Auxiliary optimization of wastewater monitoring in infectious diseases

Mingshuang Gu[*], Mengmeng Zhao, Bo Chen

College of Engineering, Tibet University, Lhasa, 850000, China
[*]Corresponding author: 19934196350@163.com

Abstract: *In order to reduce the cost of infectious disease surveillance and to quickly identify the areas of occurrence of infectious diseases, Wastewater Based Epidemiology (WBE) is applied to infectious disease analysis. To address the problem of wastewater monitoring, this paper downloaded the data of 701 sampling points of wastewater monitoring in the United States from the Center for Disease Control and Prevention website, and analyzed whether the location of wastewater monitoring stations was reasonable and adjusted them by using the fuzzy comprehensive judgment method and TSP algorithm. In order to be able to use the monitoring data of the monitoring sites more accurately to determine the development of infectious diseases in their areas, this paper analyzed the wastewater data (from CDC) of the monitoring sites, studied their change patterns, and established SVR models for prediction.*

Keywords: *Wastewater Monitoring, Infectious Diseases, Fuzzy Integrated Judgment, TSP Algorithm, SVR Forecast*

1. Introduction

With the global ravages of infectious diseases, it will not only pose a threat to people's health, but also increase the economic burden of the country. Several scholars have used various prediction models to forecast infectious diseases, for example, Shunyong Li and Jinli He proposed a combined infectious disease prediction model based on adaptive noise-complete ensemble empirical modal decomposition (CEEMDAN) and fuzzy entropy (FE) improved long and short term memory network (LSTM), but there are limitations in parameter search [1]. Dai Haoyun et al. constructed an autoregressive moving average model (ARIMA) for influenza to predict influenza incidence trends [2]. Fang Kuangnan et al. concluded that the use of dynamic SEIR model to study the transmission trend of sudden outbreak infectious diseases is more consistent with the viral transmission characteristics of infectious diseases and has better prediction effect [3]. For sudden outbreaks of large scale major infectious diseases, wastewater-based epidemiology (WBE) studies have now become a valuable tool that is widely adopted around the world. Since most of the infectious disease data have nonlinear characteristics such as small samples and irregularities, and SVR has been widely used in machine learning since its introduction, it is often used to solve small sample and nonlinear problems, and SVR also has strong nonlinear fitting performance and generalization performance, so the processing of such data using support vector machines has specific advantages. In this study, wastewater data are used as the main study data, and the data are nonlinearly mapped into the feature space of high latitude by SVR model, and a linear model is obtained by linear fitting, which can be better for early warning of large scale emergence of infectious diseases.

2. Selection of wastewater monitoring points based on fuzzy comprehensive evaluation and TSP

2.1 Fuzzy integrated evaluation

In 1965, Zadeh, a professor in the Department of Electrical Engineering and Computer Science at the University of California, Berkeley and an expert in automatic control, published the article "Fuzzy Sets", which successfully described fuzzy concepts using exact mathematical methods for the first time, thus announcing the birth of fuzzy mathematics, and the fuzzy sets it introduced provided a new method for analyzing complex systems [4]. Academicians such as Guan Zhaozheng, Pu Baoming, Li Guoping and Wang Peizhuang promoted the development and research of fuzzy mathematics in China and were the pioneers and pioneers of fuzzy set theory research in China [5].

2.1.1 Principle of fuzzy comprehensive evaluation

In the evaluation, certain evaluation indicators carry a certain degree of fuzziness and do not have very clear boundaries, thus lacking precise feedback. Fuzzy comprehensive evaluation is also a fuzzy mathematical algorithm established in the evaluation process to quantify and synthesize the realistic non-linear evaluation and finally obtain comparable quantitative results. The use of fuzzy mathematical methods for comprehensive evaluation will also be closer to the reality [6]. Fuzzy mathematics is the use of mathematical methods and analysis of objective fuzzy phenomena, fuzzy mathematics makes a general evaluation of things or phenomena affected by multiple factors, and is more effective in judging complex problems with multiple factors and levels [7].

2.1.2 Fuzzy comprehensive evaluation method model construction

(1) Determine the domain of factors of the evaluation object, which can be set n evaluation indicators:

$$U = U1, U2, \dots, Un \tag{1}$$

In this paper that is the water sample siting index evaluation system contained in the line corridor, the overall layout, price cost... Water source conditions and other evaluation indicators, that is: U = {line corridor, general layout, price cost, ... The above indicators are generally different to some extent. The above indicators generally have a certain degree of uncertainty, in the fuzzy comprehensive evaluation, usually using the maximum affiliation to express.

(2) Determine the rubric level domain, the rubric level domain is a set consisting of the different levels of evaluation results obtained by the evaluator on the evaluated object as elements, i.e.

$$V = \{V_1, V_2, \dots, V_n\} \tag{2}$$

(3) Build fuzzy relationship matrix

The comprehensive evaluation of the evaluation object should start from the single-factor fuzzy evaluation, that is, after determining the fuzzy affiliation of each factor The fuzzy affiliation of the evaluated object for each rating level in the rubric set under each factor, so as to obtain the single-factor fuzzy evaluation matrix. Assume that the evaluation results of single factor can be expressed as

$$R_i = (r_{i1}, r_{i2}, \dots, r_{in}) \tag{3}$$

After constructing the hierarchical fuzzy subset, the evaluated things are quantified one by one from each factor $\mathrm{Ui}(\mathrm{i} = 1,2\dots,\mathrm{n})$, that is, the affiliation degree (R/O) of the evaluated things to the hierarchical fuzzy subset from a single factor is determined and thus the fuzzy relationship matrix is obtained as follows:

$$\mathrm{R} = \begin{bmatrix} R| & u_1 \\ R| & u_2 \\ \cdots & \cdots \\ R| & u_n \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix} \tag{4}$$

(4) Determine the weight vector of evaluation factors

In fuzzy comprehensive evaluation, the weight vectors of evaluation factors are determined, and this paper uses hierarchical analysis to determine the evaluation indexes among the relative order of importance of the evaluation indicators is determined so that the weight coefficients are determined and normalized before synthesis [8].

(5) Synthesis of fuzzy integrated evaluation result vector

The weight set W obtained by using the hierarchical analysis method is synthesized with the fuzzy relationship matrix R of the evaluated thing (a generalized synthesis operation is indicated here, i.e., the synthesis between the weight vector and the single-factor evaluation matrix of the evaluated object), and the fuzzy comprehensive evaluation result vector S of each evaluated thing is obtained, namely.

$$W^{\circ}R = (w_1, w_{2,} \cdots, w_p) \begin{bmatrix} r & r & \cdots & r \\ r & r & \cdots & r \\ \cdots & \cdots & \cdots & \cdots \\ r & r & \cdots & r \end{bmatrix} = (s_1, s_2, \cdots s_m) = S \tag{5}$$

In equation (5): where S1, is obtained from the i-th column operation of W and R, which indicates the affiliation degree of the evaluated thing to the fuzzy subset of V hierarchy level from the overall view.

(6) Analysis of fuzzy comprehensive evaluation result vector

The evaluation object is evaluated according to the principle of maximum affiliation. However, it can be used reluctantly in some cases and more information is lost, and even make the evaluation results unreasonable. In order to improve the accuracy, this paper adopts the method of weighted average to find the affiliation level:

$$V_f = b_1 \times v_1 + b_2 \times v_2 + \cdots b_n \times v_n \tag{6}$$

Based on the magnitude of the Vf result, the degree of merit of the scheme is determined and the best address is selected among the schemes.

2.2 TSP algorithm

The TSP algorithm whose name is Traveling salesman problem (TSP problem) for short, is a mathematical planning problem proposed in 1959, TSP belongs to the typical NP-complete problem, the language of the TSP problem is described as the solution space S of SP is the set of all loops that traverse each set point exactly once and is the set of all permutations of set points. The solution space S of the TSP problem can be expressed as the set of all permutations of {1,2,...,n} [3], and the optimal solution of the simulated TSP algorithm is independent of the initial state, so the initial solution is a random function generating a random permutation of {1,2,...,n} as S_0. Its cost function is shown in equation (7):

$$C(c_1, c_2, \cdots, c_n) = \sum_{i=1}^{n+1} d(c_i, c_{i+1}) + d(c_1, c_n) \tag{7}$$

Now the solution of the TSP problem is to find the minimum of the objective function C(c1,c2,...,cn) by the simulated annealing algorithm , and accordingly, s*=(c*1,c*2,...,c*n) is the optimal solution of the TSP problem New solution generation New solution generation is very important for the solution of the problem. The new solution can be generated either separately or alternatively by the following 2 methods: exchanging the access order between u and v. If the solution before the exchange is si= (c1,c2,...,cn), the path after the exchange is the new path, i.e., as shown in Equation (8) can either.

$$S_i = (c_1, \cdots c_{u-1}, c_v, c_{v-1}, \cdots, c_{u+1}, c_u, c_{v+1}, \cdots, c_n) \tag{8}$$

The difference between the solution before the transformation of the objective function and the objective function after the transformation is calculated as

$$\Delta c' = c(s_i') - c(s_i) \tag{9}$$

where the Metropolis acceptance criterion accepts si′ as the new current solution si based on the difference of the objective function and the probability exp(-ΔC′/T), and the acceptance criterion Equation (10)

$$p = exp(-\Delta c'/T), -\Delta c' > 0 \tag{10}$$

2.3 Model solving

2.3.1 Fuzzy integrated evaluation model solving

1) Determine the weights of the first level indicators

Monitoring the site selection on the line corridor U1-1, general layout U1-2, price and cost U1-3, water conditions U1-4. The following table 1 shows the matrix of indicators.

Table 1: Primary indicator judgment matrix

Programs	Line corrodor	General layout	Fee	Water conditions
	U1-1	U1-2	U1-3	U1-4
U1-1	1	1/3	3	1/2
U1-2	3	1	3	1
U1-3	1/3	1/3	1	1/2
U1-4	2	1/2	1/3	1/3

From the judgment matrix, it can be derived that W1(0.181,0.333,0.163,0.178), the maximum eigenvalue λmax=6.666, and by the consistency CI/IR=0.094<0.1, which passed the consistency test, therefore, the weight values of each factor of the scheme and indicators are as follows Table 2 below

Table 2: Weighting of primary indicators

Serial numbr	Tier 1 indicators	Indicator Weighting
1	Line corrodor	0.334
2	General layout	0.065
3	Fee	0.075
4	Water conditions	0.036

2.3.2 TSP algorithm for solving sewage monitoring address

Our group has obtained some influencing factors affecting wastewater monitoring by reviewing the literature and combining with the actual situation. They are Human factor, time factor, network distribution for this purpose, the processed data are imported into Matlab, and then according to the principle of TSP algorithm The relevant data are obtained according to the principle of TSP algorithm. Before data training, the data need to be normalized, and the normalization mainly deals with the presence of singular samples. These samples can affect the convergence speed of the model. These samples can affect the convergence speed of the model, and sometimes even cause the model to fail to converge, so the samples need to be normalized before training. Therefore, we need to normalize the samples before training to ensure the training accuracy and convergence of the model.

After the data of the table, the data cannot be used prematurely need to find the time results of the efficiency fit analysis The analysis principle for the output set and the sample set of the degree of fit verification, the verification formula (11) is:

$$y(t) = f\big(x(t-1), \dots, x(t-m), y(t-1) \dots, y(t-n)\big) \tag{11}$$

The results of calculating the data set using Matlab are shown in Figure 1.

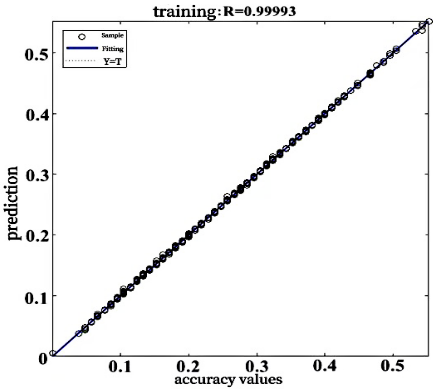

Figure 1: Data fitting table of efficiency rationality

Analysis of Figure 1 shows that there is a positive relationship between the output set and the expected set, i.e., the numerical fit is good, which means that it can be be used. Through the analysis, it is determined that its increase of ten points should be added in as shown in Figure 2(map source: CDC website).

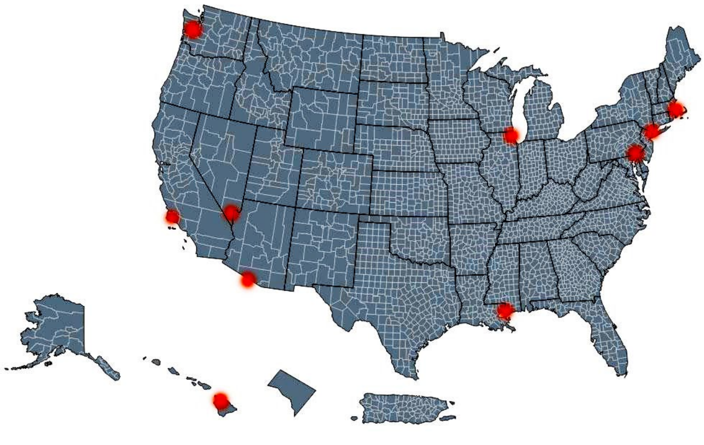

Figure 2: Add point distribution map

Most of the increased points are in the developed coastal cities, where the flow of people is greater and the contact with things to people is more complicated. Therefore, the 10 points set up are mainly concentrated in places where people are more concentrated. Therefore, the 10 points set up are mainly concentrated in places where people are more concentrated.

(1) Infectious disease prediction based on SVR support vector machine

Support Vector Machine (SVM) was originally created by Vapnik et al. based on the principle of structural risk minimization Vapnik et al. created a general-purpose learning method based on the principle of structural risk minimization and the VC-dimensional theory of statistical theory [9]. Support vector machines were first proposed to solve. It is characterized by the introduction of the structural risk function into the classification problem. The introduction of structural risk function improves the generalization ability of machine learning and cleverly solves the nonlinear classification problem, which can effectively achieve the introduction of structural risk function improves the generalization ability of machine learning and cleverly solves the nonlinear classification problem, which can effectively achieve data mining effects such as classification and regression. Support vector machines have two kinds of applications in real life, namely support vector with the in-depth research of support vector machine theory, many scholars have proved support vector machine by numerous examples. Support vector regression is one of the most popular methods. Support Vactor Regression (SVC) is a regression algorithm based on the statistical learning theory machine learning algorithm support vector machine SVM idea. Support Vactor Regression is a regression algorithm based on the statistical learning theory machine learning algorithm support vector machine SVM idea, and its basic idea is that the selected sampling point data are mapped to the high-latitude characteristic space by nonlinear mapping. The basic idea is that the selected sampling point data are mapped to the characteristic space of high latitude by nonlinear mapping, and the linear model is obtained by linear fitting in the space of high latitude. Since its introduction, SVR has been widely used in machine learning to solve small-sample, nonlinear problems, and has strong nonlinear fitting and generalization properties.

Advantages: Compared with neural networks, the structure of SVR is a risk minimization strategy, which is superior in dealing with small samples and nonlinear problems. The training results are as follows Figure 3:

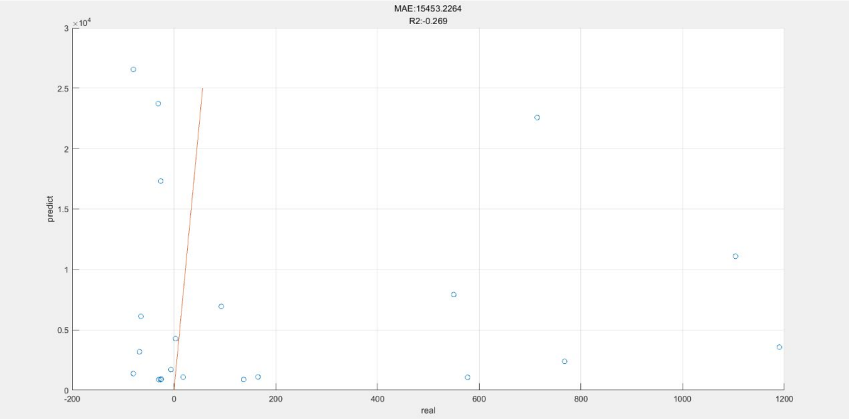

Figure 3: Training effect of SVR neural network

Urgent issues to be addressed:

Support vector machine regression has some shortcomings in building models for large training samples. The next step is to optimize the model algorithm to ensure that the support vector machine regression model is more suitable for handling large data problems.

Algorithm:

Let the linear regression function established in the high-latitude eigenspace.

$$f(x) = w\Phi(x) + b \tag{12}$$

$\Phi(X)$ is a nonlinear mapping function.

Defining ε as a linearly insensitive loss function, the

$$L(f(x), y, \varepsilon) = \begin{cases} 0, \ldots |\mathrm{y} - \mathrm{f(x)}| \leq \varepsilon \\ |\mathrm{y} - \mathrm{f(x)}| - \varepsilon \ldots |\mathrm{y} - \mathrm{f(x)}| > \varepsilon \end{cases} \tag{13}$$

f(x) is the regression function returning the predicted value; y is the corresponding true value.

Introducing the slack variables ξi, ξi* , the w,b problem is described in mathematical terms as

$$\begin{cases} \underset{w,b,c}{\mathrm{Min}} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{N} (\xi_i - \xi_i{}^*) \\ \mathrm{s.t} \quad y_i - (w\Phi(x_i) + b) \leq \varepsilon + \xi_i, \quad \mathrm{i} = 1,2 \cdots, \\ -y_i + (w\Phi(x_i) + b) \leq \varepsilon + \xi_i{}^*, \quad \mathrm{i} = 1,2 \cdots, \\ \xi_i \geq 0, \xi_i{}^* \geq 0 \end{cases} \tag{14}$$

C is the penalty factor, and the magnitude of C value is positively correlated with the size of the sample penalty of ε; ε puts the regression function error into is specified, and a smaller ε indicates a smaller error of the regression function.

Introducing the Largrange function, which can be converted to the pairwise form [10].

$$\begin{cases} \max_{a,a^*} \left[-\frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} (a_i - a_i^*)(a_j - a_j^*) K(x_i, x_j) - \sum_{i=1}^{1} (a_i + a_i^*) \varepsilon + \sum_{i=1}^{l} (a_i) \right] \\ s.t. \begin{cases} \sum_{i=1}^{l} (a_i - a_i^*) = 0 \\ 0 \leq a_i \leq C \\ 0 \leq a_i^* \leq C \end{cases} \end{cases} \tag{15}$$

$K = (x_i, x_j) = \Phi(x_i)\Phi(x_j)$is the kernel function.

Optimal solution $\alpha = [\alpha_1, \alpha_2, \alpha_3, \cdots \alpha_l], \alpha^* = [\alpha_1^*, \alpha_2^*, \alpha_3^*, \cdots \alpha_l^*]$

$$\boldsymbol{W}^* = \sum_{\boldsymbol{i=1}}^{\boldsymbol{l}} (a_i - a_i^*) \Phi(x_i) \tag{16}$$

$$\mathrm{b}^* = \frac{1}{N_{nsv}} \left\{ \sum_{0<a_{j<C}} \left[y_i - \sum_{x_i \epsilon \mathrm{SV}} (a_i - a_i^*) \mathrm{K}(x_i, x_j) - \varepsilon \right] + \sum_{0<a_{i<C}} \left[y_i - \sum_{x_j \epsilon \mathrm{SV}} (a_j - a_i^*) \mathrm{K}(x_i, x_j) + \varepsilon \right] \right\} \tag{17}$$

$\boldsymbol{N_{nsv}}$ is the number of support vectors.

The regression function can be expressed as

$$f(x) = w^*\Phi(x) + b^* = \sum_{i=1}^{l} (a_i - a_i^*) \Phi(x_i)\Phi(x) + b^* = \sum_{i=1}^{l} (a_i - a_i^*) K(x_i, x) + b^* \tag{18}$$

Some of the parameters $(a_i - a_i^*)$are not zero, corresponding to the sample x_i as the support vector in the problem.

(2) Analysis of experimental results

According to the solution of the model, the R2 of the fit reached 0.91, which is a good fit and can be used as a basis for measuring the development of the epidemic. The model has a good fit and can be used as a basis for predicting the development of the epidemic. The model evaluation results are presented in Table 3.

Table 3: SVR model evaluation

Model Evaluation	MAE	MSE	RMSE	R^2
SVR	151.70	3556.40	188.60	0.91

3. Conclusions

In this paper, we analyzed whether the location of wastewater monitoring stations is reasonable by analyzing data from 701 wastewater monitoring sampling points in the United States and establishing SVR prediction support vector regression models, and obtained the following main conclusions:

(1) The location of monitoring sites has a greater impact on the efficiency and quality of new crown monitoring, and the location of monitoring sites is very important.

(2) The TSP algorithm was used to find the best location for the 10 new points, so that the selected monitoring point locations have good monitorability and representativeness, and can be well monitored at the same time.

(3) The SVR support vector machine was used to predict the development of the epidemic, and the fit of the SVR model, R^2, was above 0.91, which was a good fit.

References

[1] Li Shunyong, He Jinli. CEEMDAN-FE-LSTM infectious disease prediction based on empirical modal decomposition [J]. Henan Science, 2022, 40(08), 1205-1212.

[2] Dai Haoyun, Zhou Nan, Ren Xiang, Luo Piaoyi, Yi Shanghui, Quan Meifang, Cha Wenting, Lv Yuan. Prediction of epidemic characteristics and trends of various subtypes of influenza based on autoregressive moving average model [J]. Disease Surveillance, 2022, (10), 1338-1345.

[3] Fang Kuangnan, Ren Rui, Zhu Jianping, Ma Shuangqi, Wang Xiaofeng. Infectious disease prediction and policy evaluation based on dynamic SEIR model [J]. Journal of Management Science, 2022, 25(10), 114-126.

[4] Zadeh L.A. Fuzzy sets [J]. Information and control, 1965 (8), 338-353. Fangfang. Research on power load forecasting based on Improved BP neural network [D]. Harbin Institute of Technology, 2011.

[5] Gao Zhenhua. Research on substation site selection based on fuzzy comprehensive evaluation method [D]. Shandong University of Science and Technology, 2019.

[6] Li Hongping. A fuzzy mathematical model for evaluating the quality of higher mathematics classroom teaching [J]. Journal of Hunan Institute of Science and Technology, 2007, (12), 16-17+20.

[7] Jin Bo, Lu Lei. Research on the improvement method of distribution transformer condition evaluation weight coefficient [J]. Transmission and Distribution Engineering and Technology, 2019, 8(3), 102-111.

[8] Cao Xiaohong, Yang Changde, Meng He, Abibail Buybaiti. Analysis of factors influencing the stability of north and south slopes based on fuzzy comprehensive evaluation method [J]. Gansu Science and Technology, 2019, 35(24), 57-62.

[9] E Xu, Zhou Yi, Li Shyzhu. Fish toxicity prediction model based on artificial bee colony algorithm optimized SVM [J]. Journal of Bohai University (Natural Science Edition), 2021, 42(04), 357-362.

[10] Gao Baocheng, Tao Bowen. Concrete strength prediction based on SVR algorithm [J]. Urban Housing, 2019, 26(04), 143-146.