
| F1        | 0.0137  | 3     | 13.24%                      | 45.80%                       |
| F9        | 0.0128  | 4     | 12.37%                      | 58.16%                       |
| F7        | 0.0117  | 5     | 11.30%                      | 69.47%                       |
| F3        | 0.0106  | 6     | 10.24%                      | 79.71%                       |
| F10       | 0.0065  | 7     | 6.28%                       | 85.99%                       |
| F6        | 0.0059  | 8     | 5.70%                       | 91.69%                       |
| F5        | 0.0047  | 9     | 4.54%                       | 96.23%                       |
| F8        | 0.0025  | 10    | 2.42%                       | 98.65%                       |
| F11       | 0.0014  | 11    | 1.35%                       | 100.00%                      |

The results in Table 11 reveal that 11 input variables have different influences on the prediction of economic loss due to waterlogging disaster in subway station engineering project. Among them, the MIV of F4 is negative, which indicates that the input variable is negatively correlated with disaster loss. This is consistent with the calculation results presented in Section 3.2. Other input variables are positively correlated with disaster losses. Among them, the relative contribution rates of F2, F4, F1, F9, F7 and F3 are all higher than 10%, while the relative contribution rates of the remaining input variables are lower than 10%. Therefore, F2, F4, F1, F9, F7 and F3 were selected as the key input variables for forecasting the disaster losses caused by waterlogging in subway station projects. Among these six key variables, four belong to disaster loss rate, and only two belong to asset value.

Step 7. According to the six key variables identified in Step 6, Steps 3–5 are re-executed. The adaptive function curve is shown in Figure 2, and the optimal parameter combination of the LSSVM is shown in Table 9. The RMSE was calculated to be 2.49%, while the MAPE was found to be 1.25%. Moreover, the relative error of Sample 41 is reduced from 12.5 to 7.5%, which meets the requirements of engineering practice. Compared with the original input variable system, the calculation accuracy of the SSA-LSSVM is significantly improved after MIV processing. In addition, in the same computing environment, the average computing time of 100 calculations when using the original input system was 322.17 s, while when using the six key variables the average computing time was 264.53 s. Thus, after the index screening, the calculation time is reduced by 57.64 s.

A 10-fold cross-validation is often used to test the accuracy of an algorithm [38]. The basic idea is to randomly divide the data set into 10 parts, and then take 9 parts as training data and 1 part as test data. There are 45 sets of data in the present case study, and the sample set cannot be randomly divided into 10 parts. Therefore, the data are randomly divided into nine groups, of which eight are used as training data and one is used as test data in turn. The calculation results are shown in Table 12.
Table 12. Calculation results of cross-validation.

| No. | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| MAPE| 1.25% | 1.42% | 1.68% | 0.65% | 1.36% | 1.07% | 1.45% | 1.37% | 1.49% |
| RMSE| 2.49% | 2.75% | 3.17% | 1.30% | 2.75% | 1.79% | 2.64% | 2.86% | 2.68% |

The calculation results shown in Table 12 reveal that the nine prediction results have good prediction accuracy and calculation stability, which shows that the calculation results of this case analysis are accurate and there is no overfitting phenomenon. In addition, in nine calculations, the relative errors of five test sets meet the engineering requirements, as they are all less than 10%.

The Bland–Altman analysis was performed on the predicted and measured values [39], as shown in Figure 3. The results in Figure 3 show that the predicted values of the five groups are within \((-1.96SD, +1.96SD)\). According to the Bland–Altman analysis method, 95% of the predicted points are within the consistent range. Therefore, from the perspective of engineering application, the prediction method in this study has good reliability.

![Figure 3. Results of the Bland–Altman analysis.](image)

The calculation results presented in Table 13 show that with the decrease of the sample set, the calculation accuracy gradually decreases. When the ratio of the sample set to test
set is 55.56%:44.44%, the RMSE is above 5%, which fails to pass the Bland–Altman test. Therefore, it is suggested that the sample set ratio should not be lower than 66.67%.

5. Discussion

Although in Section 3 the model proposed in this paper was successfully applied to 45 engineering projects, this section will discuss the calculation accuracy and stability of different prediction models and the influence of different key index screening methods on the calculation results. These discussions are useful to highlight the advances of the prediction methods proposed in this paper. In addition, this study mainly has the following three limitations: (1) More heuristic algorithms could have been used in this study. (2) If different input variable identification methods are adopted, the calculation results may be affected. (3) Only 45 projects were selected in this paper, and larger data sets are needed to verify the conclusions of this paper.

5.1. Analysis of the Calculation Accuracy of Different Prediction Models

To highlight the advancement of the model proposed in this paper, this study selected the multiple regression analysis (MRA) [10], BPNN [14], PSO-BPNN [14,31], SSA-BPNN [14,21], LSSVM [18], GA-LSSVM [18], PSO-LSSVM [19] and GWO-LSSVM [20] methods for comparative analysis. These previously developed methods were used to deal with the data in the case study, i.e., the data analyzed by the prediction model proposed in this paper.

When the MRA was used, the first 40 groups of data in Table 5 were taken as training sets, which were calculated by the Excel 2016 software (Microsoft Corp., Redmond, WA, USA):

\[
Y = 0.32 \times F_1 - 7.09 \times F_2 - 0.98 \times F_3 - 0.38 \times F_5 + 0.53 \times F_6 + 0.07 \times F_7 - 1.94 \times F_8 - 1.81 \times F_9. \tag{20}
\]

According to Equation (19), when using the MRA method, the disaster prediction value is only related to the change of f value of eight inputs, such as F1 and F2, but is not affected by F4, F10 and F11. The regression coefficient of F2 is the largest, which is consistent with the calculation results based on the MIV described in Section 4.3.

Then, the last five groups of data were taken as test sets, and the related calculation results are shown in Table 14.

| Model       | Number of Key Input Variables | RMSE    | MAPE   | Average Calculation Time | Bland–Altman |
|-------------|-------------------------------|---------|--------|--------------------------|--------------|
| MRA         | 8                             | 79.31%  | 48.23% | 5 s                      | Failed       |
| BPNN        | 6                             | 24.51%  | 18.59% | 21.45 s                  | Failed       |
| PSO-BPNN    | 5                             | 8.37%   | 6.92%  | 342.17 s                 | Failed       |
| SSA-BPNN    | 6                             | 7.14%   | 6.04%  | 325.45 s                 | Failed       |
| LSSVM       | 5                             | 18.31%  | 14.70% | 10.65 s                  | Failed       |
| GA-LSSVM    | 6                             | 10.49%  | 7.31%  | 482.64 s                 | Failed       |
| PSO-LSSVM   | 6                             | 5.12%   | 3.97%  | 304.56 s                 | Succeed      |
| GWO-LSSVM   | 5                             | 4.31%   | 2.57%  | 295.90 s                 | Failed       |
| SSA-LSSVM   | 6                             | 2.49%   | 1.25%  | 264.53 s                 | Succeed      |

Table 14. Calculation results of different models.

In the PSO, the maximum number of iterations is 200, the weight factor varies linearly from 0.4 to 0.9 and the acceleration constants C1 and C2 are both 2. In the BPNN, the sigmoid activation function is adopted, and the minimum mean square error is 10–5. The calculation accuracy of the LSSVM is related to the kernel function and penalty factor parameters. Ten groups of calculation parameters are randomly selected in this study, and the average values of RMSE and MAPE of ten calculation results are taken. In the GA, the number of individuals is 50, the maximum iteration algebra is 200, the generation gap is 0.95, the crossover probability is 0.7, the mutation probability is 0.01, the selection method is random traversal sampling, the crossover operator is the single point crossover and the mutation operator is the basic bit mutation. In the GWO, the search space is set as the regularization parameter \( \lambda \in [0.1, 100] \), kernel function width coefficient \( \sigma \in [0.01, 1000] \), population size is set as 30 and the maximum number of iterations is set as 200. The
convergence curves of the GA-LSSVM, PSO-LSSVM, GWO-LSSVM and SSA-LSSVM are shown in Figure 2 (in Section 4.3). The calculation results of different models are shown in Table 14. Except for the MRA, the MIV method was used to screen key variables in all prediction models shown in Table 14.

The results in Table 14 reveal that the optimization performance of the SSA is better than that of the GA, PSO and GWO. According to the BPNN, the RMSE and MAPE of the PSO-BPNN are 8.37 and 6.92%, respectively, and their calculation accuracy is clearly lower than that of the SSA-BPNN. Compared with the SSA-BPNN, the PSO-BPNN is also slower. According to the LSSVM, the RMSE and MAPE of the SSA-LSSVM are 2.49 and 1.25%, respectively, and their calculation accuracy is clearly higher than that of the GA-LSSVM, PSO-LSSVM and GWO-LSSVM. Additionally, the calculation speed of the SSA-LSSVM is also the fastest. The LSSVM is also superior to the BP in computing performance. Furthermore, the PSO-LSSVM and SSA-LSSVM have better computational performance than the PSO-BP and SSA-BP. In addition, only the PSO-LSSVM and SSA-LSSVM passed the Bland–Altman test.

5.2. Stability Analysis of Different Prediction Models

The better the stability, the higher the application and popularization value of the prediction model. The BPNN, PSO-BPNN, SSA-BPNN, LSSVM, GA-LSSVM, PSO-LSSVM, GWO-LSSVM and SSA-LSSVM models were calculated 200 times. The standard deviations of the RMSE and MAPE and the average calculation time of the 200 calculation results are shown in Table 15.

**Table 15. Stability analysis of different calculation methods.**

| Model     | Standard Deviation of RMSE | Standard Deviation of MAPE | Standard Deviation of Average Calculation Time |
|-----------|---------------------------|---------------------------|-----------------------------------------------|
| BPNN      | 0.039272                  | 0.03641                   | 0.4124                                        |
| PSO-BPNN  | 0.024760                  | 0.02732                   | 0.6213                                        |
| SSA-BPNN  | 0.010906                  | 0.01238                   | 0.2348                                        |
| LSSVM     | 0.024263                  | 0.03034                   | 0.0783                                        |
| GA-LSSVM  | 0.016311                  | 0.01831                   | 0.8312                                        |
| PSO-LSSVM | 0.006712                  | 0.00847                   | 0.5489                                        |
| GWO-LSSVM | 0.004271                  | 0.003412                  | 0.2642                                        |
| SSA-LSSVM | 0.001123                  | 0.000931                  | 0.0124                                        |

The calculation results in the Table 15 show that the standard deviation of the SSA-LSSVM model is the smallest and its calculation results are the most stable. In the optimization of the LSSVM, the computational stability of the SSA, GWO, PSO and GA decreases in turn. This shows that the order of the computational stability of these metaheuristic algorithms is SSA > GWO > PSO > GA, which is the same as previously reported [22–24].

5.3. Comparison of Calculation Results of Different Key Index Screening Methods

Currently, the commonly used index screening methods are PCA [26], MLRM [27] and FAM [28]. In this study, the above methods were used to identify the key indicators and bring them into the SSA-LSSVM model for calculation. The calculation results are shown in Table 16.

According to the calculation results shown in Table 16, when the MIV method is used, the calculation accuracy is the highest (RMSE = 2.49%, MAPE = 1.25%). The average calculation time is related to the number of input variables, and the larger the number of input variables, the longer the average calculation time. Only when the MIV method is used can the calculation results be analyzed by the Bland–Altman plot. After the index screening, the calculation performance of the SSA-LLSVM is improved. This highlights the necessity of screening key indicators. The comprehensive analysis shows that when compared with PCA, MLRM and FAM, the MIV is more suitable for identifying the key indicators in this study.
Table 16. Calculation results of different key index screening methods.

| Model         | The Key Input Variables | RMSE | MAPE | Average Calculation Time | Bland–Altman |
|---------------|-------------------------|------|------|--------------------------|--------------|
| Initial input system | All variables          | 3.54% | 2.03% | 322.17 s                 | Failed       |
| PCA           | F1, F2, F3, F5, F6, F7, F8, F9 | 4.24% | 3.58% | 283.59 s                 | Failed       |
| MLRM          | F1, F2, F4, F6, F8, F9, F10 | 3.03% | 2.57% | 289.32 s                 | Failed       |
| FAM           | F1, F2, F3, F5, F7, F9, F11 | 3.24% | 2.36% | 276.57 s                 | Failed       |
| MIV           | F1, F2, F3, F4, F7, F9    | 2.49% | 1.25% | 264.53 s                 | Succeed      |

6. Conclusions

Due to the many factors affecting the prediction of waterlogging disasters in subway station engineering projects and the nonstrong linear characteristics of the sample data, the traditional method has some shortcomings, such as low prediction accuracy and low calculation efficiency. In this study, a new intelligent forecasting method based on the SSA, LSSVM and MIV was developed, and an empirical study was conducted. The model combines the global convergence of the SSA and the nonlinear analysis ability of the LSSVM, and it improves the accuracy and stability of the LSSVM. This study shows that for the 45 subway station projects of lines 11 and 27 of the Chengdu Metro system and line 8 of the Wuhan Metro system, the linear relationship between the 11 input variables and the output variable is not sufficient. Only the catchment area and the contribution of accumulated completed output value have a significant linear correlation with disaster losses. The nonlinear modeling method should be given priority when predicting the economic loss due to the waterlogging disaster in a subway station engineering project. In this study, the MIV method was used to screen the characteristic variables that affect the prediction of economic losses due to waterlogging disaster in 45 subway stations. It was found that the maximum rainfall in 24 h, catchment area, unwelcome geological features, height of the retaining wall, total project investment and the number of large machines on the construction site are the key factors influencing the economic losses due to storm waterlogging disaster in this case study project. Compared with other prediction models (MRA, BPNN, PSO-BPNN, SSA-BPNN, LSSVM, GA-LSSVM, PSO-LSSVM and GWO-LSSVM), the proposed model had better prediction accuracy and computational stability. The future research direction in this field is the construction of a unified and universal input variable of waterlogging disaster loss.

Author Contributions: H.W. and J.W. conceived the presented idea. H.W. developed the theory and performed the computations. H.W. verified the analytical methods. J.W. completed the data collection and case analysis, which were the main technical difficulty of this study. H.W. and J.W. completed the revision of this manuscript together. All authors discussed the results and contributed to the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Project of Wuhan Urban and Rural Construction Bureau, grant number 201943.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The case analysis data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wu, X.; Cao, Y.; Xiao, Y.; Guo, J. Finding of urban rainstorm and waterlogging disasters based on microblogging data and the location-routing problem model of urban emergency logistics. *Ann. Oper. Res.* **2020**, *290*, 865–896. [CrossRef]

2. Zhou, L.; Wu, X.; Ji, Z.; Gao, G. Characteristic analysis of rainstorm-induced catastrophe and the countermeasures of flood hazard mitigation about Shenzhen city. *Geomat. Nat. Hazards Risk* **2017**, *8*, 1886–1897. [CrossRef]
3. Wu, H.; Wang, J. Assessment of Waterlogging Risk in the Deep Foundation Pit Projects Based on Projection Pursuit Model. *Adv. Civ. Eng.* 2020, 2020, 1–11. [CrossRef]

4. Zhao, G.; Wu, Z.G. Research on Cost Prediction Algorithm of Road Restoration and Reconstruction and Sudden Earthquake. *J. Catastrophol.* 2020, 35, 38–41.

5. Zhao, Y.; Gong, Z.; Wang, W.; Luo, K. The comprehensive risk evaluation on rainstorm and flood disaster losses in China mainland from 2004 to 2009: Based on the triangular gray correlation theory. *Nat. Hazards* 2014, 71, 1001–1016. [CrossRef]

6. Guo, J.; Wu, X.; Wei, G. A new economic loss assessment system for urban severe rainfall and flooding disasters based on big data fusion. *Environ. Res. 2020, 188, 109822.* [CrossRef]

7. Tan, L.; Wu, X.H.; Xu, Z.S.; Li, L.S. Comprehensive economic loss assessment of disaster based on CGE model and IO mod-el-A case study on Beijing “7.21 Rainstorm”. *Int. J. Disaster Risk Reduct.* 2019, 39, 101246. [CrossRef]

8. Kang, L.; Ma, L.; Liu, Y. Evaluation of farmland losses from sea level rise and storm surges in the Pearl River Delta region under global climate change. *J. Geogr. Sci.* 2016, 26, 439–456. [CrossRef]

9. Yang, F.; Ma, C.; Fang, H.J. Research progress on vulnerability: From theoretical research to comprehensive practice. *Acta Ecol. Sin.* 2019, 39, 441–453.

10. Song, Y.S.; Park, M.J. Development of Damage Prediction Formula for Natural Disasters Considering Economic Indicators. *Sustainability 2019, 11, 868.* [CrossRef]

11. Cao, Y.; Yin, K.D.; Li, X.M. Prediction of direct economic loss caused by marine disasters based on the improved GM(1,1) model. *J. Grey Syst.* 2020, 32, 133–145.

12. Yin, K.; Zhang, Y.; Li, X. Research on Storm-Tide Disaster Losses in China Using a New Grey Relational Analysis Model with the Dispersion of Panel Data. *Int. J. Environ. Public Health* 2017, 14, 1330. [CrossRef]

13. Sun, H.; Burton, H.V.; Huang, H.L. Machine learning applications for building structural design and performance assessment: State-of-the-art review. *J. Build. Eng.* 2021, 33, 101816. [CrossRef]

14. He, S.H.; Yang, B.; Dai, M.S. Flood disaster loss prediction based on dynamic recurrent neural network. *Chin. Econ.* 2014, 2014, 16–19.

15. Wang, T.T.; Liu, Q. The assessment of storm surge disaster loss based on BAS-BP model. *Mar. Environ. Sci.* 2018, 37, 457–463.

16. Zhao, X.; Li, H.; Ding, L.; Liu, M. Research and application of a hybrid system based on interpolation for forecasting direct economic losses of marine disasters. *Int. J. Disaster Risk Reduct.* 2019, 37, 101121. [CrossRef]

17. Zhao, Y.-P.; Wang, J.-J.; Li, X.-Y.; Peng, G.-J.; Yang, Z. Extended least squares support vector machine with applications to fault diagnosis of aircraft engine. *ISA Trans.* 2020, 97, 189–201. [CrossRef]

18. Cai, Z.; Xu, W.; Meng, Y.; Shi, C.; Wang, R. Prediction of landslide displacement based on GA-LSSVM with multiple factors. *Bull. Int. Assoc. Eng. Geol.* 2016, 75, 637–646. [CrossRef]

19. Xue, X.H. Evaluation of concrete compressive strength based on an improved PSO-LSSVM model. *Comput. Concr. 2018, 21, 505–511.*

20. Zeng, B.; Guo, J.; Zhang, F.; Zhu, W.; Xiao, Z.; Huang, S.; Fan, P. Prediction Model for Dissolved Gas Concentration in Transformer Oil Based on Modified Grey Wolf Optimizer and LSSVM with Grey Relational Analysis and Empirical Mode Decomposition. *Energies 2020, 13, 422.* [CrossRef]

21. Xue, J.; Shen, B. A novel swarm intelligence optimization approach: Sparrow search algorithm. *Syst. Sci. Control. Eng.* 2020, 8, 22–34. [CrossRef]

22. Lv, X.; Mu, X.D.; Zhang, J.; Wang, Z. Chaos sparrow search optimization algorithm. *J. Beijing Univ. Aeronaut. Astronaut.* 2020, 12, 1–10.

23. Tang, A.D.; Han, T.; Xu, D.W.; Xie, L. Path planning method of unmanned aerial vehicle based on chaos sparrow search al-gorithm. *J. Comput. Appl.* 2020, 11, 1–11. (In Chinese) [CrossRef]

24. Koppen, M.; Wolpert, D.; Macready, W. Remarks on a recent paper on the “no free lunch” theorems. *IEEE Trans. Evol. Comput.* 2001, 5, 295–296. [CrossRef]

25. Kapetanakis, D.S.; Mangina, E.; Finn, D.P. Input variable selection for thermal load predictive models of commercial build-ings. *Energy Build.* 2017, 137, 13–26. [CrossRef]

26. Zong, R.; Zhi, Y.; Yao, B.; Gao, J.; Stec, A.A. Classification and identification of soil source with principal component analysis and back-propagation neural network. *Aust. J. Forensic Sci.* 2013, 46, 224–233. [CrossRef]

27. Cui, R.X.; Liu, Y.D.; Fu, J.D. Estimation of winter wheat biomass using visible spectral and BP based artificial neural net-works. *Spectrosc. Spectr. Anal.* 2015, 35, 2596–2601.

28. Ding, S.; Jia, W.; Su, C.; Zhang, L.; Liu, L. Research of neural network algorithm based on factor analysis and cluster analysis. *Neural Comput. Appl.* 2010, 20, 297–302. [CrossRef]

29. Yan, C.H.; Zhang, T.L.; Sun, Y.Q.; Tang, H.S.; Li, H. A hybrid variable selection method based on wavelet transform and mean impact value for calorific value determination of coal using laser-induced breakdown spectroscopy and kernel ex-treme learning machine. *Spectrochim. Acta Part B At. Spectrosc.* 2019, 154, 75–81. [CrossRef]

30. Sun, S.; Lu, H.; Tsui, K.-L.; Wang, S. Nonlinear vector auto-regression neural network for forecasting air passenger flow. *J. Air Transp. Manag.* 2019, 78, 54–62. [CrossRef]

31. Gu, Y.; Zhang, Z.; Zhang, D.; Zhu, Y.; Bao, Z.; Zhang, D. Complex lithology prediction using mean impact value, particle swarm optimization, and probabilistic neural network techniques. *Acta Geophys.* 2020, 68, 1727–1752. [CrossRef]

32. Liu, L.J.; Wu, H.; Wang, J.W.; Yang, T.Y. Research on the evaluation of the resilience of subway station projects to waterlog-ging disasters based on the projection pursuit model. *Math. Biosci.* 2020, 17, 7302–7331. [CrossRef]
33. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [CrossRef]

34. Leppink, J.; Pérez-Fuster, P. We need more replication research-A case for test-retest reliability. *Perspect. Med. Educ.* **2017**, *6*, 158–164. [CrossRef] [PubMed]

35. Xu, H.; Deng, Y. Dependent Evidence Combination Based on Shearman Coefficient and Pearson Coefficient. *IEEE Access* **2017**, *6*, 11634–11640. [CrossRef]

36. Wang, X.; Wang, B. Research on prediction of environmental aerosol and PM2.5 based on artificial neural network. *Neural Comput. Appl.* **2018**, *31*, 8217–8227. [CrossRef]

37. Khatib, T.; Mohamed, A.; Sopian, K.; Mahmoud, M. Estimating ambient temperature for malaysia using generalized regression neural network. *Int. J. Green Energy* **2012**, *9*, 195–201. [CrossRef]

38. Dankwa, S.; Zheng, W.F. Special issue on using machine learning algorithms in the prediction of kyphosis disease: A comparative study. *Appl. Sci.* **2019**, *9*, 3322. [CrossRef]

39. Francq, B.G.; Govaerts, B. How to regress and predict in a Bland-Altman plot? Review and contribution based on tolerance intervals and correlated-errors-in-variables models. *Stat. Med.* **2016**, *35*, 2328–2358. [CrossRef]