WILEY | Hindawi

Research Article

Fitting Method of Optimal Energy-Running Time Curve Based on Train Operation Data of an Urban Rail Section

Lianbo Deng (ID)**, Hongda Mei** (ID)**, Wenliang Zhou** (ID)**, and Enwei Jing** (ID)

School of Traffic and Transportation Engineering, Rail Data Research and Application Key Laboratory of Hunan Province, Central South University, Changsha, Hunan 410075, China

Correspondence should be addressed to Lianbo Deng; lbdeng@csu.edu.cn

Due to the complexity of the operation control of urban rail transit and diversity requirements for section running time standards, based on actual train operation data, this paper proposes a curve fitting method to find the interrelation between running time and energy consumption. According to features of the energy consumption-running time curve, the discriminant criterion of outliers is constructed to select the candidate fitting data set from the original data set. To fit the energy consumption-running time curve from two-dimensional scatter points, we propose a B-spline curve fitting method based on a genetic algorithm and the fitting method is proven to have high fitting accuracy and convergence speed. Furthermore, we propose an optimization method for the fitting curve based on dynamic adjustment of the fitting data set which is selected from the candidate fitting data set to obtain the optimal energy-running time curve. The validation of Guangzhou Metro's actual operation data shows that the energy-running time curve fitted and optimized by our method has lower energy and better continuity and smoothness and could be used for evaluation of train drivers' performance and energy consumption of train operation diagram.

1. Introduction

Under the background of developing green traffic and building energy-saving cities, urban rail transit has become one of the main means of solving the problem of urban traffic congestion due to its low energy consumption and low pollution. In most Chinese cities, urban rail transit has developed rapidly in the last few years. By the end of 2019, about 40 cities in Mainland China had opened 208 urban rail transit lines, with a total length of 6,736.2 km of which 5,180.6 km was subway lines, accounting for 76.9%. China's urban rail transit has developed by leaps and bounds, but it also faces many problems such as high energy costs and low management level. Therefore, reducing the energy consumption level of urban rail transit will help to give play to the advantages of urban rail transit and maintain its competitiveness.

Energy consumption directly related to urban rail transit is mainly generated by stations and trains. Train energy consumption includes operation energy consumption and auxiliary energy consumption such as lighting, ventilation, and air conditioning. Operation energy consumption is the main source of energy consumption of urban rail trains, which generally accounts for more than 50% of the train energy consumption and takes a large proportion of the operating expenditure. The effective reduction of operation energy consumption not only meets the requirements of developing green traffic but is also an effective way for urban rail transit enterprises to reduce costs and improve benefits. Therefore, it is of great significance to study the traction energy consumption of urban rail transit.

Generally, the operation curve of urban rail trains is optimized to minimize the operation energy consumption to obtain the optimal running time and energy consumption considering the requirements of section running time. In the trial operation stage of the train, the manufacturer presets several operation curves for selection during formal operation according to some parameters such as the vehicle performance, the horizontal and vertical section of lines, the section length, and the user requirements of the urban rail enterprise. Nowadays, there has been much research on the energy-saving operation strategy of urban rail trains. As

early as 1980, Milroy [1] began to study the train operation energy consumption optimization problem. Based on the maximum principle, he proposed a short-distance energy-saving operation strategy for urban rail transit which included three stages: traction, coasting, and braking and established the optimal control model to minimize the energy consumption under a constant slope, laying the foundation for modern urban rail train optimal control theory. Khmelnitsky [2], on the basis of an analytical method, further considered the change in line slope, the difference of section speed limits, and the relationship between traction force and braking force with changes in speed and proposed a numerical algorithm for optimizing the operation strategy. Chang and Sim [3] established a multiobjective optimization control model considering comfort, punctuality, and energy consumption and used an improved genetic algorithm (GA) to calculate the switching point of coasting mode so as to achieve the effect of saving energy by reasonably increasing the coasting time. Ke et al. [4] optimized the section operation strategy of urban rail transit by using the maximum-minimum ant colony cloning optimization algorithm and proved that their proposed algorithm had a higher computational efficiency than other intelligent algorithms. Liu et al. [5] used the maximum principle to optimize the energy-saving operation strategy of the train and then used a numerical algorithm to solve the switching point of the train operation mode, also achieving good results. Villalba et al. [6] proposed an optimization model based on the train speed relationship and set a speed limit for trains traveling between stations to minimize energy consumption, achieving a 19% reduction in energy consumption compared to current levels. Bocharnikov et al. [7] designed a fitness function with variable weightings which was used to identify optimal train trajectories by running a series of simulations in parallel with a genetic algorithm search method and optimized traction energy consumption during a single-train journey by the optimal train trajectories.

However, the optimization of the operation curve is a complex optimization problem for which it is difficult to obtain the optimal solution. Meanwhile, operation curves preset by the manufacturer are limited and unable to cover all running time requirements of daily train operation. Therefore, aiming at the theoretical limitations of optimization of operation energy consumption at a certain running time, this paper attempts to study the relationship between section running time and reasonable energy consumption from a data-driven perspective by using the abundant train operation data formed in the operation process of urban rail enterprises.

In order to establish the relationship between section running time and optimal energy consumption, the data fitting method is very suitable. The fitting curve can visually show the changing trend of discrete data and has a wide range of applications in engineering practice. The selection of the fitting data set directly affects the accuracy and effect of the final fitting curve, so it is very important to choose a suitable fitting data set. M. Rza Mashinchi et al. [8] designed the granularity box regression method based on border regression to preprocess the data set containing outliers, eliminated outliers deviating from the fitting curve, and then conducted linear regression analysis on the data. Hossein Hassain et al. [9] demonstrated the importance of eliminating noise from the data set in plant growth curve fitting and proposed that using singular spectrum analysis to process data can effectively eliminate noise. Sanpeng Zheng et al. [10] improved the classical moving least squares method and could automatically identify outliers from the discrete data set to reduce the influence of outliers on the fitting curve through a weight function and to ensure the fitting effect. Ping Chen et al. [11] proposed a Gibbs sampling algorithm to detect additive outliers and patches of outliers in bilinear time series models based on the Bayesian view and demonstrated the efficacy of detection and estimation by Monte Carlo methods. Galvez et al. [12] applied the firefly algorithm, a powerful metaheuristic nature-inspired algorithm, to compute the approximating explicit B-spline curve for a given set of noisy data points. Trejo-Caballero et al. [13] proposed a linear combination of radial basis functions (RBFs) to tackle the curve fitting problem with a set of data points including noises.

Considering the influence of the passenger loading rates on the correlation between the running time and the energy consumption, we construct the data set of train operation based on a given loading-rate standard. The train operation data in this paper refer to the operation information of each train in the research section and operation direction, including the section operation curve, the section running time, and the corresponding energy consumption. Ignoring specific details such as the operation speed, acceleration, operation mode, and other parameters of the operation curve, we construct energy consumption-running time data points (E-T points) by taking the section running time as the abscissa and the energy consumption as the ordinate to study the change laws between running time and energy consumption and obtaining the optimal energy consumption-running time curve (the optimal E-T curve).

The optimal E-T curve shows the lowest operation energy consumption in different section running times. Meanwhile, the corresponding train operation curves of E-T points can provide abundant running curve support for the operation of the train under different running times. And due to the optimality of each point of the curve, the reasonableness of the operation strategy adopted by a train in the section can be evaluated accordingly by the comparison between the actual operation energy consumption and the optimal operation energy consumption. What is more, with the optimal E-T curve of each section in the train operation diagram, the optimal energy consumption of the entire operation diagram can also be calculated to evaluate the energy consumption level of the existing operation diagram and make up the optimal energy consumption timetable.

To obtain the optimal E-T curve, we first construct the discriminant criterion of the energy consumption level of the train operation data to eliminate the obvious unreasonable data with high energy consumption from the original train operation data set and obtain a candidate set of fitting data points after the preliminary screening to improve

the accuracy of curve fitting. Secondly, based on the candidate fitting data set, a B-spline fitting method is adopted. Finally, based on the feedback of the fitting results, we establish an optimization method to improve the quality of the fitting curve by the dynamic adjustment of the fitting data set.

2. Candidate Fitting Data Set

The section operation curve of urban rail trains generally includes several parts such as "maximum traction-cruising (or coasting)-maximum braking." However, minor changes in operating strategy and control parameters will cause changes in the curve, running time, and energy consumption, resulting in a diversity of section operation data of trains. Even the same running time may correspond to several different train operation curves, as shown in Figure 1.

As described above, the train operation strategy (the operation curve) of the train operation data and the running time under the strategy are defined as E-T points. Because of the diversity of train operation data in rail sections, E-T points are relatively unordered and scattered in the coordinate system. In order to study the relationship between section running time and energy consumption and to obtain the optimal E-T curve, it is necessary to construct the discriminant criterion of outliers of E-T points, remove unreasonable energy consumption data from the original data set, and filter out a better data set to ensure the fitting effect of the E-T curve.

Although there may be a one-to-many relationship between section running time and the train operation curve, each running time should have the unique optimal energy consumption [14, 15]. Figure 2 shows that optimal energy consumption and running time have a negative correlation and the uniqueness of the optimal energy consumption in a certain running time. Thus, the outliers of E-T points can be removed according to the interrelation of the train operation data.

Taking the four data points in Figure 3 as an example, according to the approximate inverse relation between optimal energy consumption and running time, the optimal energy consumption EB at running time TB must be less than EA because TB > TA, so point B should be removed. Similarly, compared with point C, point D needs to be removed.

Setting the original data set of E-T points as $P = \{p_i(T_i, E_i), i = 1, 2, 3, \}$, according to the above analysis, the optimal data set $P^* = \{p_i^*(T_i^*, E_i^*), \quad i = 1, 2, 3, \ldots\}$ should satisfy the following law: $\forall p_1^*(T_1^*, E_1^*), p_2^*(T_2^*, E_2^*) \in P^*$, if $T_2^* \geq T_1^*$, then $E_2^*/T_2^* \leq E_1^*/T_{21}^*$, $E_2^* \leq E_1^*$, and generally, $\forall p_1^*(T_1^*, E_1^*)$, $p_2^*(T_2^*, E_2^*) \in P^*$, if $T_2^* > T_1^*$, then $E_2^*/T_2^* < E_1^*/T_{21}^*$, $E_2^* < E_1^*$.

Based on the features of the optimal data set, the discriminant criterion of outliers of the original data set is constructed as follows:

$\forall p_1 T_1$, E_1, $p_2 T_2$, $E_2 \in P$, if $E_1 = E_2, T_1 > T_2$, then p_1 is worse than p_2. $\forall p_1(T_1, E_1), p_2(T_2, E_2) \in P$, if $T_1 = T_2, E_1 > E_2$, then p_1 is worse than p_2. $\forall p_1(T_1, E_1), p_2(T_2, E_2) \in P$, if $T_1 \geq T_2, E_1 > E_2$ or $(E_1/T_1) > (E_2/T_2)$, then p_1 is worse than p_2.
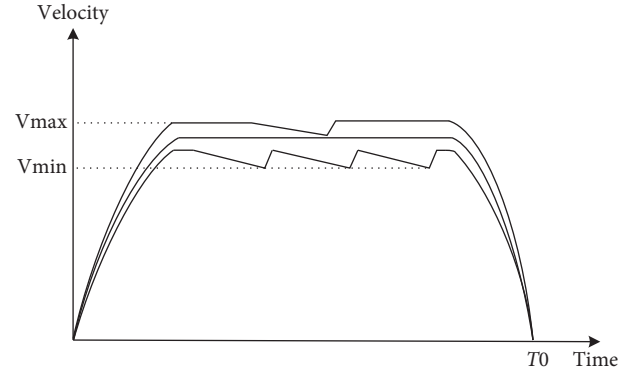

FIGURE 1: Multiple operation curves can correspond to the same running time.

The above rules can simply distinguish and eliminate some obvious bad points, but there may be some points of poor quality that cannot be eliminated. Therefore, we define the inferiority to measure the quality of these points, and the greater the inferiority, the worse the energy consumption of one point. $\forall p(T, E) \in P$, the neighboring point with less running time of point p is $p_1(T_1, E_1), T_1 < T$, and $\overrightarrow{p_1'}$ is the direction vector of the tangent at point p_1 on the E-T curve, so the inferiority of point p is defined as follows:

$$\delta(p) = \frac{\delta\left(\overrightarrow{p_1'}, \overrightarrow{p_1 p}\right)}{\delta\left(\overrightarrow{p_1'}, 0\right)}, \tag{1}$$

where $\delta(\overrightarrow{p_1'}, \overrightarrow{p_1 p})$ is the angle between vector $\overrightarrow{p_1'}$ and $\overrightarrow{p_1 p}$, $\delta(\overrightarrow{p_1'}, 0)$ is the angle between vector $\overrightarrow{p_1'}$ and the horizontal axis, and $\delta(p) \in (0, 1]$.

Based on the discriminant criterion of outliers, we can preliminarily remove outliers and points with unreasonable energy consumption from the original data set $P = \{p_i(T_i, E_i), i = 1, 2, 3, \ldots\}$ and obtain the candidate fitting data set $P^s = \{p_i^s(T_i^s, E_i^s), i = 1, 2, 3, \ldots\}$. The specific steps of the algorithm are as follows:

Step 1. Order all E-T points in the original data set in ascending order according to the running time. $|P|$ is the size of set P, and the initial number of elements in set P is Num.

Step 2. Remove outliers based on the discriminant criteria. Set $i = 0$, and compare the energy consumption of p_i and p_{i+1}, where $p_i, p_{i+1} \in P$. If $T_{i+1} = T_i, E_{i+1} > E_i$, or $T_{i+1} > T_i, E_{i+1} \geq E_i$, then $P = P - \{p_{i+1}\}, |P| = |P| - 1$. Set $i = i + 1$.

Step 3. If $i < \text{Num} - 1$, return to Step 2; otherwise, set $P^s = P$ d and take P^s as the candidate fitting data set.

In this algorithm, we obtain the candidate fitting data set by the discriminant criterion of outliers, and the definition of inferiority is used in the dynamic adjustment of the fitting data set later.
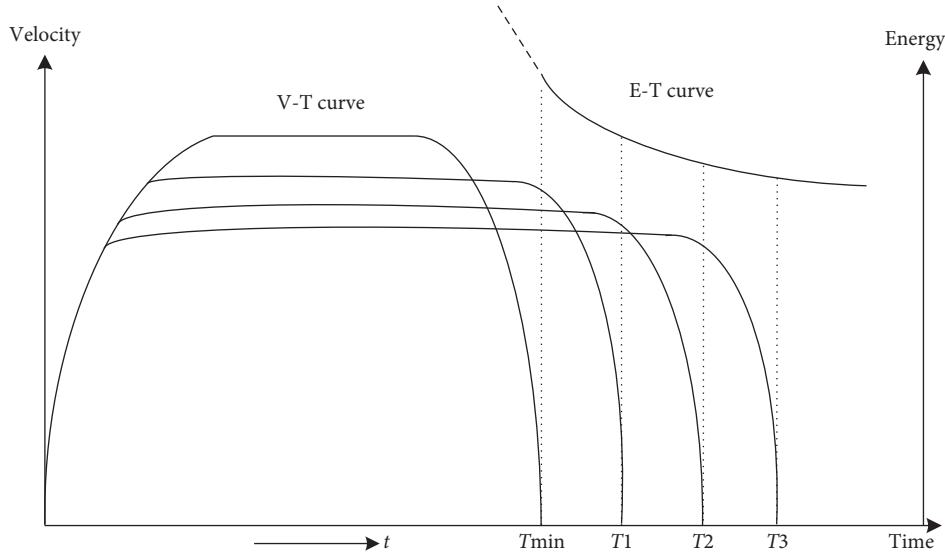
FIGURE 2: The interrelation between section running time and energy consumption.
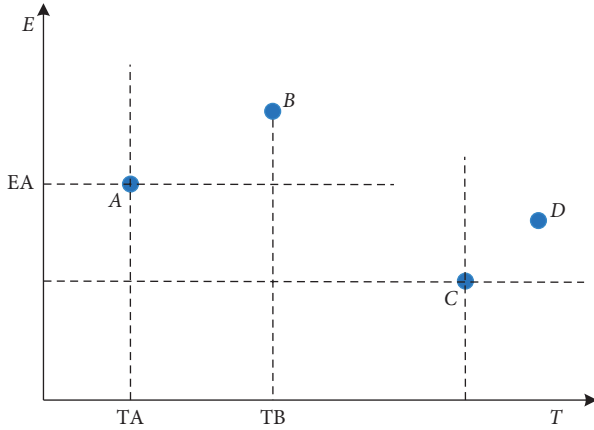

FIGURE 3: Removal of points with high energy consumption.

3. A B-Spline Curve Fitting Method Based on a Genetic Algorithm for the Optimal E-T Curve

For the selected train operation data set, the running time and energy consumption of the data constitute scattered points that could be fitted as an E-T curve. B-spline has the powerful function of expressing and designing free-form curves and surfaces and is one of the most popular mainstream methods for the mathematical description of shapes. So the B-spline curve can be used to fit a set of two-dimensional data points based on the train operation data [16, 17].

3.1. B-Spline Curve Fitting Method for the Selected Data Set. An ordered data set $Q = q_i\{(T_i, E_i), i = 1, \ldots, m\}$ is selected from the candidate fitting data set P^s $(Q \subseteq P^s)$, and the parameter vector of Q is $T = \{t_i, i = 1, \ldots, m\}$. The mathematical definition of the B-spline curve over the knot vector $U = \{u_0 = \cdots = u_k \leq u_{k+1} \leq \cdots \leq u_n \leq u_{n+1} = \cdots = u_{n+k+1}\}$ is shown as follows:

$$B(t) = \sum_{j=0}^{n} P_j N_{j,k}(t), \tag{2}$$

where k is the degree of curve, P_j $(j = 0, 1, \ldots, n)$ are control points, and $N_{j,k}(t)$ $(j = 0, 1, \ldots, n)$ are the B-spline basis functions. Basis functions are calculated using the following equations:

$$N_{j,0}(t) = \begin{cases} 1, & u_j \leq t < u_{j+1}, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

$$N_{j,k}(t) = \frac{t - u_j}{u_{j+k} - u_j} N_{j,k-1}(u) + \frac{u_{j+k+1} - u}{u_{j+k+1} - u_{j+1}} N_{j+1,k-1}(u). \tag{4}$$

If necessary, the convention $(0/0) = 0$ in equation (4) is applied. When the data set Q falls on the B-spline curve, $\forall q_i \in Q$ should be satisfied:

$$q_i = B(t_i) = \sum_{j=0}^{n} P_j N_{j,k}(t_i), \quad i = 1, \ldots, m, \tag{5}$$

which is written in matrix form as follows:

$$Q = NP, \tag{6}$$

where **Q** is the data set matrix and **N** is the B-spline basis function matrix that could be calculated by the parameter vector T and the knot vector U.

The B-spline curve should go through the start point and end point and then

$$\begin{aligned} Q_0 &= B(t_0) = P_0, \\ Q_1 &= B(t_1) = P_1. \end{aligned} \tag{7}$$

Aiming at the minimum square sum of error (*SSE*) between the fitting data points and actual data points, the objective function can be expressed as

$$f = \sum_{i=1}^{m-1} [Q_i - B(t_i)]^2 = \sum_{i=1}^{m-1} \left[R_i - \sum_{j=1}^{n-1} P_j N_{j,k}(t_i) \right]^{2'}, \quad (8)$$

where

$$R_i = Q_i - Q_0 N_{0,k}(t_i) - Q_m N_{m,k}(t_i), \quad i = 1, 2, \ldots, m-1. \quad (9)$$

According to the least squares principle, calculate the partial derivative of control points $P_l (l = 1, 2, \ldots, n-1)$ as follows:

$$\frac{\partial f}{\partial D_l} = \sum_{i=1}^{m-1} \left[-2R_i N_{l,k}(t_i) + 2N_{l,k}(t_i) \sum_{j=1}^{n} P_j N_{j,k}(t_i) \right], \quad (10)$$

and then

$$-\sum_{i=1}^{m-1} R_i N_{l,k}(t_i) + \sum_{i=1}^{m-1}\sum_{j=1}^{n-1} D_j N_{j,k}(t_i) N_{l,k}(t_i) = 0, \quad (11)$$

$$\sum_{i=1}^{m-1} \left(\sum_{j=1}^{n-1} N_{j,k}(t_i) N_{l,k}(t_i) \right) D_j = \sum_{i=1}^{m-1} R_i N_{l,k}(t_i). \quad (12)$$

Transform equation (12) into matrix form and then

$$\left(\mathbf{N}^T \mathbf{N} \right) D = R, \quad (13)$$

where

$$\mathbf{N} = \begin{bmatrix} N_{1,k}(t_1) & \cdots & N_{n-1,k}(t_1) \\ \vdots & \ddots & \vdots \\ N_{1,k}(t_{m-1}) & \cdots & N_{n-1,k}(t_{m-1}) \end{bmatrix},$$

$$\mathbf{R} = \begin{bmatrix} N_{1,k}(t_1) R_1 + & \cdots & + N_{1,k}(t_{m-1}) R_{m-1} \\ \vdots & \ddots & \vdots \\ N_{n-1,k}(t_1) R_1 + & \cdots & N_{n-1,k}(t_{m-1}) R_{m-1} \end{bmatrix}, \quad (14)$$

$$\mathbf{D} = \begin{bmatrix} D_1 \\ \vdots \\ D_{n-1} \end{bmatrix}.$$

The control point matrix $\mathbf{P}$ can be calculated approximately as follows:

$$P = \left(\mathbf{N}^T \mathbf{N} \right)^{-1} \mathbf{N}^T Q. \quad (15)$$

Furthermore, the mathematical expression of the fitting curve Q^c can be obtained as

$$Q^c = N \left(\mathbf{N}^T \mathbf{N} \right)^{-1} \mathbf{N}^T Q. \quad (16)$$

Letting $q_i^C \in Q^c$ be the point on the fitting curve corresponding to parameter t_i, and the sum of squares of the least squares error SSE is calculated as

$$\text{SSE} = \sum_{i=1}^{m} \left| q_i - q_i^C \right|^2. \quad (17)$$

According to the above theory, the key of the B-spline curve fitting method is to find out the parameter vector T and the knot vector U of the fitting data set. In previous researches, the knot vector is always fixed, and then, the parameter vector c be selected by methods such as uniform parameterization, the Gauss–Newton approach, and centripetal model parameterization. Alternatively, the parameter vector is first determined by the cumulative chord length parameterization, and then, the knot vector is calculated [18, 19]. However, the accuracy of the fitting curve obtained by these methods is not satisfying, and it is difficult to obtain the optimal fitting curve. Therefore, in this paper, we combine the GA and B-spline curve fitting method to solve the nonlinear problem by changing the parameter vector and node vector simultaneously [20]. Meanwhile, considering the internal relationship between the parameter vector and the knot vector, when the parameter vector and the number of control points are determined, the appropriate knot vectors can be directly calculated by the average ordered parameter method [21, 22]. In this way, each adjacent knot interval corresponds to at least one data point that ensures that the fitting curve has high fidelity. Therefore, the fitting problem of the B-spline curve is transformed into the problem of using a GA to find out the optimal parameter vector and control points without coding the parameter vector and knot vector at the same time, which reduces the complexity of the algorithm.

3.2. Genetic Algorithm Design in B-Spline Curve Fitting

3.2.1. Notation.
All the relevant notations used in the genetic algorithm are listed in Table 1.

3.2.2. Selection of the Initial Population and Chromosome Coding.
The initial population of the genetic algorithm is generally generated randomly. According to the features of the B-spline curve, we generate the initial population of size N randomly in this paper. Coding methods usually include binary coding and real coding. In order to reflect the increasing characteristic of the parameter vector in the B-spline curve more intuitively, real coding is adopted in this paper. The chromosome of each individual in the population is coded as an $(m + 1)$-dimensional increasing real vector in the following equation:

$$\{G_1, G_2, \ldots, G_m, G_{m+1}\}, \quad (18)$$

where $G_1 = 0, G_2 = 1$, and $G_{m+1} \in (0, 1]$.

The top m genes of the chromosome represent the parameter vector corresponding to the fitting data, which are increasing and randomly selected within the interval $[0, 1]$. While the $(m + 1)th$ gene of the chromosome represents the number of control points, the number cannot be less than four because the degree of the B-spline curve is three. So we select G_{m+1} randomly within the interval $[4, m]$.

3.2.3. The Fitness Function.
The fitness function is directly related to the quality of the final result and the optimization efficiency of the genetic algorithm. In order to obtain a fitting

TABLE 1: Notation in genetic algorithm.

	Parameters
N	Population size
ρ_c	Crossover probability
ρ_{c1}, ρ_{c2}	Parameters of the crossover probability
ρ_u	Mutation probability
ρ_{u1}, ρ_{u2}	Parameters of the mutation probability
f_{mean}	Average fitness
f_{max}	Maximum fitness
f_{min}	Minimum fitness
H	Set of parent chromosomes $H = \{h_i \mid i = 1, 2, \ldots, N\}$
C_c	Chromosomes produced by crossover
C_u	Chromosomes produced by mutation
C	Set of offspring chromosomes $C = C_c + C_u$
$\{G_1, G_2, \ldots, G_m, G_{m+1}\}$	Individual chromosome gene

curve with the least squares error with as few control points as possible, we design the fitness function as

$$\text{fitness} = \frac{1}{1 + \text{SSE} + \lambda \times n},\qquad(19)$$

where SSE is the least squares error and λ is the weight factor of control points.

λ affects the number of control points of the fitting curve. Generally, the more control points there are, the higher the precision of curve fitting, but it is easy to fall into "over-fitting" if there are too many control points.

3.2.4. Crossover.
Common crossover methods include single-point crossover, two-point crossover, uniform crossover, and linear crossover. In this paper, the top m genes of individual chromosomes represent the parameter vector, so the new chromosome obtained after crossover should also maintain the increasing feature of parameter vectors, and thus, linear crossover is more suitable.

Two individuals are randomly selected from the set of parent chromosomes, random number r is generated within the interval $[0, 1]$, and the linear crossover is operated if $r < \rho_c$ as follows:

$$\begin{cases} h'_1 = sh_1 + (1-s)h_2, \\ h'_2 = (1-s)h_1 + sh_2, \end{cases}\qquad(20)$$

where s is a random number in the interval $[0, 1]$ and h'_1 and h'_2 are the new chromosomes after crossover. In this way, the top m genes of the new chromosome keep increasing, and we also need to round the $(m + 1)$th gene of the chromosome to an integer because G_{m+1} is the number of control points.

Meanwhile, based on the theory of inheritance of superiority, we set the crossover probability of good chromosomes with a higher fitness to be larger than that of chromosomes with lower fitness so as to ensure that superior genes can be passed on to their offspring and improve the optimization ability of the algorithm. The dynamic crossover probability ρ_c is calculated by the fitness of the population as follows:

$$\rho_c = \begin{cases} \dfrac{\rho_{c1} - (\rho_{c1} - \rho_{c2})(f_{mean} - f_2)}{(f_{mean} - f_{min})}, & f_1 \leq f_{mean}, \\[3mm] \dfrac{\rho_{c1} - p_{c2}(f_{max} - f_2)}{(f_{max} - f_{min})}, & f_2 \leq f_{mean} < f_1, \\[3mm] \rho_{c1}, & f_2 > f_{mean}, \end{cases}\qquad(21)$$

where f_1 is the larger fitness value of the two chromosomes to be crossed while f_2 is the smaller fitness.

3.2.5. Mutation.
According to the coding method and the characteristics of chromosomal genes, we adapt the single-point mutation method, which means every gene on the chromosome of every individual in the population may mutate.

Let $h_1 = \{G_1, G_2, \ldots, G_m, G_{m+1}\}$ be one chromosome from the set of parent chromosomes. For each gene G_i of h_1, we generate a random number r within the interval $[0, 1]$ and operate single-point mutation if $r < \rho_u$ as

$$G'_i = \begin{cases} 0, & i = 1, \\ sG_{i+1}, & i = 2, \\ G_{i-1} + s(G_{i+1} - G_{i-1}), & 2 < i < m, \\ 1, & i = m, \\ \text{Round}(G_i + s), & i = m + 1, \end{cases}\qquad(22)$$

where s is a random number in the interval $[0, 1]$, G'_i is the new chromosome after mutation, and $\text{Round}(G_i + s)$ when $i = m + 1$ means that G_{m+1} is rounded to an integer.

Also based on the theory of inheritance of superiority, in order to preserve the superior genes, we set the good chromosomes with high fitness to mutate with low probability, while the chromosomes with low fitness mutate with high probability to seek new optimization directions. The dynamic mutation probability ρ_u is calculated by the fitness of the population as

$$\rho_u = \begin{cases} \dfrac{\rho_{u1}\,(f_{\max} - f)}{(f_{\max} - f_{\mathrm{mean}})}, & f \geq f_{\mathrm{mean}}, \\[2mm] \rho_{u2}, & f < f_{\mathrm{mean}}, \end{cases} \qquad (23)$$

where f is the fitness of the chromosomes to be mutated.

3.2.6. Selection. The study shows that the convergence of the GA mostly lies in the selection operator, which may be roulette selection, the expected value selection method, and the sorting selection method. We adapt the classic roulette selection in this paper. We first calculate the fitness of the set of parent chromosomes and the set of offspring chromosomes generated by crossover and mutation, respectively, and then determine the selection probability of the individual chromosome based on its corresponding fitness ratio in all chromosomes, and finally, select the new generation population according to the selection probability.

3.2.7. Algorithm Flow

Step 1. Input the degree of the B-spline curve, the fitting data set Q, the size of the population N, the maximum iteration number $t_{\max}$, and the fitting precision ε. After the initial population is randomly generated, we set iteration number $t = 1$.

Step 2. Calculate the fitness parameters including $f_{\mathrm{mean}}, f_{\max},$ and $f_{\min}$.

Step 3. Select a pair of chromosomes from the set of parent chromosomes randomly, calculate the crossover probability ρ_c, generate new chromosomes by the linear crossover operation, and add the new chromosomes into C_c.

Step 4. Select a chromosome from the set of parent chromosomes in turn, calculate the mutation probability ρ_u, operate single-point mutation on every gene of the chromosome, and add the mutated chromosome into C_u.

Step 5. Calculate the fitness of the set of offspring chromosomes C and select a new generation of size N from the set H and C by the roulette selection method.

Step 6. If $t > t_{\max}$ or $SSE \leq \varepsilon$, the algorithm is terminated. Otherwise, set $t = t + 1$ and repeat steps 2 to 5.

The algorithm flowchart is shown in Figure 4.

4. Fitting Curve Optimization Based on Dynamic Adjustment of the Fitting Data Set

4.1. Optimization Model of the E-T Fitting Curve. We could preliminarily screen the original data set by the discriminant criterion of outliers described in Section 2 and obtain the candidate fitting data set. However, some points with large inferiority in the candidate fitting data set may make the curve fitting effect poor and are therefore not suitable for inclusion in the fitting data set. For example, as shown in Figure 5, points B and D could be eliminated by the discriminant criterion of outliers so the candidate fitting

data set includes points $A, E, F, C,$ and G, but points E and F have relatively large inferiority compared with other points and are bound to affect the curve trend, which means the energy consumption of the fitting curve is not optimal. Therefore, it is important to select the optimal fitting data set from the candidate fitting data set to obtain the optimal E-T curve, so we propose an optimization model for the E-T fitting curve calculated by the fitting method in Section 3.

Each point on the optimal E-T curve fitted by the B-spline curve fitting method should satisfy the following:

$$B(T_i) \leq E_{p_i^s}, \quad \forall p_i^s (T_i, E_i) \in P^s, \qquad (24)$$

where $B(T_i)$ is the ordinate value of the point on the fitting curve when its abscissa value is T_i.

Based on the approximate inverse relation between the optimal energy consumption and running time, the optimal E-T fitting curve should satisfy the monotonicity and continuity of the first and second derivatives, expressed mathematically in equations (25) and (26), respectively:

$$B'(T_i) \leq 0, B'(T_i) < B'(T_{i+1}), \quad \forall q_i (T_i, E_i) \in Q, \qquad (25)$$

$$B''(T_i) \geq 0\ B''(T_i) > B''(T_{i+1}), \quad \forall q_i (T_i, E_i) \in Q. \qquad (26)$$

At the same time, considering the feasibility of fitting, the number of points in the fitting data set should be guaranteed, shown as follows:

$$|Q| \geq 3. \qquad (27)$$

On the basis of satisfying the above constraints, as many points as possible must be included. Therefore, the objective function of the optimization of the E-T fitting curve is put forward as follows:

$$\max |Q| - \lambda_1 \mathrm{SSE}, \qquad (28)$$

where λ_1 is the penalty factor of the least squares error SSE.

In summary, the optimization model of the E-T fitting curve is as follows:

$$\begin{cases} \max & |Q| - \lambda_1 \mathrm{SSE}, \\ & B(T_i) \leq E_{p_i^s}, \forall p_i^s (T_i, E_i) \in P^s, \\ \text{s.t.} & B'(T_i) \leq 0, B'(T_i) < B''(T_{i+1}), \forall q_i (T_i, E_i) \in Q, \\ & B''(T_i) \geq 0\ B''(T_i) > B''(T_{i+1}), \forall q_i (T_i, E_i) \in Q, \\ & |Q| \geq 3. \end{cases}$$
$$(29)$$

4.2. Optimization Algorithm Design Based on Dynamic Adjustment of the Fitting Data Set. Firstly, the fitting data set Q is selected from the candidate data set P^s randomly, and according to the E-T curve fitted by the B-spline curve fitting method based on the GA, we constantly adjust the fitting data set to improve the fitting results. Due to the fact that the fitting precision ε is a tiny number and $SSE < \varepsilon$, SSE can be
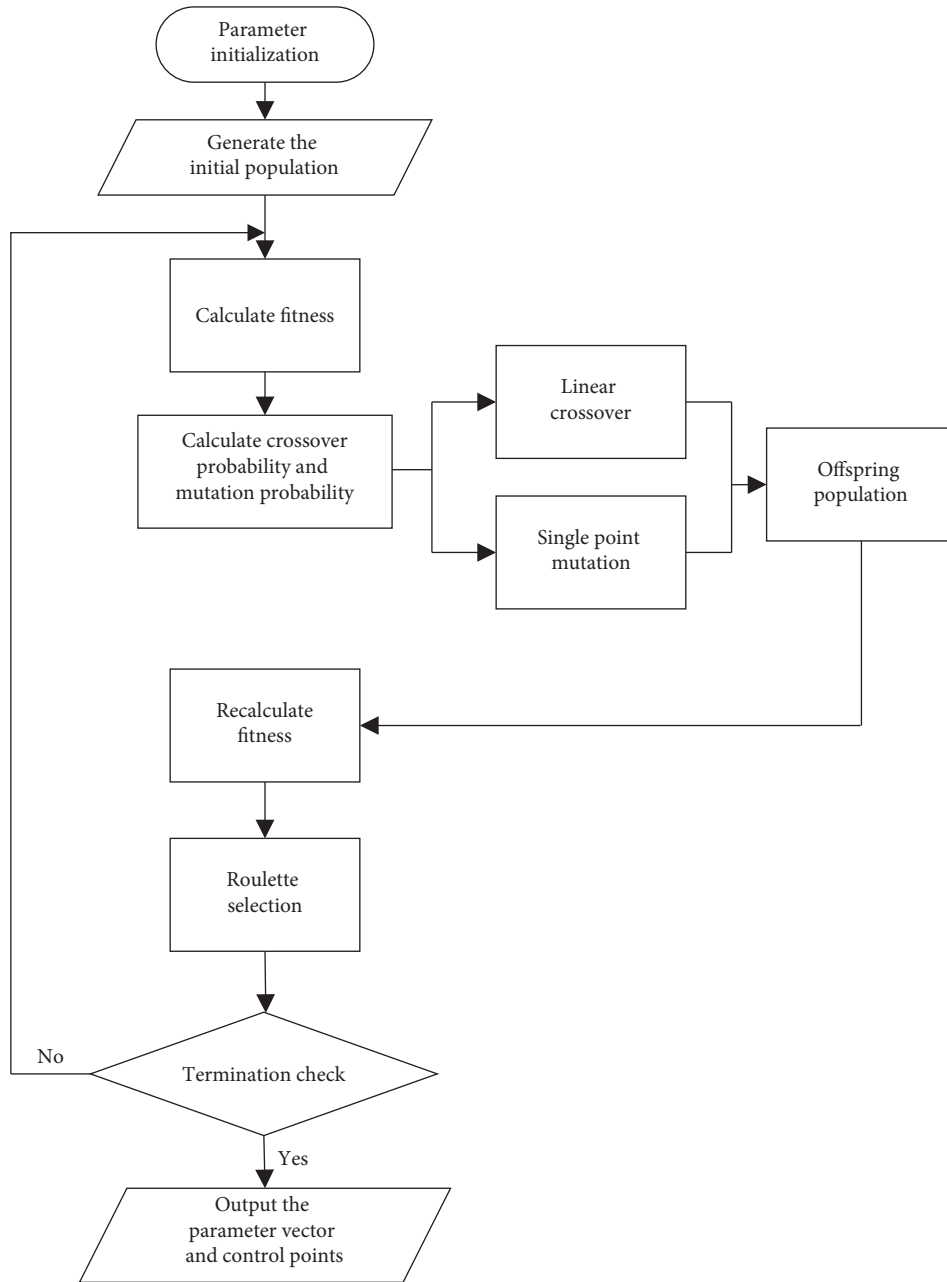
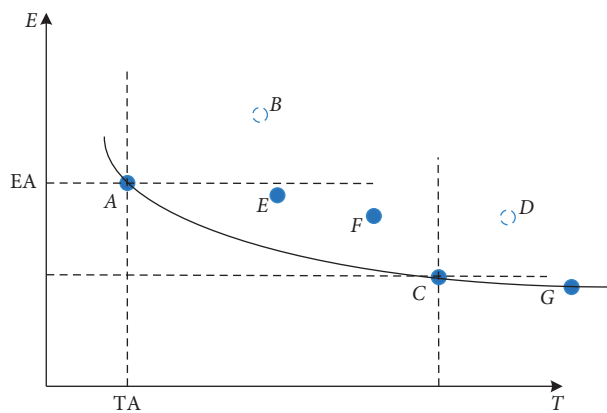Figure 4: Algorithm flowchart.

Figure 5: Elimination of points with larger inferiority.

regarded as zero in the objective function. The penalty factors of the constraints in equations (24), (25), and (26) are added to the objective function in the following equation:

$$\max|Q| - \lambda_2 \sum_{\substack{p_i^s(T_i,E_i)\in P^s-Q \\ B(T_i)-E_{p_i^s}>0}} \left(B(T_i)-E_{p_i^s}\right)^2 - \lambda_3 \sum_{\substack{q_i(T_i,E_i)\in Q \\ B'(T_i)>B_I(T_{i+1}) \\ B''(T_i)<0}} (\delta(T_i))^2 - \lambda_4 \sum_{\substack{q_i(T_i,E_i)\in Q \\ B''(T_i)<0}} (B''(T_i))^2, \tag{30}$$

where λ_2, λ_3, and λ_4 are the penalty factors, and generally, $\lambda_4 < \lambda_3 < \lambda_2$ because constraint in equation (24) relates to the optimization of energy consumption, while the constraints in equations (25) and (26) relate to the smoothness of the fitting curve.

We design the optimization algorithm on the basis of the theory of tabu search. We define bidirectional tabu lists including the tabu list Q^+ ($Q^+ \in P^s - Q$) for adding points to the fitting data set and the tabu list Q^- ($Q^- \in Q$) for removing points from the fitting data set. For $q \in Q^+$ or $q \in Q^-$, $\eta(q)$ is the tabu step size of point q and η_{max} is the maximum tabu step size.

The steps to add points to the fitting data set Q are as follows:

Step 1. Calculate the data set $Q_0^+ = \{p_i^s(T_i, E_i) | p_i^s \in P^s - Q, B(T_i) - E_{p_i^s} > 0\}$ that does not meet the constraint in equation (24).

Step 2. Select the point to be added. If $Q_0^+ = \varnothing$, do not add points to the fitting data set Q. Otherwise, if $Q_0^+ \neq \varnothing, Q_0^+ - Q^+ \neq \varnothing$, select point $p \in Q_0^+ - Q^+$ by the roulette selection method and add point p into data set Q. The selection probability is calculated as follows:

$$\rho^+(p) = \frac{\left[B(T_i) - E_{p_i^s}\right]^2}{\sum_{p\in Q_0^+-Q^+}\left[B(T_i) - E_{p_i^s}\right]^2}. \tag{31}$$

If $Q_0^+ \neq \varnothing, Q_0^+ - Q^+ = \varnothing$, select one point randomly from the data set PTS $= \{p|p \in Q^+, \eta(p) = 1\}$ and add this point into the data set Q.

Step 3. Update the bidirectional tabu lists. As for the new point p added into data set Q, set $Q^- = Q^- \cup \{p\}$ and $\eta(p) = \eta_{max}$. For $\forall q \in Q^+$, set $\eta(q) = \eta(q) - 1$, and if $\eta(q) = 0$, set $Q^+ = Q^+ - \{q\}$.

The steps to remove points from the fitting data set Q are as follows:

Step 1. Calculate the data set $Q_0^- = \{p(T_i, E_i) | p \in Q, B'(T_i) > B'(T_{i+1}), B'(T_i) < 0\}$ that does not meet the constraints in equations (25) and (26).

Step 2. Select the point to be removed. If $Q_0^- = \varnothing$, do not remove points from the fitting data set Q. Otherwise, if $Q_0^- \neq \varnothing, Q_0^- - Q^- \neq \varnothing$, select point $p \in Q_0^- - Q^-$ by the roulette selection method and remove point p from data set Q. The selection probability is calculated as

$$\rho^-(p) = \frac{[\delta(p)]^2}{\sum_{p\in Q_0^--Q^-}[\delta(p)]^2}. \tag{32}$$

If $Q_0^- \neq \varnothing, Q_0^- - Q^- = \varnothing$, select one point randomly from data set MTS $= \{p|p \in Q^-, \eta(p) = 1\}$ and remove this point from data set Q.

Step 3. Update the bidirectional tabu lists. As for the point p removed from data set Q, set $Q^+ = Q^+ \cup \{p\}$ and $\eta(p) = \eta_{max}$. For $\forall q \in Q^-$, set $\eta(q) = \eta(q) - 1$, and if $\eta(q) = 0$, set $Q^- = Q^- - \{q\}$.

The adjustment of the fitting data set Q terminates when $Q_0^+ = \varnothing$ and $Q_0^- = \varnothing$. The algorithm flowchart is shown in Figure 6.

5. Experimental Examples

The train operation data samples in this paper are mainly composed of actual train operation data and simulated data, and the section running time range is 80–120 s. Generally, the data samples are based on the actual operation data; however, in the actual train operation process, the value range of section running time is relatively limited (82–85 s). We simulate some data samples as supplements by the software named "Urban rail transit train traction calculation and operation diagram energy consumption evaluation" which is used by the operation department of Guangzhou Metro. For each train operation data, we take the unit distance (such as 0.1 m) as the calculation step, and the energy consumption is calculated by accumulating the power (the traction force multiplies the distance) at all distance steps. We extracted a bunch of E-T points to form the original data set, as shown in Figure 7.

The energy consumption of most data points shown in Figure 7 is within a reasonable range; however, there are also some points whose energy consumption value is obviously too high. Based on the discriminant criterion of outliers, we preliminarily remove outliers and points with unreasonable energy consumption from the original data set, and all remaining points after filtering constitute the candidate fitting data set in Figure 8.

Although the overall trend of the candidate data set conforms to the characteristics of the E-T curve, it can be seen from the partially enlarged view that some points are unordered and have great inferiority that will definitely affect
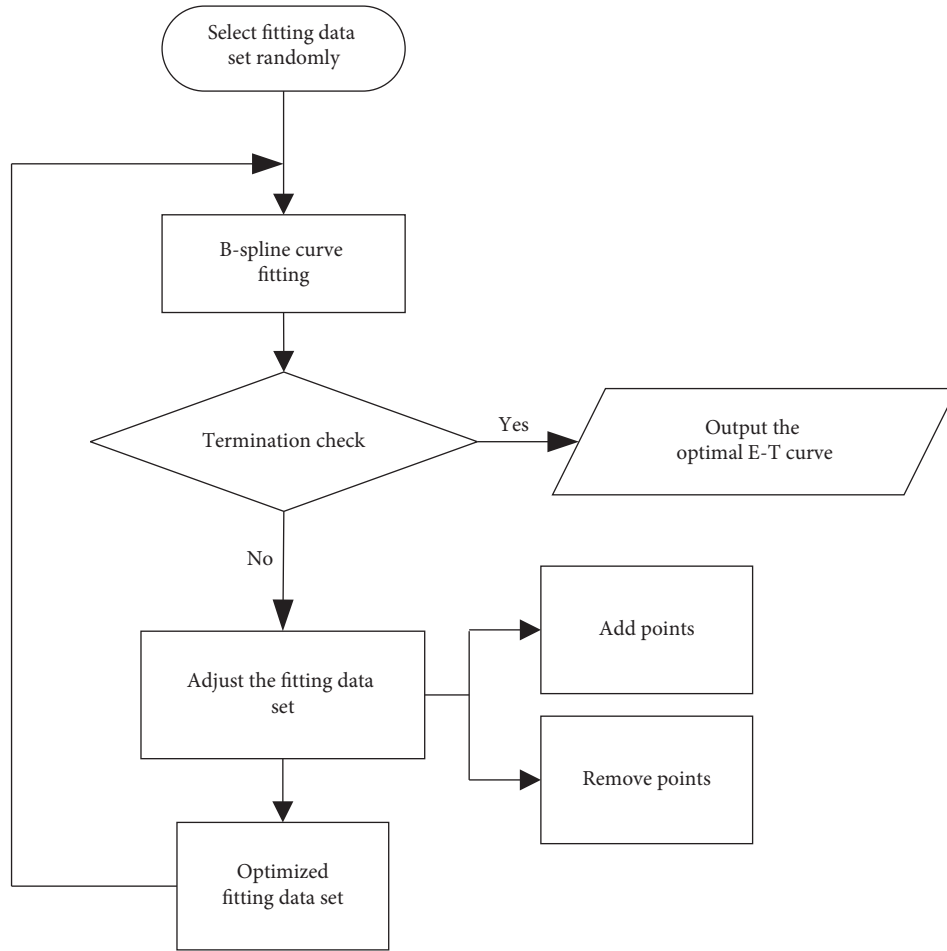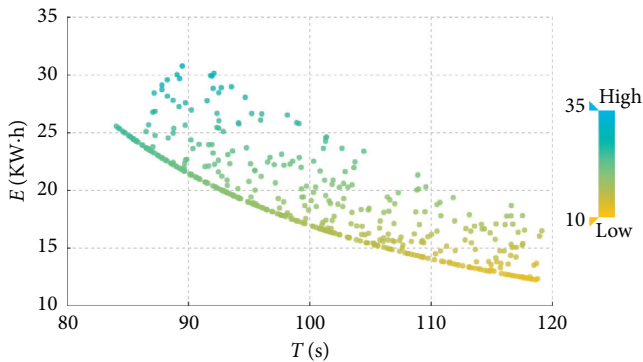
Figure 6: Optimization algorithm flowchart.

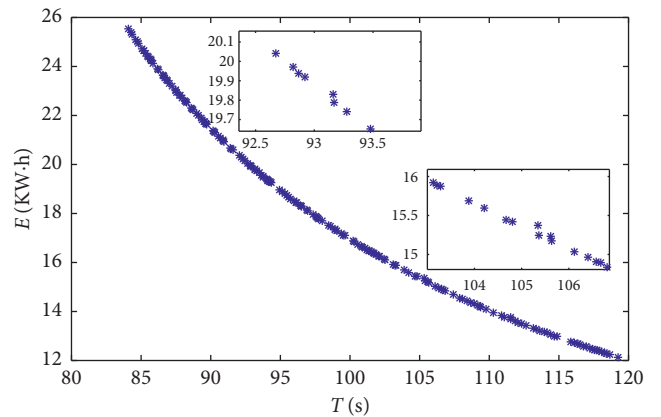Figure 7: Scatter chart of the original data set.

Figure 8: Scatter chart of the candidate data set.

the fitting effect of the E-T curve and could not ensure optimal energy consumption of the curve. Therefore, we have to select the optimal fitting data set from the candidate data set by the optimization model of the fitting curve proposed in Section 4, and the model parameter settings are shown in Table 2.

We randomly select 68 points for the fitting data set Q. From the candidate data set P^s, and as the number of iterations increases, data set Q is constantly adjusted and eventually includes 82 E-T points when $Q_0^+ = \varnothing$ and $Q_0^- = \varnothing$. The results are shown in Figures 9 and 10.

In Figure 9, although the value of the objective function fluctuates, the overall trend is to increase with the number of iterations, and the data set Q_0^+ that does not meet the constraint in equation (17) drops significantly and finally drops to zero, indicating that the energy consumption of the fitting curve points basically reaches the optimal value. Because we have already removed some outliers from the original data set, the number of points in the data set Q_0^- that does not meet the

TABLE 2: Parameter values in the model.

Parameter	Value
Degree of B-spline curve k	3
Initial population size N	40
Parameters of the crossover probability ρ_{c1}, ρ_{c2}	$\rho_{c1} = 0.9, \rho_{c2} = 0.2$
Parameters of the mutation probability ρ_{u1}, ρ_{u2}	$\rho_{u1} = 0.3, \rho_{u2} = 0.8$
Fitting precision ε	0.001
Maximum tabu step size $\eta_{\max}$	3
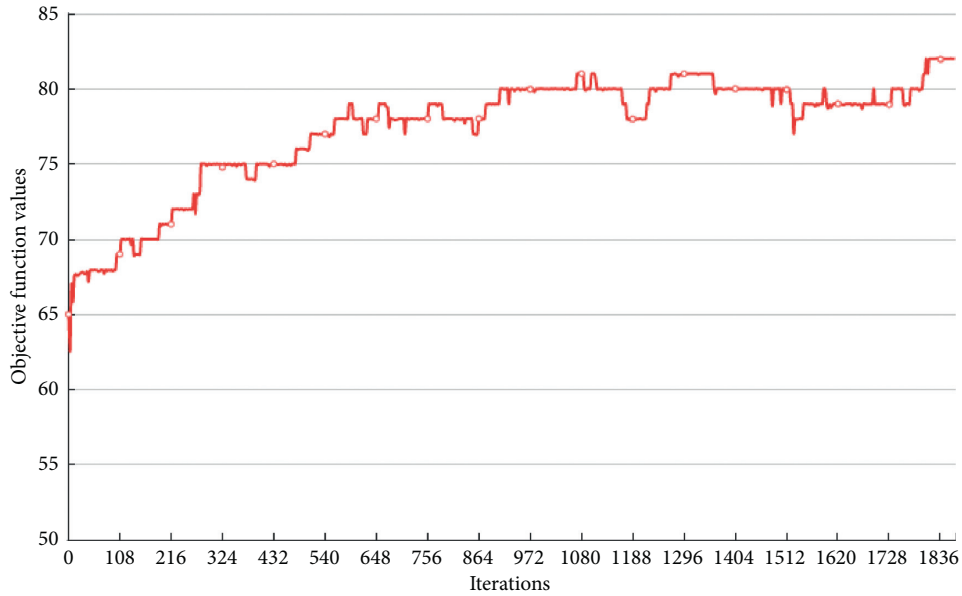
FIGURE 9: Change trend of the value of the objective function.

constraints in equations (18) and (19) stays low all the time and eventually decreases to zero as well, which ensures that the fitting curve has good continuity and smoothness.

The optimal energy consumption curve fitted by the optimized fitting data set using the B-spline curve fitting method based on the GA is shown in Figure 11. Figure 11(a) shows the changing trend of the least squares error and the average fitness of the population, and Figure 11(b) is the optimal E-T fitting curve.

The least squares error of the fitting curve decreases rapidly, and the average fitness of the population increases as the number of generations increases. In the 166th generation, the algorithm is terminated when the least squares error SSE = 0.00943, which meets the requirements of fitting accuracy, and the average fitness of the population also stabilizes. This proves that the B-spline fitting method based on the GA proposed in this paper has strong optimization ability, high fitting accuracy, and high convergence speed.

Meanwhile, the comparison between the optimized fitting curve and the original fitting curve fitted by the candidate data set is shown in Figure 12.

The original fitting curve is fitted by the candidate fitting data set, while the optimized fitting curve is fitted by the fitting data set optimized and adjusted by the optimization algorithm in Section 4. Although the overall trends of the two curves are similar, in a partially enlarged

view, the optimized fitting curve is smoother after eliminating some points with great inferiority which have an influence on the curve, and by contrast, the optimized fitting curve is below the original fitting curve on the whole, which means that the goal of energy consumption optimization has been achieved well. Taking 0.1s as the time interval, we, respectively, select 353 sample points with the same running time from the original fitting curve and the optimal fitting curve and all sample points are from 84.1 s to 119.2 s. By accumulating the difference ratio of all sample points, the energy consumption of the optimal fitting curve is 0.69 KW h less than that of the original fitting curve, and the maximum energy consumption difference among all sample points is 0.16%. From the calculation, the energy consumption of the optimized fitting curve is lower than that of the original fitting curve and the optimal fitting data set has fewer fitting data points, which proves that the optimization method of the fitting curve proposed in this paper can select the optimal fitting data set from a large number of original data points and obtain the optimal E-T curve.

The optimal E-T curve can reflect the lowest operation energy consumption under different section running times, so each data point on the optimal E-T curve corresponds to the optimal operation curves, such as the velocity-distance (V-T) curve and time-distance (T-S) curve. In Figure 13, we
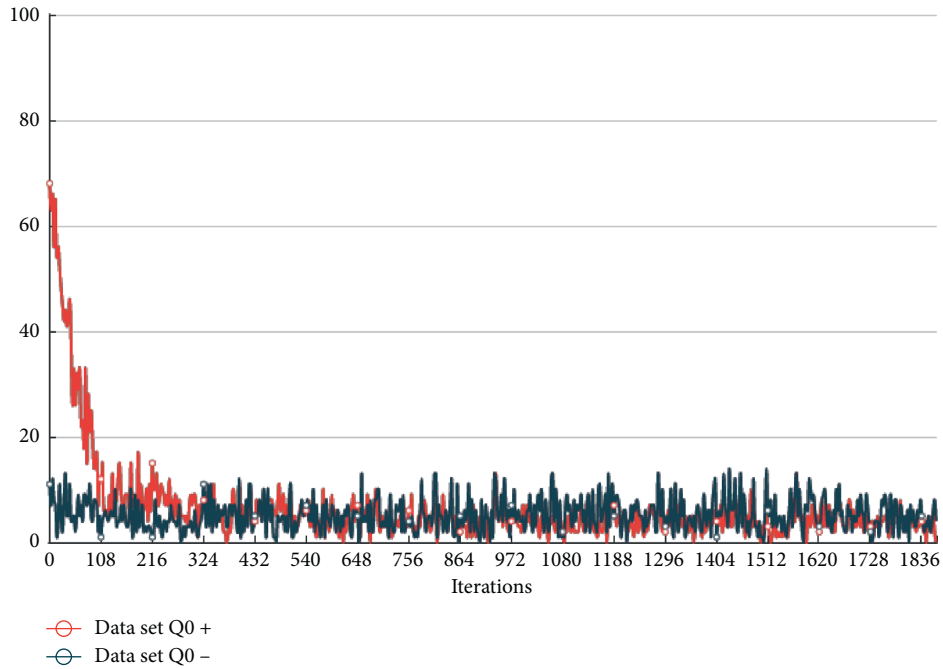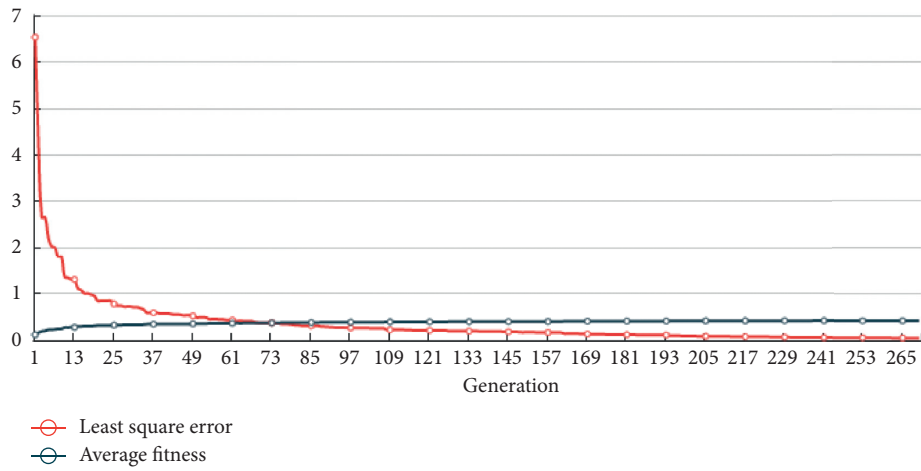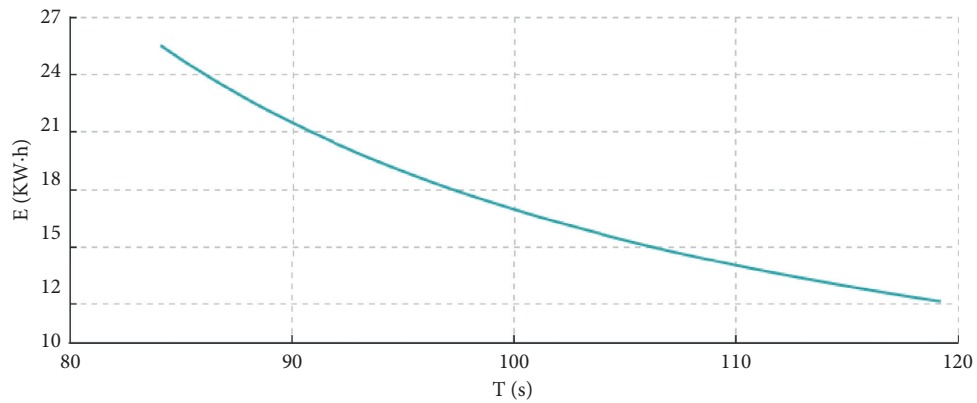
FIGURE 10: Change trend of data sets that do not meet the constraints.

(a)

(b)

FIGURE 11: B-spline fitting results: (a) change trend of the least square error; (b) change trend of the average fitness of the population.
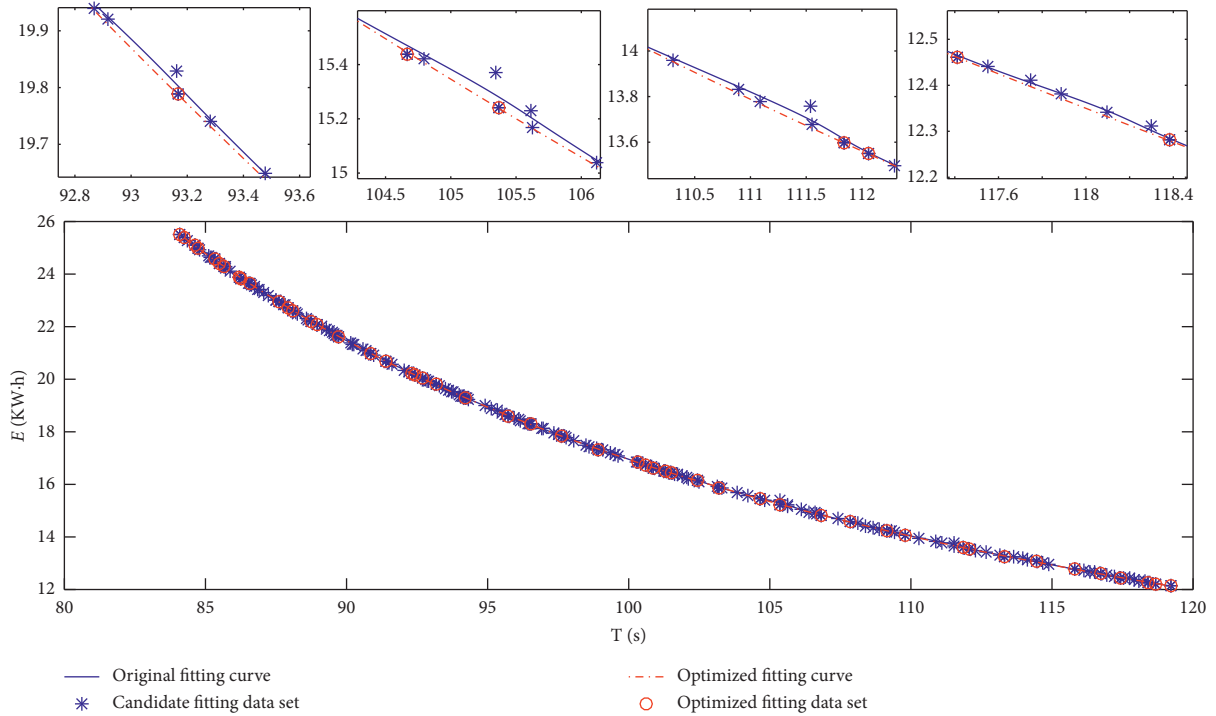
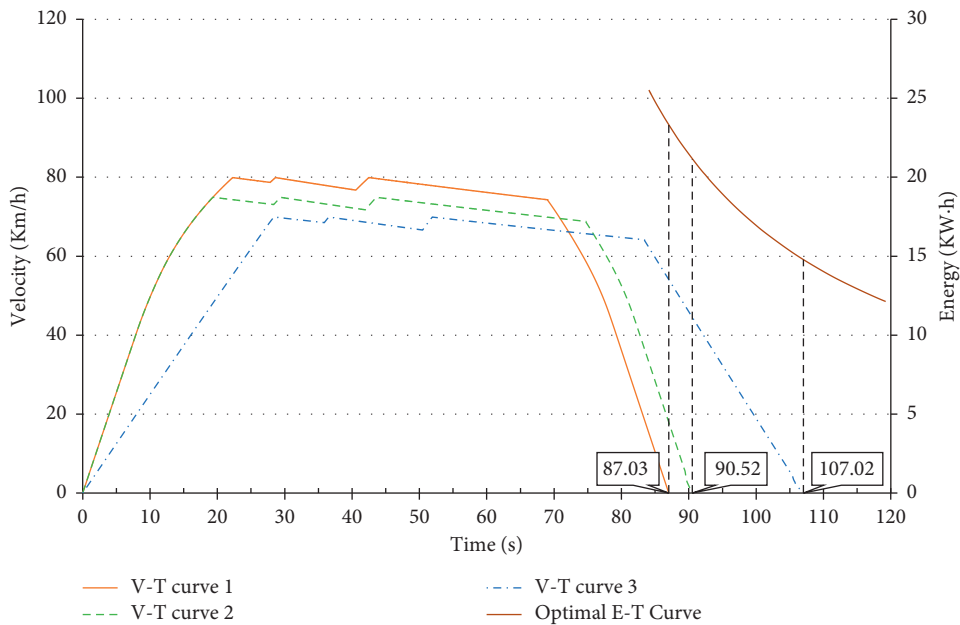FIGURE 12: Comparison of fitting curves fitted by different fitting data sets.

FIGURE 13: V-T curves corresponding to the optimal E-T points.

randomly select three E-T points from the optimal fitting data set and extract the corresponding V-T curves that have the optimal energy consumption.

The V-S curves with the optimal energy consumption can provide abundant running curve support for the operation of the train under different running time requirements.

6. Conclusions

In this paper, we propose a fitting method of the optimal energy consumption-running time curve of an urban rail section based on train operation data. The main work completed includes the following:

(1) Based on the features of the section operation curve of urban rain trains and correlations between the running time and corresponding energy consumption, the discriminant criterion of outliers is proposed to select the candidate fitting data set from many original data points, which reduces the scale of the train operation data set as well as guaranteeing the curve fitting quality.

(2) An improved B-spline curve fitting method is proposed in which the parameter vector and knot vector are optimized by the genetic algorithm, which has a higher fitting accuracy and faster convergence speed.

(3) On the basis of tabu search, we construct an optimization model of the fitting curve by defining bidirectional tabu lists to adjust and optimize the fitting data set from the candidate data set dynamically. It is proposed that the optimization method could obtain the optimal E-T curve and ensure the continuity and smoothness of the fitting curve at the same time.

The research on the optimal E-T curve based on operation data of urban rain trains has certain practical significance beyond theoretical limitations, and the optimal E-T fitting curve could be used in the selection of section running time, evaluations of the energy consumption of the train operation diagram, and performance appraisal of train drivers. Further research will focus on the optimization ability of the algorithm of fitting curve optimization.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

References

[1] I. P. Milroy, *Aspects of Automatic Train Control*, Loughborough University, Loughborough, UK, 1980.

[2] E. Khmelnitsky, "On an optimal control problem of train operation," *IEEE Transactions on Automatic Control*, vol. 45, no. 7, pp. 1257–1266, 2000.

[3] C. S. Chang and S. S. Sim, "Optimising train movements through coast control using genetic algorithms," *IEE Proceedings - Electric Power Applications*, vol. 144, no. 1, pp. 65–73, 1997.

[4] B. R. Ke, *Block-layout Design Using MAX-MIN Ant System for Saving Energy on Mass Rapid Transit Systems*, IEEE Press, New York, NY, USA, 2014.

[5] R. Liu and I. M. Golovitcher, "Energy-efficient operation of rail vehicles," *Transportation Research Part A: Policy and Practice*, vol. 37, no. 10, pp. 917–932, 2003.

[6] I. V. Sanchis and P. S. Zuriaga, "An energy-efficient Metro speed profiles for energy savings: application to the valencia Metro," *Transportation Research Procedia*, vol. 18, pp. 226–233, 2016.

[7] Y. V. Bocharnikov, A. M. Tobias, C. Roberts, S. Hillmansen, and C. J. Goodman, "Optimal driving strategy for traction energy saving on DC suburban railways," *IET Electric Power Applications*, vol. 1, no. 5, pp. 675–682, 2007.

[8] M. R. Mashinchi, A. Selamat, S. Ibrahim et al., "Outlier elimination using granular box regression," *Information Fusion*, vol. 27, pp. 161–169, 2016.

[9] H. Hassani, M. Zokaei, D. von Rosen, S. Amiri, and M. Ghodsi, "Does noise reduction matter for curve fitting in growth curve models?" *Computer Methods and Programs in Biomedicine*, vol. 96, no. 3, pp. 173–181, 2009.

[10] S. Zheng, R. Feng, and A. Huang, "A modified moving least-squares suitable for scattered data fitting with outliers," *Journal of Computational and Applied Mathematics*, vol. 370, p. 112655, 2020.

[11] P. Chen, L. Li, Y. Liu et al., "Detection of outliers and patches in bilinear time series models," *Mathematical Problems in Engineering*, vol. 2010, Article ID 580583, 256 pages, 2010.

[12] A. Gálvez and A. Iglesias, "Firefly algorithm for explicit B-spline curve fitting to data points," *Mathematical Problems in Engineering*, vol. 2013, pp. 206–226, 2013.

[13] G. Trejo-Caballero, H. Rostro-Gonzalez, C. H. Garcia-Capulin et al., "Automatic curve fitting based on radial basis functions and a hierarchical genetic algorithm," *Mathematical Problems in Engineering*, vol. 2015, Article ID 731207, 14 pages, 2015.

[14] S. Su, T. Tang, X. Li et al., "Optimization of multitrain operations in a subway system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 673–684, 2014.

[15] X. Li, C. F. Chien, L. Li et al., "Energy-constraint operation strategy for high-speed railway," *ICIC International*, vol. 8, no. 10, pp. 6569–6583, 2012.

[16] H. Kang, F. Chen, Y. Li, J. Deng, and Z. Yang, "Knot calculation for spline fitting via sparse optimization," *Computer-Aided Design*, vol. 58, pp. 179–188, 2015.

[17] D. D. Hearn, *Computer Graphics with OpenGL*, Publishing House of Electronics Industry, Beijing, China, 2004.

[18] W. Zheng, P. Bo, Y. Liu, and W. Wang, "Fast B-spline curve fitting by L-BFGS," *Computer Aided Geometric Design*, vol. 29, no. 7, pp. 448–462, 2012.

[19] R. Świta and Z. Suszyński, "Thermal image approximation using B-spline surfaces," *International Journal of Thermophysics*, vol. 39, no. 11, p. 127, 2018.

[20] H. SHOU and L. HU, "B-spline curve fitting algorithm for real-coded GA with normal constraint," *Journal of Zhejiang University of Technology*, vol. 44, no. 4, pp. 466–472, 2016.

[21] L. A. Piegl and W. Tiller, "Least-squares B-spline curve approximation with arbitary end derivatives," *Engineering with Computers*, vol. 16, no. 2, pp. 109–116, 2000.

[22] F. Yoshimoto, T. Harada, and Y. Yoshimoto, "Data fitting with a spline using a real-coded genetic algorithm," *Computer-Aided Design*, vol. 35, no. 8, pp. 751–760, 2003.