water

MDPI

Article

Prediction of Sewage Treatment Cost in Rural Regions with Multivariate Adaptive Regression Splines

Yumin Wang [1,*], Lei Wu [1] and Bernard Engel [2]

[1] School of Energy and Environment, Southeast University, 2# Sipailou Street, Nanjing 210096, China; wulei@seu.edu.cn

[2] Department of Agricultural and Biological Engineering, Purdue University, West Lafayette, IN 47906, USA; engelb@purdue.edu

* Correspondence: wangyumin@seu.edu.cn

Received: 14 December 2018; Accepted: 21 January 2019; Published: 23 January 2019

Abstract: In this paper, to interpret the cost structure of decentralized wastewater treatment plants (DWWTPs) in rural regions, a simple nonparametric regression algorithm known as multivariate adaptive regression spline (MARS) was proposed and applied to simulate the construction cost (CC), operation and maintenance cost (OMC), and total cost (TC). The effects of design treatment capacity (DTC), removal efficiency of chemical oxygen demand (RCOD), and removal efficiency of ammonia nitrogen (RNH_3-N) on the cost functions of CC, OMC, and TC were analyzed in detail. The results indicated that: (1) DTC is the most important parameter to determine cost structure with relative importance of 100%, followed by RCOD and RNH_3-N with relative importance of 16.55%, and 9.75%, respectively; (2) when DTC is less than 5 m^3/d, the slopes of CC and TC on DTC are constants of 1.923 and 1.809, respectively, with no relationship with RCOD and RNH_3-N; (3) when DTC is less than 20 m^3/d, the OMC is a constant of 435 RMB/year; and (4) in other cases, CC, OMC, and TC are related to RCOD and RNH_3-N besides DTC. Compared with widely used support vector machine (SVM) models and multiple linear regression (MLR) models, the MARS model has better statistical significance with greater *R* values and smaller *RMSE* and *MAPE* values, which indicated that the MARS model is a better way to approximate the cost for DWWTPs.

Keywords: multivariate adaptive regression spline (MARS); decentralized wastewater treatment plants (DWWTPs); construction cost (CC); operation & maintenance cost (OMC); total cost (TC)

1. Introduction

"New rural construction" proposed in 2005 in China, is a new policy to realize sustainable development in rural regions with a prosperous economy, perfect facilities, a beautiful environment, and a harmonious civilization [1,2]. However, increasing amount of domestic sewage was drained into rural water environment without proper treatment. To construct "new rural", economic and effective sewage treatment facilities are needed [3]. In rural regions with limited budgets, the cost structure of wastewater treatment including construction cost (CC), operating and maintenance cost (OMC), and total cost (TC) require better understanding to help create economically feasible water quality management programs in the future, and to help in the planning of wastewater treatment plants [4–6].

In the past, cost structures of municipal wastewater treatment plants (MWWTPs) were studied in lots of literatures, and wastewater treatment capacity was the primary consideration. Regression methods, such as simple linear regression, multiple linear regression, non-linear regression were applied to evaluate the relationship between treatment capacity and treatment cost [4,7]. Since uncertainties exist in cost estimations, such as wastewater generation, treatment, and reuse, fuzzy technology was integrated into regression models to generate fuzzy linear regression models, fuzzy

nonlinear regression models, and fuzzy goal regression models. The input and output variables and regression coefficients were taken as fuzzy numbers in fuzzy regression models [4,7–10]. The effects of treatment level (secondary treatment and advanced secondary treatment) or treatment units (anaerobic pond, anaerobic tank, constructed wetland) combined with treatment capacity on CC and OMC of MWWTPs were also analyzed and modeled [7–9,11]. In rural regions, the CC and OMC of wastewater treatment plants (WWTPs) are higher than urban regions due to a smaller and dispersed population, intermittent sewage discharge, lower water supply standard, significant fluctuation of water consumption, smaller sewage treatment scale, more complex topographic conditions, and more difficult to collect wastewater. To decrease CC and OMC of WWTPs in rural regions, the decentralized wastewater treatment plants (DWWTPs), which means that sewage is collected according to the zoning and the sewage is treated separately in each zone, are suitable. Therefore, DWWTPs are drawing wider interest from all over the world, especially water-deficient countries and regions [12,13]. However, only treatment capacity is given to the cost structure of DWWTPs [12–14]. In MWWTPs, it was reported that removal of biological oxygen demand (BOD) and nutrients affects the cost significantly due to the substantial increase of tank volume for nitrification [15]. In this paper, the effects of removal of organic pollutants and nutrients on the cost of DWWTPs were analyzed by proposing a multivariate adaptive regression spline (MARS) model. By partitioning training data sets into separate piecewise linear segments (splines) of differing gradients (slopes), a MARS model fit the relationship between a set of input variables and dependent variables in high-dimensional data [16], which has been successfully applied to predict and classify in many engineering fields [17–21]. The MARS model has the advantages of suitable for nonlinear systems, easy to interpret, and variables appear in the resulting model directly. Additionally, MARS does not need to assume the distribution of the predictor variables, which is very important because the variables in cost models are not normally distributed [20]. However, MARS has not been applied to predict wastewater treatment costs to date.

In this paper, the MARS model was applied to predict CC, OMC, and TC for DWWTPs in rural regions, which can help analyze cost-efficiency more accurately. In the MARS model, factors taken into account included not only the design treatment capacity (DTC), but also the removal efficiency (R) of chemical oxygen demand (COD) and ammonia nitrogen (NH_3-N), termed as RCOD and RNH_3-N, respectively.

2. Methods

2.1. Data Set

The cost data for rural sewage treatment systems were collected from 215 sets of DWWTPs located in Changshu region of Jiangsu Province in China, which were in operation. The research scope included 11 districts: Bixi District, Southeast Street, Yushan Town, Meili Town, Haiyu Town, Guli Town, Shajiabang Town, Zhitang Town, Dongbang Town, Shanghu Town, and Xinzhuang Town. Since rural wastewater has the characteristic of fair biological treatability without toxic or harmful substances, four treatment technologies were adopted: membrane bioreactor (MBR), sequencing batch reactor (SBR), biological filter and artificial wetland (BFAW), and purification tank (PT). In the cost model of MWWTPs, treatment capacity and treatment level are regarded as two most important drivers [8,22]. However, in rural regions, the sewage is primarily consisted of domestic wastewater, in which COD and NH_3-N are the main pollution factors [23]. Therefore, RCOD and RNH_3-N were selected to represent treatment level. The mean values of parameters with various treatment capacities are shown in Table 1.

Table 1. Mean value of parameters for the selected samples.

x_1 (DTC) (m^3/d)	x_2 (RCOD)	x_3 (RNH_3-N)	y_1 (CC) (10^4 RMB/Year)	y_2 (OMC) (10^4 RMB/Year)	y (TC) (10^4 RMB/Year)
1 (109)	0.57 (0.14)	0.63 (0.15)	0.08 (0.02)	0.01 (0.00)	0.09 (0.02)
2 (9)	0.64 (0.11)	0.65 (0.11)	0.25 (0.13)	0.02 (0.01)	0.27 (0.13)
5 (36)	0.62 (0.17)	0.58 (0.15)	0.92 (0.14)	0.06 (0.03)	0.98 (0.15)
10 (23)	0.64 (0.17)	0.55 (0.16)	1.57 (0.39)	0.11 (0.08)	1.68 (0.38)
15 (16)	0.66 (0.08)	0.61 (0.07)	1.63 (0.45)	0.12 (0.11)	1.76 (0.47)
20 (9)	0.69 (0.14)	0.59 (0.12)	1.94 (0.23)	0.21 (0.11)	2.15 (0.27)
45 (3)	0.64 (0.09)	0.52 (0.17)	4.34 (0.85)	0.38 (0.21)	4.72 (1.01)
50 (3)	0.74 (0.05)	0.68 (0.04)	5.45 (1.58)	0.6 (0.3)	6.05 (1.86)
60 (5)	0.62 (0.17)	0.61 (0.10)	5.53 (0.67)	0.3 (0.18)	5.83 (0.78)
100 (1)	0.61 (0)	0.57 (0)	6.01 (0)	0.5 (0)	6.52 (0)
110 (1)	0.71 (0)	0.63 (0)	11.54 (0)	0.57 (0)	12.11 (0)

Note: In column 1, numbers in parenthesis refer to the set numbers of specific design treatment capacity, and in other columns, numbers in parenthesis refer to standard deviations.

In Table 1, x_1 refers to DTC (ranges from 1 m^3/d to 110 m^3/d), x_2 stands for RCOD, x_3 represents RNH_3-N, y is TC including y_1 (CC) and y_2 (OMC). In this paper, CC and OMC refer to annual construction cost, and annual operation and maintenance cost, respectively. The annual construction cost is obtained using Equation (1):

$$\mathrm{CC} = \frac{r \cdot (1+r)^t}{(1+r)^t - 1} \cdot IC \tag{1}$$

where IC is the investment cost (10^4 RMB); r is the discount rate, which is set to be 0.035; and t is the expected life of the plant, which is assumed to be 10 years.

The construction costs were supplied by the Construction Bureau in Changshu City, and the operation and maintenance costs were supplied by Suzhou Hongyu Wastewater Treatment Engineering Limited Corporation. In the MARS model, the dataset was divided into two subsets: a training set with 160 samples for developing the MARS model and a testing set with 55 samples for verifying the developed MARS model. The training set included 88 sets with DTC of 1 m^3/d, 9 sets with DTC of 2 m^3/d, 21 sets with DTC of 5 m^3/d, 23 sets with DTC of 10 m^3/d, 7 sets with DTC of 15 m^3/d, 5 sets with DTC of 20 m^3/d, 2 sets with DTC of 45 m^3/d, 2 sets with DTC of 50 m^3/d, 2 sets with DTC of 60 m^3/d, and 1 set with DTC of 110 m^3/d. The other data were selected for the testing set.

The data were normalized between 0 and 1 by Equation (2) as follows:

$$d_{\mathrm{norm}} = \frac{d - d_{\mathrm{min}}}{d_{\mathrm{max}} - d_{\mathrm{min}}} \tag{2}$$

where d_{norm} is the normalized value of the dataset, d is the input/output variable, d_{min} is the minimum value of the dataset, and d_{max} is the maximum value of the dataset. In the following discussion, if there is no special explanation, all the variables refer to the variables being normalized.

2.2. Multivariate Adaptive Regression Spline (MARS)

Multivariate adaptive regression spline (MARS) was introduced by Friedman, which is a nonparametric regression modeling procedure that can approximate the relationship between a dependent variable (y) and a set of independent variables (x_1, x_2, ..., x_n) with a piecewise regression [16,19,23,24]. Functions fitted in piecewise regression are called basis functions (BFs) of the MARS methods. BFs can be either single spline function or a product of two or more spline

functions for different explanatory variables [19,20,24–30]. The form of MARS is expressed based on multivariate spline basis functions as follows:

$$Y = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(X) = \beta_0 + \sum_{m=1}^{M} \beta_m \prod_{i=1}^{Km} \max\{S_{im}(x_{j(i,m)} - t_{im}), 0\} \quad (3)$$

where Y represents the predicted value of the response; β_0 is the constant; β_m is the coefficient of the mth term of the basis function $B_m(X)$; M is the number of basis functions; $S_{im} = \pm 1$; $x_{j(i,m)}$ is the explanatory variables associated with the basis function $B_m(x_1, x_2, \cdots, x_n)$, i.e., the values of the jth explanatory variables at the ith node of the mth basic function; K_m is the level of interaction between $j(i,m)$ variables; and t_{im} indicates the node locations for $B_m(x_1, x_2, \cdots, x_n)$, which are the interface points between pieces, called knots in the MARS model. In this paper, $X = (x_1, x_2, x_3)$ and $Y = (y, y_1, y_2)$.

The definition of each BF is selected from the collection C where:

$$C = \{\{\max\{(x_j - t), 0\}, \max\{(t - x_j), 0\}\}_{t,j} : t \in \{x_{j1}, \cdots, x_{jn}\}, j = 1, \cdots, k\}. \quad (4)$$

Each basis function is piecewise linear with a knot t at x_{ji}, which can be multiplied together to form non-linear functions. The location and number of the needed spline basis functions were found through a second-order forward/backward stepwise regression procedure. For example, a two-sided basis function with knot t of 0.5 is shown in Figure 1.

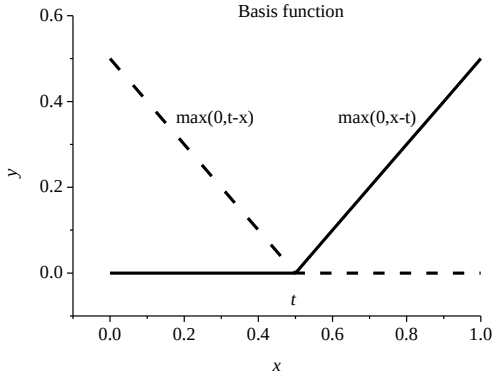


Figure 1. A graphical representation of a spline basis function: The left spline (max (0, $t - x$)) is shown as a dashed line and the right spline (max (0, $x - t$)) is shown as a solid line.

The basis functions are generated through two steps, which are the forward phase and backward pruning phase, detailed as follows.

2.3. Step 1 (Forward Phase)

In the forward stage, MARS becomes larger by considering a great number of basis functions and all possible variables among the predictor variables. In this phase, potential knots are continuously found to be added into basis functions to improve the performance until the model reaches a predetermined allowable maximum number of basis functions. Consequently, an over-fit model is generated as follows:

$$y' = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(x) \quad (5)$$

where y' is the predicted value for the response variable.

The regression coefficients $\beta_m, m = 0, 1, \cdots, M$ are estimated using the MARS method to obtain the center of the dependent variable.

2.4.Step 2 (Backward Pruning Phase)

In this phase, the basis function with the least contribution to the model performance was deleted one by one, leading to a simplified and generalized MARS model. Generalized cross-validation (GCV) criterion is used to assess the importance of variables, which can be expressed as follows:

$$\mathrm{GCV}(M) = \frac{\frac{1}{N}\sum_{i=1}^{N}[y_i - f(x_i)]^2}{\left[1 - \frac{C(M)}{N}\right]^2} \tag{6}$$

in which N is the number of observations, and $f(x_i)$ is the predicted values of the MARS model. $C(M)$ is a complexity penalty that increases with number of basis functions in the model, defined as:

$$C(M) = (M+1) + d \times M \tag{7}$$

where M is the number of basis functions, and d is the penalizing parameter. With the rise of the d value, fewer knots are obtained and function estimation becomes smoother. The optimal value of d is among 2 to 4 [16,29]. In this study, a default value of 3 is assigned to the penalizing parameter.

The importance of the variable can be obtained by assessing the decrease in the GCV values when the variable is removed from the model. The most important variable with the highest decrease in the GCV values is assigned a score of 100. The scores of the other variables are obtained according to the ratio of the decrease in the GCV values by these variables to that of the most important variable.

The effect of the input variables on the output variables can be explained well using analysis of variance (ANOVA) decomposition of the calculation results. In this paper, the relative importance of the input variabls x_1, x_2, and x_3 to the output variables y, y_1, and y_2 can be identified using ANOVA decomposition. The ANOVA decomposition of the developed MARS model is given by Equation (8) as follows:

$$f(x) = \beta_0 + \sum_{B=1} f_i(x_i) + \sum_{B=2} f_{ij}(x_i, x_j) + \sum_{B=3} f_{ijk}(x_i, x_j, x_k) \tag{8}$$

where $\sum_{B=1} f_i(x_i)$ is the overall basis function involving only a single variable, $\sum_{B=2} f_{ij}(x_i, x_j)$ is the overall basis function involving exactly two variables, and $\sum_{B=3} f_{ijk}(x_i, x_j, x_k)$ is the overall basis function involving three variables. The MARS model is created using the Software Salford Predictive Modeler 8.2, Salford Systems Company, San Diego, U.S.

3. Results and Discussion

3.1. Choose the Maximum Basis Function Number and Order Number

The maximum basis function number determined the performance of the MARS model in the forward step. The effects of numbers of maximum basis functions on the training performance are significantly different for model order in the MARS model, shown in Figure 2. Model order refers to the maximum interaction number of the basis functions.

The training performance was assessed in terms of coefficient R. The value of R was determined using Equation (9) as follows:

$$R = \frac{\sum_{i=1}^{n}(y_i - y_{i0})(y'_i - y'_{i0})}{\sqrt{\sum_{i=1}^{n}(y_i - y_{i0})^2}\sqrt{\sum_{i=1}^{n}(y'_i - y'_{i0})^2}} \tag{9}$$

where y_i and y_i' are the actual and predicted y values, respectively, and y_{i0} and y_{i0}' are the means of actual and predicted y values corresponding to n patterns. In this study, the value of n was 160 and 55 for the training and testing datasets, respectively.

It can be shown from Figure 2 that for the first order curve, the training performance increased with the maximum basis function number initially, and then remained invariable when the maximum basis function number was greater than 5, which means that the maximum basis function number can be chosen to be greater than 5. For the second order curve and the third order curve, we found that the two curves kept the same variation trend when the maximum basis function number was lower than 20. When the maximum basis function number was greater than 20, the third order training performance curve was better than that of the second order. However, the advantage was not apparent. The coefficient R was 0.981 with a maximum basis function number of 20 for the second order, which was close to 1. Consequently, to simplify the MARS model, model order was set to be 2, and the maximum basis function number was set to be 20. The input and output datasets were normalized using Equation (2).

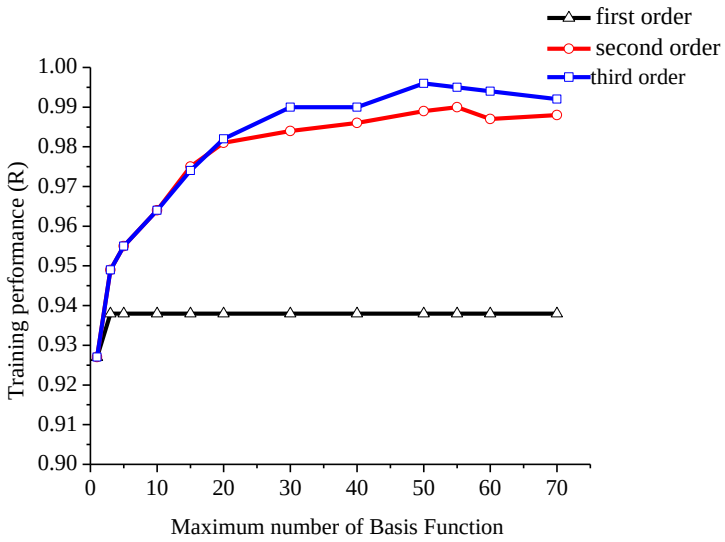


Figure 2. Effect of maximum number of basis functions as well as order on model performance R based on the training set of y.

3.2. Basis Functions and ANOVA Decomposition

3.2.1. Construction Cost (CC)

Through the forward phase and backward pruning phase, six basis functions were used to reach the minimum GCV value, which can represent the construction cost (CC) with the best solution. The MARS expression of CC (y_1) was given according to Equation (10) and Table 2:

$$y_1 = 0.071 + \sum_{m=1}^{6} \beta_m B_m(x) \tag{10}$$

Table 2. Basis functions of y_1 and corresponding coefficients.

$B_m(x)$	Equations	β_m
$B_1(x)$	$\max(0, x_1 - 0.037)$	1.19
$B_2(x)$	$\max(0, 0.037 - x_1)$	−1.923
$B_3(x)$	$B_1(x) \times \max(0, x_2 - 0.746)$	−15.299
$B_4(x)$	$B_1(x) \times \max(0, x_2 - 0.716)$	10.63
$B_5(x)$	$\max(0, x_1 - 0.083)$	−0.937
$B_6(x)$	$B_5(x) \times \max(0, x_3 - 0)$	0.802

The details of the basis functions in the MARS model for construction cost (y_1) are shown in Table 2. The effects of variables x_1, x_2, and x_3 on y_1 were determined using the slopes and intervals of basis functions.

(1) In Table 2, the first column $B_m(x)$ (m = 1, 2, . . . , 6) refers to the basis functions in the MARS model, the second column describes the equation form for $B_m(x)$ (m = 1, 2, . . . , 6), and the third column is the coefficient for $B_m(x)$ (m = 1, 2, . . . , 6). For example, for $B_1(x)$ in Table 2, if $(x_1 - 0.037)$ is greater than 0, i.e., DTC greater than 5 m^3/d, then the value of $B_1(x)$ is equal to $(x_1 - 0.037)$; and $B_1(x)$ is equal to 0 if $(x_1 - 0.037)$ is less than or equal to 0. A positive estimated coefficient β_m for the basis function indicated an increased construction cost, and a negative estimated coefficient β_m indicated a decreased construction cost. From this information, the effect of x_1 on y_1 had three impacts. When x_1 was less than 0.037 (DTC less than 5 m^3/d), then y_1 (CC) had no relationship with either x_2 (RCOD) or x_3 (RNH_3-N), and increased by 1.923 for each 1% increase in x_1.

(2) When x_1 was greater than 0.037 and less than 0.083 (DTC greater than 5 m^3/d and less than 10 m^3/d), then y_1 (CC) depended on both x_1 and x_2 (RCOD) with no relationship with x_3 (RNH_3-N).

   (i) When x_2 was less than 0.716 (RCOD less than 69.5%), then y_1 (CC) depended on x_1 without relationship with x_2, and increased by 1.19 for each 1% increase in x_1.

   (ii) When x_2 was greater than 0.716 and less than 0.746 (RCOD greater than 69.5% and less than 72.0%), then y_1 (CC) had a relationship with both x_1 and x_2, and increased by **1.19** to 1.509 for each 1% increase in x_1 corresponding to x_2 values of 0.716 and 0.746, respectively. The slope of y_1 increased with the increase of x_2.

   (iii) When x_2 was greater than 0.746 (RCOD greater than 72.0%), then y_1 (CC) increased by 0.323 to 1.507 for each 1% increase in x_1 corresponding to x_2 values of 1 and 0.746, respectively. The slope of y_1 decreased with the increase of x_2.

In summary, when x_1 was greater than 0.037 and less than 0.083 (DTC greater than 5 m^3/d and less than 10 m^3/d), the maximum slope of CC on DTC was a constant of 1.509.

(3) When x_1 was greater than 0.083 (DTC more than 10 m^3/d), then y_1 (CC) also depended on x_1, x_2 (RCOD), and x_3 (RNH_3-N) together, which are described in detail as follows:

   (i) When x_2 was less than 0.716 (RCOD less than 69.5%), then y_1 (CC) had a relationship with both x_1 and x_3 without consideration of x_2, and increased by 0.253 (x_3 = 0) to 1.055 (x_3 = 1) for each 1% increase in x_1. The effect of x_1 on y_1 increased with the increase of x_3.

   (ii) When x_2 was greater than 0.716 and less than 0.746 (RCOD greater than 69.5% and less than 72%), then y_1 (CC) is related to x_1, x_2 and x_3 together. With an increase of x_2 and x_3, the slope of y_1 on x_1 increased accordingly, and increased by 0.253 (corresponding to x_2 = 0.716 and x_3 = 0) to 1.374 (corresponding to x_2 =0.746 and x_3 = 1.0) for each 1% increase in x_1.

   (iii) When x_2 was greater than 0.746 (RCOD greater than 72%), then y_1 (CC) was also related to x_1, x_2, and x_3 together. However, the slope of y_1 on x_1 increased with the increase of x_3 and the decrease of x_2, and increased by -0.614 (corresponding to x_2 = 1 and x_3 = 0) to 1.374 (corresponding to x_2 = 0.746 and x_3 = 1.0) for each 1% increase in x_1.

In summary, when x_1 is greater than 0.083 (DTC greater than 10 m^3/d), the maximum slope of CC on DTC was 1.374.

Therefore, conclusions can be drawn as follows:

(1) The construction cost (CC) increased with design treatment capacity (DTC), and the maximum slope of CC on DTC decreased gradually from 1.923 to 1.374 in accordance with x_1 from 0 to 1.0. The variation of slope was also determined by x_2 and x_3.

(2) When x_1 was less than 0.037 (DTC less than 5 m^3/d), the slope of y_1 kept a constant of 1.923, and had no relation with neither x_2 (RCOD) nor x_3 (RNH_3-N). The result indicated that when DTC was less than 5 m^3/d, the relationship between CC and DTC was linear with a coefficient of 1.923.

(3) When x_1 was greater than 0.037 and less than 0.083 (DTC ranged from 5 m^3/d to 10 m^3/d), the slope of y_1 had a relationship with x_1 (DTC) and x_2 (RCOD) without any consideration of x_3 (RNH_3-N).

(4) When x_1 was greater than 0.083 (DTC greater than 10 m^3/d) and x_2 was less than 0.716 (RCOD less than 69.5%), and the slope of y_1 had a relationship with DTC (x_1) and RNH_3-N (x_3) without any consideration of RCOD (x_2).

The ANOVA decomposition of the MARS model aims to put together the basis functions with the same input variables. The ANOVA decomposition of y_1(CC) is shown in Table 3, from which it is clear that the variable DTC had the maximum effect on y_1, which has a maximum value of GCV, indicating the importance of the corresponding ANOVA function.

Table 3. Results of ANOVA decomposition in construction cost.

Function	Standard Deviation	GCV	Basis	Variable
1	0.068	0.0034	3	DTC
2	0.019	0.0009	2	DTC, RCOD
3	0.057	0.0007	1	DTC, RNH_3-N

In Table 3, the ANOVA function number is listed in the first column; the second column provides the standard deviation of this function, which gives an indication of its relative importance to the overall model and can be interpreted in a manner similar to the standardized regression coefficient in a linear model; the third column also gives an indication of the importance of the corresponding ANOVA function by listing the GCV score for a model with all BFs corresponding to that particular ANOVA function removed; the fourth column gives the number of BFs comprising the ANOVA function; and the last column gives the particular input variables associated with the ANOVA function.

3.2.2. Operation and Maintenance Cost

The MARS expression of operation and maintenance cost is given by Equation (11). Only two basis functions were obtained by the forward phase and backward pruning phase of the MARS model to get the minimum GCV value.

$$\begin{aligned} y_2 &= 0.044 + \sum_{m=1}^{2} \beta_m B_m(x) \\ &= 0.044 - 1.829 \times \max(0, x_1 - 0.174) + 4.221 \times \max(0, x_1 - 0.174) \times \max(0, x_3 - 0) \end{aligned} \quad (11)$$

It can be shown that the effect of x_1 has two impacts:

(1) When x_1 was less than 0.174 (DTC less than 20 m^3/d), then y_2 was a constant of 0.044, i.e., OMC was a constant of 435 RMB/year.

(2) When x_1 was greater than 0.174 (DTC greater than 20 m^3/d), then slope of y_2 increases from -1.829 to 2.392 (i.e., -1.829 + 4.221) corresponding to an x_3 value of 0 (RNH_3-N of 3.42%) and 1.0 (RNH_3-N of 91.89%). When the value of x_3 increases from 0 to 0.13 (RNH_3-N increased from 3.42% to 14.9%), the slope of y_2 increased from -1.829 to 0, and the value of y_2 decreased with the increase of x_1 due to negative slopes. When the value of x_3 increased from 0.13 to 1.0 (RNH_3-N increased from 14.9% to 91.89%), the slope of y_2 increased from 0 to 2.392, and the value of y_2 increased with the increase of x_1 due to positive slope values.

Therefore, the conclusion about operation and maintenance cost (OMC) can be drawn as follows:

(1) When DTC was less than 20 m^3/d, y_2 was a constant of 0.444 (OMC is 435 RMB/year) without relationship with the value of DTC. When DTC was greater than 20 m^3/d, and x_3 was less than 0.13 (RNH_3-N less than 14.9%), y_2 decreased with an increase of x_1 due to the negative slope of y_2. In contrast, when DTC was greater than 20 m^3/d, and x_3 was greater than 0.13 (RNH_3-N greater than 14.9%), y_2 increased with the increase of x_1 due to a positive slope of y_2.

(2) The value of y_2 had no relationship with x_2 (RCOD).

The ANOVA decomposition of y_2 (OMC) showed that only variable DTC had an effect on the MARS model of y_2 (OMC) separately with a standard deviation of 0.095 and GCV of 0.015.

3.2.3. Total Cost

The MARS expression of total cost is given by Equation (12). The details of basis functions in the MARS model for total cost y are shown in Table 4.

Table 4. Basis functions of y and corresponding coefficients.

$B_m(x)$	Equation	β_m
$B_1(x)$	$\max(0, x_1 - 0.037)$	1.336
$B_2(x)$	$\max(0, 0.037 - x_1)$	−1.809
$B_3(x)$	$B_1(x) \times \max(0, x_2 - 0.818)$	−64.45
$B_4(x)$	$B_1(x) \times \max(0, x_3 - 0.709)$	74.024
$B_5(x)$	$B_1(x) \times \max(0, x_2 - 0.8)$	43.353
$B_6(x)$	$B_1(x) \times \max(0, x_2 - 0.844)$	19.429
$B_7(x)$	$B_1(x) \times \max(0, x_3 - 0.703)$	−78.757
$B_8(x)$	$\max(0, x_1 - 0.083)$	−0.828
$B_9(x)$	$B_8(x) \times \max(0, x_3 - 0.648)$	9.071

$$y = 0.067 + \sum_{m=1}^{9} \beta_m B_m(x) \quad (12)$$

The MARS model of TC (y) combined the models of y_1 (CC) and y_2 (OMC) with consideration of all the variables.

The ANOVA decomposition of y (TC) is given in Table 5. Similar to y_1 (CC), the effect of DTC on y (TC) was the most significant in all three variables with the maximum value of GCV.

Table 5. Results of ANOVA decomposition for total cost.

Function	Standard Deviation	GCV	Basis	Variable
1	0.1	0.0069	3	DTC
2	0.034	0.0012	3	DTC, RCOD
3	0.026	0.0008	3	DTC, RNH_3-N

The effects of x_1 (DTC) on y (TC) are three-fold (shown in Figure 3a,b):

(1) When x_1 was less than 0.037 (DTC less than 5 m^3/d), then y (TC) increased by 1.809 for each 1% increase in x_1, and variables x_2 (RCOD) and x_3 (RNH_3-N) had no effects on slope of y, which can also be seen in Figure 3a,b.

(2) When x_1 was less than 0.083 and greater than 0.037 (DTC was less than 10 m^3/d and greater than 5 m^3/d), then y (TC) depended on x_1 (DTC), x_2 (RCOD), and x_3 (RNH_3-N) together, which is described in detail as follows:

    (i) When x_2 was less than 0.8 (RCOD less than 76.6%), then y (TC) depended on both x_1 and x_3 together without consideration of x_2 (shown in Figure 3a). The slope of y was a constant of 1.336 when x_3 was less than 0.703. When x_3 ranged from 0.703 to 0.709 (i.e., RNH_3-N

from 65.6% to 66.1%), the slope of y decreased from 1.336 to 0.863 accordingly. When x_3 ranged from 0.709 to 1.0 (RNH_3-N from 66.1% to 91.89%), the slope of y decreased from 0.863 to −0.514 accordingly.

(ii) When x_3 was less than 0.703 (RNH_3-N less than 65.6%), then y (TC) depended on both x_1 and x_2 together without consideration of x_3 (shown in Figure 3b). When x_2 ranged from 0.8 to 0.818 (RCOD from 76.6% to 78.1%), the slope of y increased from 1.336 to 2.117 accordingly. When x_2 ranged from 0.818 to 0.844 (RCOD from 78.1% to 80.3%), the slope of y decreased from 2.117 to 1.568 accordingly. When x_2 ranged from 0.844 to 1.0 (RCOD from 80.3% to 93.47%), the slope of y decreased from 1.568 to 1.308 accordingly.

(iii) When x_2 was greater than 0.8 and x_3 was greater than 0.703, then the effect x_1 on y was connected with the effect of both x_2 and x_3.

(3) When x_1 is greater than 0.083 (DTC greater than 10 m^3/d), then y (TC) depended on x_1, x_2, and x_3 together, which is described as follows:

(i) When x_2 was less than 0.8, then y (TC) depended on both x_1 and x_3 together without consideration of x_2 (shown in Figure 3a. When x_3 was less than 0.648 (RNH_3-N less than 60.7%), the slope of y was a constant of 0.508. When x_3 ranged from 0.648 to 0.703 (RNH_3-N from 60.7% to 65.6%), the slope of y increased from 0.508 to 1.007 accordingly. When x_3 ranged from 0.703 to 0.709 (RNH_3-N from 65.6% to 66.1%), the slope of y decreased from 1.007 to 0.588 accordingly. When x_3 ranged from 0.709 to 1.0 (RNH_3-N from 66.1% to 91.89%), the slope of y increased from 0.588 to 1.851 accordingly.

(ii) When x_3 was less than 0.648 (RNH_3-N less than 60.7%), then y (TC) depended on x_1 and x_2 together without consideration of x_3 (shown in Figure 3b). When x_2 ranged from 0.8 to 0.818 (RCOD from 76.6% to 78.1%), the slope of y increased from 0.508 to 1.288 accordingly. When x_2 ranged from 0.818 to 0.844 (RCOD from 78.1% to 80.3%), the slope of y decreased from 1.288 to 0.74 accordingly. When x_2 ranged from 0.844 to 1.0 (RCOD from 80.3% to 93.47%), the slope of y decreased from 0.74 to 0.48 accordingly.

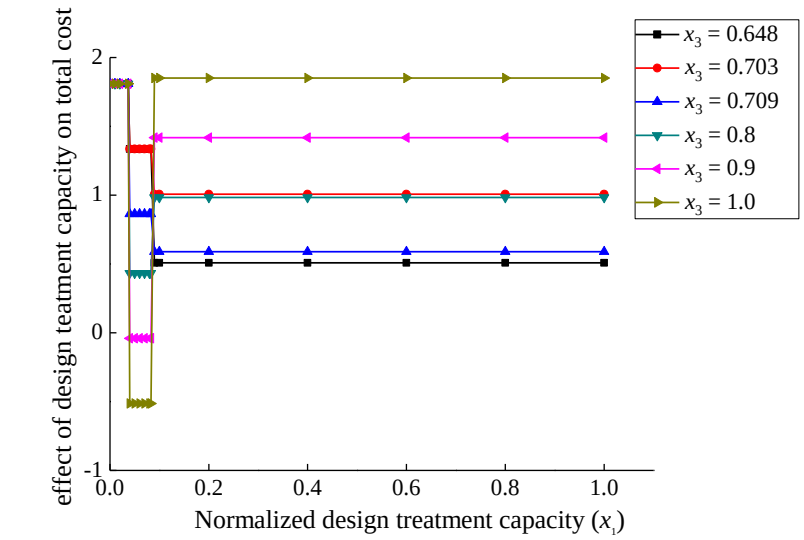
(iii) When x_2 was greater than 0.8 and x_3 was greater than 0.648, then the effect of x_1 on y was connected with the effect of both x_2 and x_3.

Therefore, the conclusions for TC can be drawn as follows:

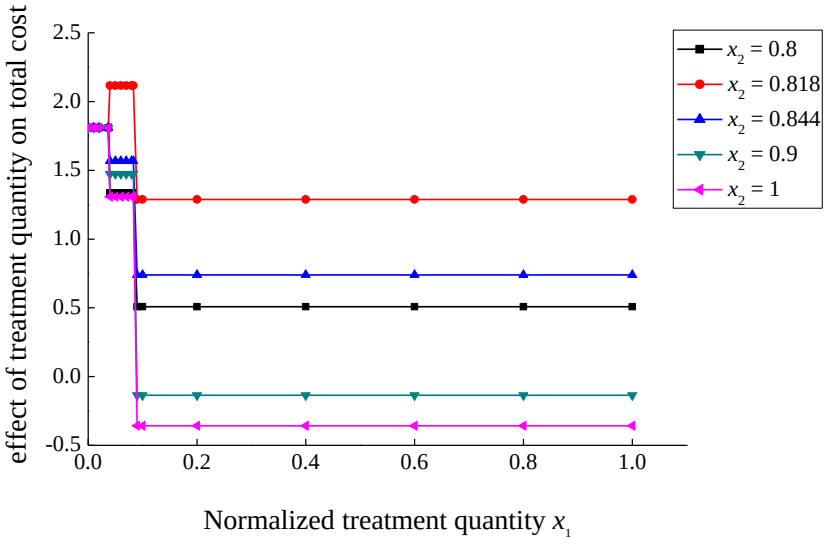
(1) When x_1 was less than 0.037 (DTC less than 5 m^3/d), the slope of y (TC) was a constant of 1.809, and had no relation with neither x_2 (RCOD) nor x_3 (RNH_3-N), which was similar to the slope of y_1 (CC).

(2) When x_1 was greater than 0.037 and less than 0.083 (DTC range from 5 m^3/d to10 m^3/d), the slope of y (TC) had a relationship with both x_2 (RCOD) and x_3 (RNH_3-N). When x_2 was less than 0.8 (RCOD less than 76.6%), the slope of y had a relationship with x_3 without consideration of x_2. When x_3 was less than 0.703 (RNH_3-N less than 65.6%), the slope of y had a relationship with x_2 without consideration of x_3. In addition, when x_2 was less than 0.8 (RCOD less than 76.6%) and x_3 was less than 0.703 (RNH_3-N less than 65.6%), the slope of y had no relationship with either x_2 or x_3.

(3) When x_1 was greater than 0.083 (DTC greater than 10 m^3/d), the slope of y had a relationship with both x_2 (RCOD) and x_3 (RNH_3-N). When x_2 was less than 0.8 (RCOD less than 76.6%), the slope of y had a relationship with x_3 without consideration of x_2. When x_3 was less than 0.648 (RNH_3-N less than 60.7%), the slope of y had a relationship with x_2 without consideration of x_3. In addition, when x_2 was less than 0.8 and x_3 was less than 0.648 (RCOD less than 76.6% and RNH_3-N less than 60.7%), the slope of y had no relationship with either x_2 or x_3.



(**a**)



(**b**)

Figure 3. (**a**) Effect of design treatment capacity x_1 on total cost under different x_3 without consideration of $x_2 < 0.8$. (**b**) Effect of design treatment capacity x_1 on total cost under different x_2 without consideration of x_3 ($x_3 < 0.648$).

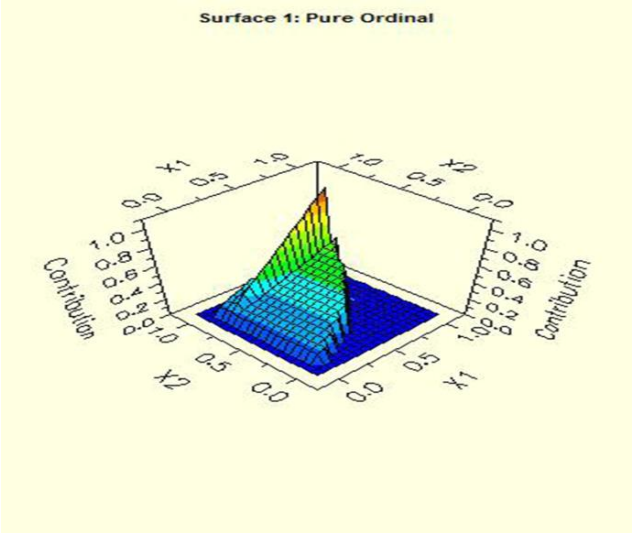
The relative importance of variables and the relationship among variables are shown in Table 6, from which we can find that DTC was the most important variable in determining the total cost of sewage treatment facilities, which was the same as the results of ANOVA decomposition.

Table 6. Relative importance of variables and their relationship.

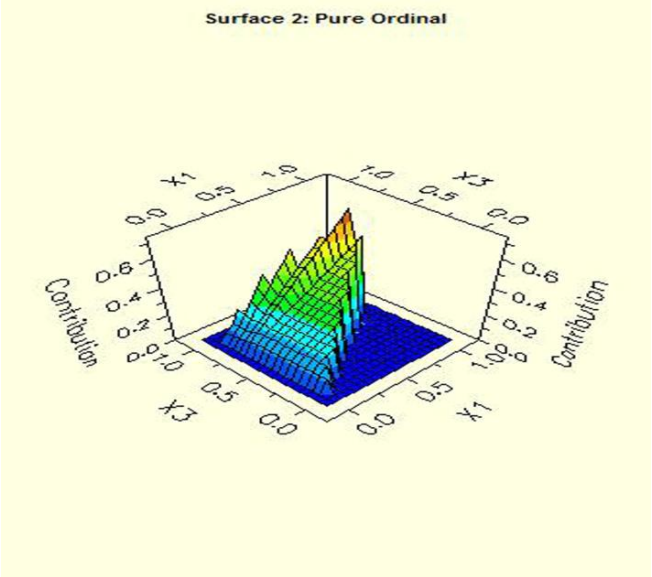
Variable	x_1	x_2	x_3	y	Relative Importance C_m
x_1	1				100
x_2	0.221	1			16.55
x_3	−0.02	0.47	1		9.75
y	0.963	0.248	−0.029	1	

From Table 6, we can also find that y (TC) had a significant relationship with x_1 (DTC), which means that TC increased with the development of DTC. Compared with x_1 (DTC), the relationship between y (TC) and x_2 (RCOD), as well as the relationship between y (TC) and x_3 (RNH_3-N), were lower. In addition, x_2 (RCOD) was more significant than x_3 (RNH_3-N).

The relationships among variables in the MARS model is shown in Figure 4a,b. The value of y (TC) rose with variables x_1 and x_2, and the contribution of x_1 was greater than x_2. A similar conclusion that the contribution of x_1 on TC (y) was greater than x_3 can be seen from Figure 4b. The results are consistent with the above analysis.



(**a**)



(**b**)

Figure 4. (**a**) Contribution to total cost y of the second order term of the variables design treatment capacity (x_1) and removal efficiency of chemical oxygen demand (x_2). (**b**) Contribution to total cost y of the second order term of the variables design treatment capacity (x_1) and removal efficiency of ammonia nitrogen (x_3).

The total cost of building wastewater treatment plants in rural regions is an important issue concerning regional sustainable development, which is very difficult to be accurately simulated due to its nonlinear characteristics. Consequently, the multivariate adaptive regression spline (MARS)

model is applied to predict the total cost of DWWTPs in this paper. The MARS model has its own advantages: (1) it does not require assumption of relationships between input and output variables, (2) automatically finds the best knots in basis functions, (3) can provide a more precise relationship between the response variable and predictor, and (4) does not require a long training process to reduce modeling time [29,30]. Therefore, the stepwise model obtained through MARS technology is a suitable method to predict total cost.

The model obtained through the MARS method was able to predict CC, OMC, and TC in DWWTPs. The comparisons of training dataset and testing dataset between real values and predicted values for CC, OMC, and TC are shown in Figure 5. In Figure 5, all the variables are in their original scale, not in the normalized scale.

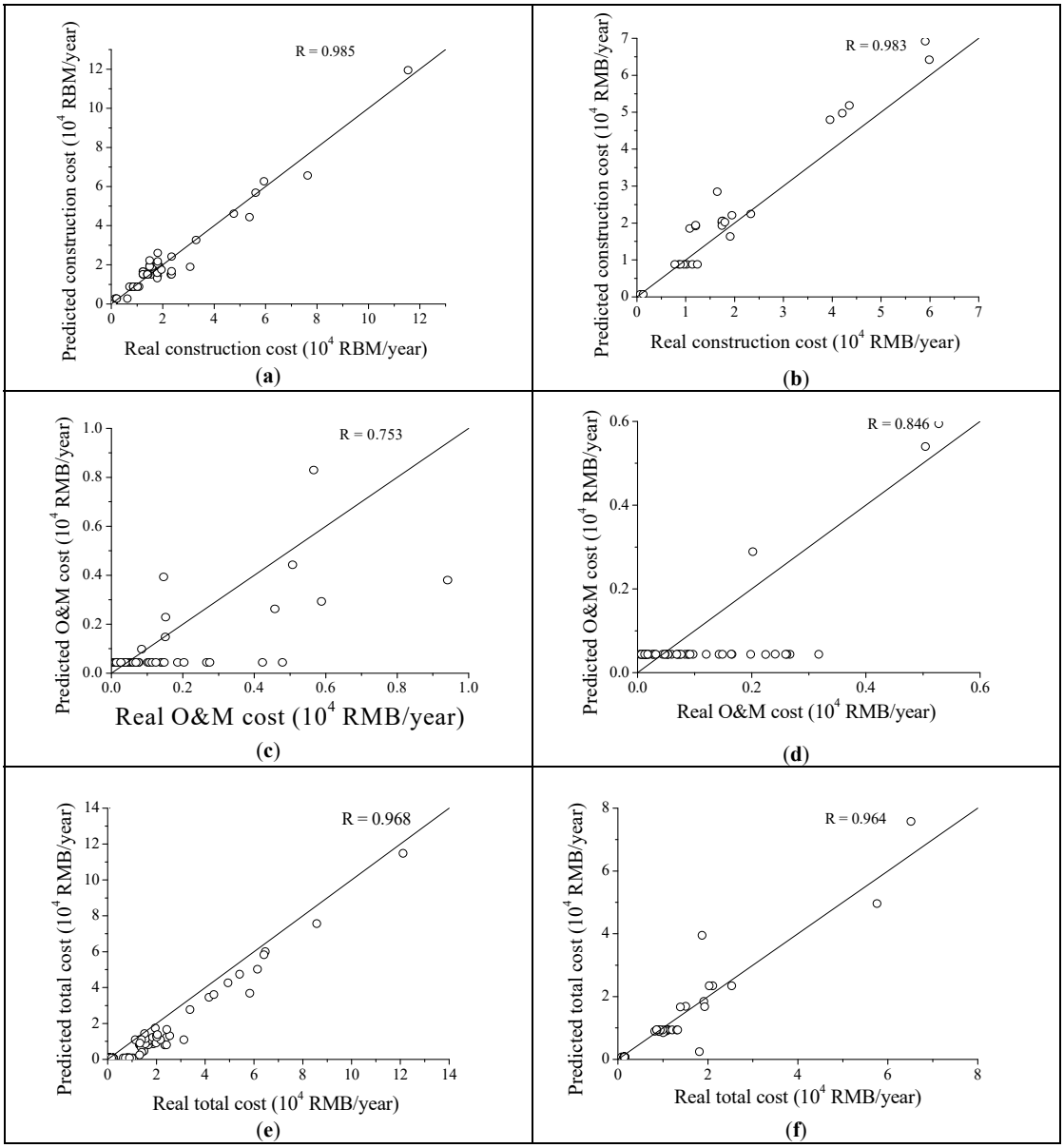


Figure 5. (**a**) Comparison of training dataset between real construction cost and predicted construction cost. (**b**) Comparison of testing dataset between real construction cost and predicted construction cost. (**c**) Comparison of training dataset between real OMC and predicted OMC. (**d**) Comparison of testing dataset between real OMC and predicted OMC. (**e**) Comparison of training dataset between real total cost and predicted total cost. (**f**) Comparison of testing dataset between real total cost and predicted total cost.

The comparisons of the training dataset and testing dataset between real and predicted values for CC are shown in Figure 5a,b, respectively. The results of CC obtained in the training dataset were better than the results obtained in the testing dataset, which can also be observed in Table 7. The value of R for the training dataset (0.985) was greater than the value of R for the testing dataset (0.983). The comparisons of the training dataset and testing dataset between real and predicted values for OMC are shown in Figure 5c,d, respectively. The results of OMC obtained in the testing dataset were better than the results obtained in the training dataset, which can also be observed in Table 7. The value of R for the testing dataset (0.846) is greater than the value of R for the training dataset (0.753). In addition, through the MARS method, the OMC model was obtained and expressed by Equation (11). When x_1 was less than 0.174 (DTC less than 20 m^3/d), OMC was a constant of 435 RMB/year. Since there were 148 samples among 160 samples in the training set with DTC less than 20 m^3/d, and there were 45 samples among 55 samples in the testing set with DTC less than 20 m^3/d, many data points in Figure 5c,d were flat. The comparisons of training and testing datasets between real and predicted total cost (TC) are shown in Figure 5e,f, respectively. Similar to CC, the results of TC obtained in the training dataset were better than the results obtained in the testing dataset, which can also be observed in Table 7. The value of R for the training testing dataset (0.968) was greater than the value of R for the testing dataset (0.964).

Table 7. Accuracy comparison among multivariate adaptive regression spline (MARS), support vector machine (SVM), and multiple linear regression (MLR) for the training and testing datasets.

Variables	Dataset	*R*			*RMSE*			*MAPE*		
		MARS	**SVM**	**MLR**	**MARS**	**SVM**	**MLR**	**MARS**	**SVM**	**MLR**
y_1	training	0.985	0.964	0.935	0.249	0.937	0.369	0.121	8.625	0.977
	testing	0.983	0.965	0.918	0.044	0.825	0.997	0.027	3.893	0.703
y_2	training	0.753	0.763	0.565	0.088	0.099	0.081	2.558	5.420	1.206
	testing	0.846	0.825	0.673	0.093	0.093	0.091	1.300	3.199	0.893
y	training	0.968	0.964	0.929	0.561	1.005	0.452	0.281	7.984	0.861
	testing	0.964	0.956	0.904	0.421	0.833	0.770	0.273	3.599	0.813

According the results of TC, shown in Figure 5e,f, the cost of treating 1 m^3 of sewage ranged from 147 RMB/year to 1512 RMB/year, with an average of 687 RMB/year.

3.3. Comparison with the Other Models

In this paper, the support vector machine (SVM) method and a multiple linear regression (MLR) model were applied to compare the results with a training set and a testing set (the same sets that were applied for validation of MARS model).

Besides the correlation coefficient of R, the accuracy performance of models were also assessed by root mean square errors (*RMSE*) and mean absolute percent error (*MAPE*), expressed by Equations (13) and (14), respectively, as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_i')^2}{n}} \quad (13)$$

$$MAPE = \frac{\sum_{i=1}^{n} |y_i - y_i'|/y_i}{n} \quad (14)$$

The comparisons of R, *RMSE*, and *MAPE* among the three methods of MARS model, SVM model, and MLR model were calculated using Equations (9), (13), and (14), respectively, and shown in Table 7.

As can be seen, R training data values of y_1 for MARS, SVM, and MLR were 0.985, 0.964, and 0.935, respectively; and R testing data values of y_1 for MARS, SVM, and MLR were 0.983, 0.965, and 0.918,

respectively. Results show that MARS was the fittest model of y_1 in terms of maximizing R values, with a value roughly 2% above SVM, and 5% above MLR. In terms of *RMSE* and *MAPE*, the MARS model was the lowest for both training and testing data. As for y_2, R training data values for MARS, SVM, and MLR were 0.753, 0.763, and 0.565, respectively; and R testing data values for MARS, SVM, and MLR were 0.846, 0.825, and 0.673, respectively. Results also indicated that the MARS model fitted y_2 better than MLR and almost the same as SVM (1% below SVM for training data). In addition, R training data values of y for MARS, SVM, and MLR were 0.968, 0.964, and 0.929, respectively; and R testing data values of y for MARS, SVM, and MLR were 0.964, 0.956, and 0.904, respectively. Results also indicated that MARS was the best model to fit y.

MARS model had better statistical results with higher R values than SVM model and the MLR method. It can also be found that the simulation results of y_2 for the three models were not as good as that for y_1 and y as seen by the relatively lower R values. In addition, although the R values of the SVM model were closer to that of the MARS model, the *RMSE* and *MAPE* values of the SVM model were greater than that of MARS model, especially for y_1 and y, which verified that the performance of the MARS model was better than the SVM model. The simulation accuracies of y, y_1, and y_2 for the training and testing datasets among the three models were in the order of MARS > SVM > MLR, except for y_2 for the training dataset, which indicated that MARS model was a more effective method to simulate the cost structure of DWWTPs than the SVM and MLR models. Moreover, the cost structure of DWWTPs had the characteristics of being stepwise and nonlinear.

4. Conclusions

In this paper, a MARS model is proposed for predicting the cost structure of DWWTPs. The model considers the effect of DTC, RCOD, and RNH_3-N on CC, OMC, and TC. The results obtained can be summarized as follows:

(1) The DTC was the most important parameter for predicting CC, OMC, and TC with a relative importance of 100, followed by RCOD and RNH_3-N with the relative parameters of 16.55 and 9.75, respectively.

(2) The slopes of CC and TC on DTC were related to DTC, RCOD and RNH_3-N, which is described in detail as follows:

   (a) When DTC was less than 5 m^3/d, the slopes of CC and TC on DTC were constants of 1.923 and 1.809 without consideration of RCOD and RNH_3-N. The constant slope means that that the relationship between CC or TC and DTC was linear. The positive and negative slopes indicated the increasing and decreasing trend, respectively. The result indicated that when DTC was less than 5 m^3/d, each of the various treatment technologies with differing RCOD and RNH_3-N can be chosen by planners from an economic point of view.

   (b) When DTC was greater than 5 m^3/d, RCOD and RNH_3-N affected the slopes of CC and TC of DWWTPs, which can help choose treatment technology:

      (i) The slopes of CC and TC on DTC had no relationship with RCOD when RCOD was less than 69.5% and 76.6%, respectively.

      (ii) When DTC was less than 10 m^3/d, the slope of CC on DTC had no relationship with RNH_3-N.

      (iii) When DTC was less than 10 m^3/d and RNH_3-N less than 60.7%, the slope of TC on DTC had no relationship with RNH_3-N.

   (c) When DTC was greater than 10 m^3/d and RNH_3-N less than 65.6%, the slope of TC on DTC had no relationship with RNH_3-N.

   (d) With the increase of DTC, the slope of CC on DTC decreased gradually from 1.923 to 1.374.

(3) The slopes of OMC on DTC were related to DTC and RNH_3-N described as follows:

(a) When DTC was less than 20 m^3/d, then OMC was a constant of 435 RMB/year, which means that 20 m^3/d was the threshold for constructing DWWTPs. It can be concluded that when the treatment scale is no more than 20 m^3/d, the OMC is the same at 435 RMB/year. The conclusion is meaningful and can help managers make budget between DTC and OMC.

(b) When DTC was greater than 20 m^3/d, then slope of OMC on DTC increased with RNH_3-N and had no relationship with RCOD.

The results obtained provide useful information to perform techno-economic analysis for planners to make decisions on treatment scale and treatment technology before construction. The developed MARS model combined the merits of a nonparametric model and traditional multiple linear regression with simplicity and good interpretation, which does not need to assume a statistical distribution of the data. The non-linear structure of the cost function captured the inherent relationship between variables, which can be expected to improve the accuracy of model. Compared with SVM and MLR models, the simulation results obtained by the MARS model were closer to the real costs. The results showed that the developed MARS model can be a valuable tool to predict CC, OMC, and TC of DWWTPs. The cost–benefit evaluation can be performed more scientifically by simulating the cost structure with the proposed MARS model. The proposed method can also be applied to other regions in China to determine CC, OMC, and TC of DWWTPs based on DTC, RCOD, and RNH_3-N, which can provide helpful and meaningful information for local governments to make reasonable and economic plans to protect the water environment, especially in rural regions.

Author Contributions: Writing-Review & Editing, Y.W.; Investigation, L.W.; Formal analysis, B.E.

Funding: The research was funded by the Special S&T Project on Treatment and Control of Water Pollution from Bureau of Housing and Urban-Rural Development of Changshu City (number as 2011ZX07301-003-05-04).

Acknowledgments: The data used were provided by Construction Bureau in Changshu City, and Suzhou Hongyu Wastewater Treatment Engineering Limited Corporation.

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

References

1. Hernandez-Sancho, F.; Molinos-Senante, M.; Sala-Garrido, R. Cost modelling for wastewater treatment processes. *Desalination* **2011**, *268*, 1–5. [CrossRef]
2. Zhou, Y.; Huang, G.; Zhu, H.; Li, Z.; Chen, J. A factorial dual-objective rural environmental management model. *J. Clean. Prod.* **2016**, *124*, 204–221. [CrossRef]
3. Mkwate, R.C.; Chidya, R.C.G.; Wanda, E.M.M. Assessment of drinking water quality and rural household water treatment in Balaka District, Malawi. *Phys. Chem. Earth* **2017**, *100*, 353–362. [CrossRef]
4. Chen, H.; Chang, N. A comparative analysis of methods to represent uncertainty in estamating the cost of constructing wastewater treatment plants. *J. Environ. Manag.* **2002**, *65*, 383–409. [CrossRef]
5. Engin, G.O.; Demir, I. Cost analysis of alternative methods for wastewater handling in small communities. *J. Environ. Manag.* **2006**, *79*, 357–363. [CrossRef] [PubMed]
6. Xin, X.; Huang, G.; Sun, W.; Zhou, Y.; Fan, Y. Factorial two-stage irrigation system optimization model. *J. Irrig. Drain. Eng.* **2016**, *142*, 04015056. [CrossRef]
7. Papadopoulos, B.; Tsagarakis, K.P.; Yannopoulos, A. Cost and land functions for wastewater treatment projects: Typical simple linear regression versus fuzzy linear regression. *J. Environ. Eng.* **2007**, *133*, 581–586. [CrossRef]
8. Friedler, E.; Pisanty, E. Effects of design flow and treatment level on construction and operation costs of municipal wastewater treatment plants and their implications on policy making. *Water Res.* **2006**, *40*, 3751–3758. [CrossRef] [PubMed]
9. Lamas, W.Q.; Silveira, J.L.; Giacaglia, G.E.O.; Reis, L.O.M. Development of a methodology for cost determination of wastewater treatment based on functional diagram. *Appl. Therm. Eng.* **2009**, *29*, 2061–2071. [CrossRef]

10. Khan, U.T.; Valeo, C. Comparing a Bayesian and fuzzy number approach to uncertainty quantification in short-term dissolved oxygen prediction. *J. Environ. Inform.* **2017**, *30*, 1–16. [CrossRef]
11. Molinos-Senante, M.; Hernández-Sancho, F.; Sala-Garrido, R. Cost-benefit analysis of water-reuse projects for environmental purposes: A case study for Spanish wastewater treatment plants. *J. Environ. Manag.* **2011**, *92*, 3091–3097. [CrossRef]
12. Chen, R.; Wang, X.C. Cost-benefit evaluation of a decentralized water system for wastewater reuse and environmental protection. *Water Sci. Technol.* **2009**, *59*, 1515–1522. [CrossRef] [PubMed]
13. Wang, Y.; Wu, L.; Feng, Y. Cost function for treating wastewater in rural regions. *Desalin. Water Treat.* **2015**, *57*, 17241–17246.
14. Naik, K.S.; Stenstrom, M.K. A feasibility analysis methodology for decentralized wastewater systems-energy-efficiency and cost. *Water Environ. Res.* **2016**, *88*, 201–209. [CrossRef] [PubMed]
15. Maurer, M.; Rothenberger, D.; Larsen, T.A. Decentralised wastewater treatment technologies from a national perspective: At what cost are they competitive? *Water Sci. Technol. Water Supply* **2006**, *5*, 145–154. [CrossRef]
16. Zhang, W.; Zhang, Y.; Goh, A.T.C. Multivariate adaptive regression splines for inverse analysis of soil and wall properties in braced excavation. *Tunn. Undergr. Space Technol.* **2017**, *64*, 24–33. [CrossRef]
17. Zarei, K.; Atabati, M.; Teymori, E. Multivariate adaptive regression splines for prediction of rate constants for radical degradation of aromatic pollutants in water. *J. Solut. Chem.* **2014**, *43*, 445–452. [CrossRef]
18. Samui, P.; Kim, D. Determination of the angle of shearing resistance of soils using multivariate adaptive regression spline. *Mar. Georesour. Geotechnol.* **2014**, *33*, 542–545. [CrossRef]
19. Nieto, P.J.G.; Fernández, J.R.A.; Lasheras, F.S.; Juez, F.J.D.; Muñiz, C.D. A new improved study of cyanotoxins presence from experimental cyanobacteria concentrations in the Trasona reservoir (Northern Spain) using the MARS technique. *Sci. Total Environ.* **2012**, *430*, 88–92. [CrossRef]
20. Menon, R.; Bhat, G.; Saade, G.R.; Spratt, H. Multivariate adaptive regression splines analysis to predict biomarkers of spontaneous preterm birth. *Acta Obstet. Gynecol. Scand.* **2014**, *93*, 382–391. [CrossRef] [PubMed]
21. Lee, T.S.; Chen, I.F. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Syst. Appl.* **2005**, *28*, 743–752. [CrossRef]
22. Molinos-Senante, M.; Garrido-Baserba, M.; Reif, R.; Hernández-Sancho, F.; Poch, M. Assessment of wastewater treatment plant design for small communities: Environmental and economic aspects. *Sci. Total Environ.* **2012**, *427–428*, 11–18. [CrossRef] [PubMed]
23. Chen, Y.; Fan, R.; Liu, Z.; Chen, P. Research progress of integrated rural domestic sewage treatment plant. *J. Anhui Agric. Sci.* **2016**, *44*, 84–88. (In Chinese)
24. Chang, L. Analysis of bilateral air passenger flows: A non-parametric multivariate adaptive regression spline approach. *J. Air Transp. Manag.* **2014**, *34*, 123–130. [CrossRef]
25. Friedman, J.H. Multivariate adaptive regression splines. *Ann. Stat.* **1991**, *19*, 1–67. [CrossRef]
26. Haghiabi, A.H. Prediction of river pipeline scour depth using multivariate adaptive regression splines. *J. Pipeline Syst. Eng.* **2017**, *8*, 04016015. [CrossRef]
27. Samui, P. Multivariate adaptive regression spline (Mars) for prediction of elastic modulus of jointed rock mass. *Geotech. Geol. Eng.* **2013**, *31*, 249–253. [CrossRef]
28. Cheng, M.; Cao, M. Estimating strength of rubberized concrete using evolutionary multivariate adaptive regression splines. *J. Civ. Eng. Manag.* **2016**, *22*, 711–720. [CrossRef]
29. Chang, L. Exploring contributory factors to highway accidents: A nonparametric multivariate adaptive regression spline approach. *J. Transp. Saf. Secur.* **2017**, *9*, 419–438. [CrossRef]
30. Fernández, J.R.A.; Nieto, P.J.G.; Muniz, C.D.; Antón, J.C.Á. Modeling eutrophication and risk prevention in a reservoir in the Northwest of Spain by using multivariate adaptive regression splines analysis. *Ecol. Eng.* **2014**, *68*, 80–89. [CrossRef]