Health diagnosis of bus operation based on multi-source data

Zhou, Xuemei; Guan, Zhen; Pang, Yusong; Ji, Xiangfeng; Lin, Xiaodan

DOI
10.1049/itr2.12169

Publication date
2022

Document Version
Final published version

Published in
IET Intelligent Transport Systems

Received: 7 May 2021 | Revised: 10 December 2021 | Accepted: 14 January 2022

DOI: 10.1049/itr2.12169

IET Intelligent Transport Systems

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Health diagnosis of bus operation based on multi-source data

Xuemei Zhou[1] | **Zhen Guan**[1] | **Yusong Pang**[2] | **Xiangfeng Ji**[3] | **Xiaodan Lin**[1]

[1] College of Transportation Engineering, Key Laboratory of Road and Traffic Engineering of the State Ministry of Education, Tongji University, Shanghai, China

[2] Section Transport Engineering and Logistics, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands

[3] Department of Management Science and Engineering, School of Business, Qingdao University, 62 Keda Branch Road, Laoshan District, Qingdao, Shandong Province, China

Correspondence
Xuemei Zhou, College of Transportation Engineering, Tongji University, Key Laboratory of Road and Traffic Engineering of the State Ministry of Education, Shanghai, China.
Email: zhouxm@tongji.edu.cn
Xiangfeng Ji, Department of Management Science and Engineering, School of Business, Qingdao University, 62 Keda Branch Road, Laoshan District, Qingdao, Shandong Province 266000, China.
Email: jixiangfeng@qdu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 61873190, 71801138

Abstract

Based on the multi-source data available for bus operations, this paper proposes a health diagnosis system for single-line bus operation systems from two aspects: The operation efficiency and stability. Firstly, the index weight has been defined and calculated based on the Entropy Method. The composite index of bus operation has been obtained, and the health classification standards that pertain to efficient and effective bus operations have also been constructed (Very Healthy, Healthy, Sub-Healthy, and Unhealthy). Secondly, the more efficient machine learning method has been used in order to establish the classification algorithm training set. The effect of the k-Nearest Neighbour and Decision Tree Classification Model has also been compared and analysed in this particular study. Finally, a bus line in Foshan is taken as a case study to verify the effectiveness of the method. This paper can effectively improve the diagnosis efficiency and accuracy by introducing the artificial intelligence algorithm into bus operation diagnosis. It provides a foundation for the development of bus operation health diagnosis decision support system with the function of "bus disease" prevention and treatment.

1 | INTRODUCTION

As a critical component of urban internal passenger transportation, the urban bus service is the primary mode of transportation for most of the urban residents in China. However, the buses in many cities are in the long-term unhealthy operation, manifesting low punctuality, speed fluctuations and other "bus diseases" [1]. Therefore, it is very urgent to improve the bus operation state. The health diagnosis of bus operation is the prerequisite of scientific management of "bus disease". An early and accurate diagnosis of any underlying issues tends to play a vital role in solving problems in a timely manner. It has become possible to monitor the bus operation status and realize the potential holographic perception regarding the future trends with the popularization of accumulating large amounts of data in bus operation systems and the maturity in artificial intelligence technology [2]. Therefore, it is a rather revolutionary phenomenon to dig deeper into the extensive amount of multi-source data that has been captured on the bus operation, diagnose the health status of bus operation and analyse operation bottlenecks addressed to the needs of the consumer.

Comparing bus operation system to human body, analysing bus operation based on multi-source data is like making a comprehensive examination for bus system. The health diagnosis based on the examination results is equivalent to the doctor's interpretation of the examination report. It provides a systematic and accurate grounds for the treatment of "bus disease". The diagnosis of bus operation should be carried out in the order of "physical examination, health judgment, disease diagnosis" like the diagnosis process of patients in the hospital [3]. In order to accurately diagnose the bus operation health status

based on multi-source data, it is necessary to select an appropriate index set first, which helps to examine the bus operation level from different aspects. Then a reasonable health classification standard is needed which helps to judge what constitutes an unhealthy state and how unhealthy it is. Finally, the health status of bus operation can be diagnosed according to the results of "physical examination".

Based on the multi-source data fusion and machine learning this paper first determines the evaluation index system of bus operation. After that, a health classification standard is discussed and determined. Then, we propose a method for bus operational health diagnosis. Finally, a case study is used to verify the effectiveness of the method.

2 | LITERATURE REVIEW

The current researches carried out on the discipline of bus operation diagnosis mainly focus on three aspects: (1) The application of multi-source data in the evaluation of bus operation; (2) bus operation evaluation indexes; (3) the health diagnosis method pertaining to bus operation.

The newer generation of high-tech systems and operative methods primarily represented by computers, mobile internet, big data and artificial intelligence, has seen a gradual but definite development over time. In this regard, data such as bus automatic vehicle location (AVL), global positioning system (GPS), bus application (APP), bus trajectory, taxi floating car, integrated circuit card (IC card), mobile phone signalling data has also been further improved. It provides the possibility for the early and effective diagnosis of complex bus operation systems [4]. In particular, the growth and application of the big data concept further supports the continuous development of experimental transportation research. Therefore, in terms of public transport, this research mainly focuses on the analysis pertaining to the residents' travel characteristics [5], travel behaviour, and passenger flow forecast [6], by using the IC card data [7]. In the era of big data, it is possible to evaluate the bus services by integrating the service capacity of the bus line network with the spatial information of crowd activities, or to analyse supply levels of bus by predicting the travel needs of citizens [8]. Moreover, in terms of data fusion, there are many available data sources that have been explored, in addition to the widely used AVL and automatic vehicle identification (AVI) data [9]. It is anticipated that there will be many complications if the GPS data pertaining to bus operations is not integrated with other data. These complications might include exposure to insufficient data samples, unbalanced spatial distribution, and insufficient coverage rate, etc. Therefore, many scholars have proposed a variety of data fusion methods, that too from various angles, such as the genetic algorithm, fuzzy logic and wavelet transform [10], in order to make the best possible analysis of the bus operation systems. Based on the floating car data fusion of taxi and bus, the accuracy of traffic state judgment is higher than that based on the single data [11], and the fusion analysis of mobile phone data and bus data [12] can be used for the evaluation of bus network optimization.

TABLE 1 Classification of evaluation indexes

Target layer	Criterion layer	Index layer
Efficiency	Based on the bus operation efficiency	Running speed
		Running speed between stations
		Running time ratio
	Based on the bus timetable	Schedule adherence
		Departure interval
		Weighted delay index
Reliability	Based on the bus operation stability	Coefficient of variation of running speed between stations
		Coefficient of variation of running time
		Stability ratio of time headway
		Coefficient of variation of time headway
Convenience	Based on passenger perception	Waiting time
		Transfer time
		Station coverage rate
		Congestion level (full load coefficient of carriages)

It is noteworthy that the evaluation of bus operation reliability [13, 14] plays an important role in public transport planning and management. The research about the bus operation evaluation indexes is mainly divided into four categories according to different focus points (the bus timetable, the operation efficiency, the operation stability and the passenger perception). When viewed from different perspectives, various evaluation indexes can be used for the evaluation, such as the schedule adherence, station coverage rate, the travel time [15–17], the running speed [18], the coefficient of variation of the bus running time (CV_t) [19], and the transfer time [20]. In this paper, the evaluation indexes are classified after comprehensive analysis, and they are summarized in the Table 1.

Once the evaluation indexes are determined, the health diagnosis of the bus operation status can be performed. The Health diagnosis is a research tool which is initially applied mainly in the field of bridges and tunnels. Some scholars had introduced it to the field of transportation in recent years. The bridge health diagnosis methods are mainly divided into non-artificial intelligence methods and artificial intelligence methods. The non-artificial intelligence diagnosis methods mainly include the signature analysis approaches [21, 22], model updating approaches [23–25], identification methods based on statistical analysis [26], genetic algorithms [27, 28], and the fuzzy method [29–31]. Whereas, the artificial intelligence methods mainly include the neural network [32, 33], and the support vector machine [34]. In addition, further development of ensemble learning, deep learning, and transfer learning also provide newer and more innovative ideas for carrying out the bridge health diagnosis [35]. At present, there have been relatively comprehensive tunnel health diagnosis standards, including the qualitative criteria, quantitative criteria, or even a combination of the two. Diagnosis based

on tunnel health, which is based on the analysis of tunnel disease investigation and testing results, is to establish a tunnel computational model to analyse the health status of the tunnel structure [36] by applying mathematical mechanics methods such as finite element model [36] and shear crack model [37].In the field of urban traffic and bus health diagnosis, it is indispensable to construct an evaluation index system, and then calculate the weights. The weight calculation methods include the Fuzzy Comprehensive Evaluation Method [38], the Analytic Hierarchy Process, Expert Decision-Making [39, 40], the Data Envelopment Analysis [41], and the Entropy Method [42].

There has been a large amount of research studies based around multi-source data fusion, bus operation evaluation, and health diagnosis. However, most of the traditional research methods focused only on the evaluation system construction and result calculation of the bus operation status. They only provided the values of the evaluation index but not the diagnostic standards which were a necessary part of the analysis. For bus management and operators, a judgment result is not enough. They are more interested in how this result is judged, what factors it based on, and how it should be solved. A machine learning based approach can give explicit judgment criteria. In fact, the logic of the machine's judgment is exactly what the human has actively made it learn. Secondly, when using non-artificial intelligence methods to assess the health status of bus operations, it is necessary to calculate all bus routes within a city. Due to the excessive amount of data, the computation time of this type of method is unacceptable. The introduction of machine learning algorithms can significantly increase the computational speed. Thirdly, the judgment in non-artificial intelligence methods is generally based on the absolute magnitude of the calculated results. It requires not only consideration of global data, but also subjective factors. Machine learning methods can effectively alleviate this problem by simply optimizing the training set. In this process, the subjective factor is greatly weakened.

The purpose of this study is to provide fast and accurate diagnosis of bus operation health by means of data fusion and efficient machine learning methods, relying only on basic data sets such as GPS. Compared to traditional evaluation methods, the method in this paper allows for autonomous classification and diagnosis capabilities through learning. It has significant advantages in terms of calculation speed, result accuracy and implementation difficulty. The method represents a very effective tool in preventing and eradicating urban "bus disease".

3 | CONSTRUCTION OF HEALTH DIAGNOSIS SYSTEM FOR SINGLE-LINE BUS OPERATION

In order to diagnose the health level of bus operation, the evaluation system should be constructed firstly. The appropriate indexes should be selected by analysing the data characteristics available and the construction principles. And then, the evaluation index is calculated to provide training samples for the subsequent classification algorithm through the processing and analysing of multi-source data of bus operation.

3.1 | Preparation of bus operation data

Based on the multi-source data like GPS data of bus operation, the bus operation data is pre-processed and the useful data is extracted. Through the fusion processing of different data, it is transformed into the data format needed for the research.

3.1.1 | Data source and overview

The multi-source data in this research comes from the bus operation data that is specific to Foshan city, from the time period spanning from November 21 to 23, 2016. It includes variables such as the bus GPS data, bus operation schedule, bus IC card data, Foshan road network vector data, and the bus network data.

The GPS data used for the purpose of this study does not include the name of the bus line, the up and down line directions, the arrival time, the departure time, and the dwell time at the station. It just includes vehicle identification number, record time, longitude, latitude, and valid. Therefore, it is necessary to fuse the GPS data with other data, in order to make the data availability better. In this specific study, the mapping relationship between the line and the license plate in the IC card data was mainly used to obtain the single-line bus operation data. Other than that, the geographic information science (GIS) data includes the Foshan road network data and the bus network data. Combining the national road network data and the Foshan boundary data, and using the ArcGIS geoprocessing function for intersection processing, the researchers also obtained the Foshan road network data. Furthermore, based on the Python Crawler Algorithm, the direction of the Foshan bus line through the application programming interface (API) of Gaode map is also mapped out. The data structure mainly includes the line name, the originating station, the terminal station, the longitude and latitude of the line direction, the station name, and its longitude and latitude information.

3.1.2 | Data pre-processing

For the original data obtained, it needs to be pre-processed to extract the useful data. The whole process of pre-processing is shown in the Figure 1 below.

Single-line data matching

For the purpose of this study, standardization of the operation schedule of each company was undertaken. After that the mapping table of bus line and the license plate number was expressed in a unified format. In case the operation schedule was missing, this study made use of the IC card data to obtain the mapping relationship. The combination of the two kinds of data [43] could then match the line and bus license plate number more accurately and avoid any wastage of the obtained data. Other than that, based on the obtained mapping relationship between the license plate and the line, the license plate number of each line has been extracted and stored as a new comma

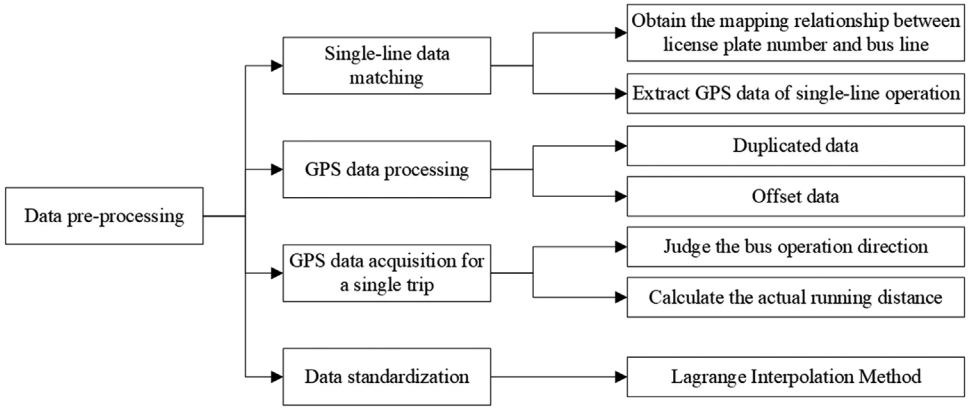

FIGURE 1 Process of data pre-processing

separated value (CSV) file that has been named after the bus line designation. Finally, the GPS data of the bus operation of each line in a day has also been obtained.

GPS data processing

It was already anticipated that there would be some errors in the process of collecting the GPS data transmission. As a result, it is necessary to pre-process the GPS data to improve the reliability of GPS data. In this regard, the cleaning process of the GPS data basically refers to the deletion or modification of the data which has a problem. This may include problems such as a completely duplicated data, partially repeated data, offset data, or erroneous data [44]. When processing with the duplicated data, the most common and effective practice is to retain the first occurrence of the data and eliminate duplicate values. For the offset data, the map matching principle is used to match the GPS data of the bus with the data of the bus network crawled by the GIS. Therefore, the Nearest Neighbour Analysis Method [45] in ArcGIS has been used to match the offset GPS data to the nearest road network. Similarly, the offset coordinates have been calculated in this manner.

GPS data acquisition for a single trip

In order to calculate the distance between each of the GPS data points and the originating station, there are a number of measures that need to be considered. First, there must be a judgement regarding whether the data points are in the originating station, in the terminal station, or in the operation range (50 m radius of the station is the judgment standard). Then, the GPS data points need to be read and interpreted cyclically to judge the bus operation direction. When the judgment is completed, the operation data of each trip is stored separately.

It is noteworthy that the bus always runs along the existing road. But the existing road is curved. Therefore, its actual running distance is related to the shape of the road. Keeping this in mind, this paper adopts the cumulative method to calculate the actual running distance of the bus. It means that the GPS data of each bus is sorted in chronological order and the distance between two points is calculated. Then all the distances are added to get the actual running distance of the bus. Finally, the relationship between the actual running distance and the time of each trip is obtained. The distance between stations can be calculated in the same way. This is the specific processing of GPS data to calculate the distance between each station and the originating station. By subtracting the distance between this station and the originating station from the distance between its next station and the originating station, the distance between the two adjacent stations can also be obtained.

Data standardization

Sometimes the dwell time at the physical location of the station lasts only a few seconds. If the dwell time happens to be between two GPS points, it may not even be recognized. In order to improve the accuracy of the dwell time, it is necessary to standardize the GPS data. Before this standardization procedure, the data collection frequency tends to be 3–10 s. After the standardization, the data update time changes to 1 s. For this purpose, Lagrange Interpolation Method [46] has been used to standardize the data.

3.2 | Construction of health diagnosis system

In Table 1 we listed a large number of indicators related to the evaluation of bus operation status. It is unrealistic to calculate and take into account all the indicators. Therefore, it is necessary to screen. Some indicators are not available through the data we have in hand, such as waiting time. Some indicators are not applicable to single-route bus systems, such as station coverage. Therefore, we selected the metrics in terms of efficiency and stability respectively, according to the operation mode of the study scenario.

Bus operation efficiency is mainly expressed in terms of speed. Therefore, we chose the average running speed of the line and the running time index as the two metrics. The stability can be mainly reflected by variance. After comparison, the coefficient of variation of time headway and the coefficient of variation of the running speed between the stations were finally chosen as the other two indices. After that, the definition and calculation of specific indices will be completed.

3.3 | Definition and calculation of health diagnosis index

3.3.1 | Running speed

The running speed (v) refers to the average running speed of a bus from the originating station to the terminal station. It is a critical index to measure the efficiency of passenger bus travel. Therefore, it can be represented by the following equation:

$$v = \frac{l}{T_{actual}} \tag{1}$$

In the equation, v represents the average running speed of the bus. l denotes the actual running distance of the bus. T_{actual} represents the actual travel time of the bus. Its value can be calculated by reading the operating data of each trip in a single line circularly, and dividing the actual running distance by the actual running time.

3.3.2 | Running time index

The running time index (T_r) refers to the ratio of the actual running time that is consumed by the bus during the entire operation, to the time required for driving under the condition of free flow velocity. The lesser the running time index is, the faster the operation of the bus. Thus, the calculation formula is as follows:

$$T_r = \frac{T_{actual}}{T_{free\ flow\ velocity}} \tag{2}$$

In the equation, T_r is the running time index. T_{actual} denotes the actual running time of a single trip. $T_{free\ flow\ velocity}$ is the running time of a single bus under the condition of free flow velocity, which is expressed by the speed limit of the road section. The running time under the free flow velocity is calculated by constructing the cost matrix in ArcGIS, while the speed limit of the road section is taken as the free flow velocity.

3.3.3 | Coefficient of variation of time headway

For bus services with short departure intervals, the reliability of bus operation is determined by the stability of the time headway. (C_{cv}) represents the coefficient of variation of time headway of vehicles, arriving at a certain station, on a specific line. Since the GPS device in the bus may not exactly collect the data at any section, only the time headway of the station section is calculated for convenience.

$$C_{cv} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(TH_{actual} - TH_{planned}\right)^2}}{TH_{planned}} \tag{3}$$

In the equation, C_{cv} is the coefficient of variation of time headway. N represents the number of stations. TH_{actual} denotes the actual time headway of station section. $TH_{planned}$ shows the planned time headway of this line. First, the arrival interval between each car and the next car is calculated. After the whole line is calculated, the standard deviation of arrival interval is calculated. Finally, it is divided by the planned departure interval to be the coefficient of variation of time headway. In order to calculate the actual time headway, it is necessary to judge the arrival time and the departure time at the station, and then calculate the arrival interval between each bus, and the next bus. In this paper, when the GPS data point is 50 m away from the bus station, it is considered that the bus has arrived or departed to or from its desired destination. In this manner, the planned time headway is obtained according to the operation schedule.

3.3.4 | Coefficient of variation of running speed between stations

V_{cv} refers to the standard deviation of the running speed between stations. It primarily describes the fluctuation of the average running speed of the different trips between the stations. It can reflect the stability of the bus operation.

$$\bar{v}_s = \frac{1}{m}\sum_{i=1}^{m} v_i \tag{4}$$

$$V_{cv} = \sqrt{\frac{1}{N}\sum_{i=1}^{N-1}\left(v_i - \bar{v}_s\right)^2} \tag{5}$$

In the equation, V_{cv} represents the coefficient of variation of the running speed between stations. v_i shows the average running speed between the i-th station and the (i+1)-th station during the statistical period. $\bar{v}_s$ denotes the average value of v_i during the statistical period. m represents the number of trips during the statistical period. N shows the number of bus stations.

4 | HEALTH DIAGNOSIS BASED ON MACHINE LEARNING

4.1 | Construction of training set based on entropy method

The most common training set construction methods include the Expert Scoring Method, the Analytic Hierarchy Process, and the Entropy Method. However, the Expert Scoring and Analytic Hierarchy Process are greatly influenced by human subjective factors. Since this study is based entirely on the data pertaining to bus operations, the Entropy Method is deemed to be more appropriate, and has been referred to in order to obtain the weight of the evaluation index, in an objective manner. This method has been widely used by the government authorities, and for professional research as well.

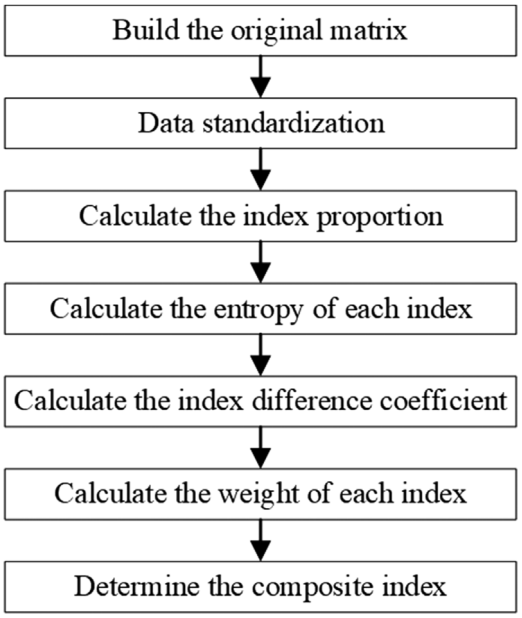

FIGURE 2 Calculation process of entropy method

4.1.1 | Weight calculation for each index

The Entropy Method can make full use of the existing data, determine the index weight according to the degree of variation of the index value, and avoid the deviation caused by human factors. The basic calculation process is shown in Figure 2.

1. Build the original matrix

It is assumed that there are n samples and m evaluation indicators. The initial data matrix (X) of the evaluation system can be calculated by the equation:

$$X = \left(x_{ij}\right)_{m\times n} \tag{6}$$

In the equation, x_{ij} represents the calculated value of the i-th sample under the j-th index. There are 455 bus operation samples and 4 evaluation indicators in this study.

2. Data standardization

Due to the different calculation methods of each index, there will be different units and orders of magnitude, which makes it impossible to directly compare each index. In order to solve this problem, the index needs to be dimensionless. In this paper, the method of data standardization is as follows:

Standardization method of Positive Indexes,

$$x'_{ij} = \frac{x_{ij} - \min\left\{x_{1j},\dots,x_{nj}\right\}}{\max\left\{x_{1j},\dots,x_{nj}\right\} - \min\left\{x_{1j},\dots,x_{nj}\right\}} \tag{7}$$

Standardization method of Negative Indexes:

$$x''_{ij} = \frac{\max\left\{x_{1j},\dots,x_{nj}\right\} - x_{ij}}{\max\left\{x_{1j},\dots,x_{nj}\right\} - \min\left\{x_{1j},\dots,x_{nj}\right\}} \tag{8}$$

3. Calculate the Proportion (P_{ij}) of the i-th sample under the j-th index:

$$P_{ij} = \frac{x'_{ij}}{\sum_{i=1}^{n} x'_{ij}}; j = 1,\dots,m \tag{9}$$

4. Calculate the Entropy (e_j) of j-th index:

$$e_j = -k\sum_{i=1}^{n} P_{ij}\ln\left(P_{ij}\right); j = 1,\dots,m \tag{10}$$

In the equation, $k = \frac{1}{\ln n}$, $0 < e_j < 1$.

5. Calculate the Difference Coefficient (d_j) of j-th. Generally, the higher the difference coefficient, the lower the entropy. So the d_j can be represented by the following method.

$$d_j = 1 - e_j \tag{11}$$

6. Calculate the Wight Index (w_j) of j-th,

$$w_j = \frac{d_j}{\sum_{j=1}^{m} d_j} \tag{12}$$

7. Calculate the Composite Index (F_i) of i-th sample based on the w_j,

$$F_i = \sum_{j=1}^{m} w_j P_{ij},\ i = 1,\dots,n \tag{13}$$

This paper has selected three typical bus lines in Foshan, as the training set to determine the weight value. These include the 391, 125, and the 123 lines. In this regard, 125 is the urban bus line, 391 is the suburban bus line of Foshan, and 123 is the bus line that connects the urban and the suburban areas. These three typical lines are capable of representing the overall health status of the bus operation in Foshan. It is noteworthy that on November 21st, 455 trips were recognized from the valid data of the three lines that have been considered. Therefore, these 455 trips were taken as the sample training set for this analysis. According to the calculation steps mentioned above and the running of Python code, the result of weight calculation is shown in Table 2.

According to the result, it can be observed that the most critical index for the health diagnosis of bus operation is the running speed, followed by the running time index, while the third

TABLE 2 Calculation results of entropy method

Health diagnosis index	Entropy e_j	Difference coefficient d_j	Wight w_j
Running speed v	0.984	0.016	0.514
Running time index T_r	0.993	0.007	0.242
Coefficient of variation of time headway C_{cv}	0.995	0.005	0.167
Coefficient of variation of running speed between stations V_{cv}	0.998	0.002	0.078

and last one in order of importance is the coefficient of variation of time headway. The coefficient of variation of running speed between stations has the least influence on the composite index. Whereas, the running speed is the most important index that characterizes the health of bus operation.

4.1.2 | Health classification standard of bus operation

The quantile method is commonly used in traffic engineering discipline studies to reflect the fluctuation of vehicle speed operating on a roadway. The composite index is closely related to speed. This method was used in the next step to analyse of the composite index. The prerequisite for using the quantile method is that the speed obeys a normal distribution during the statistical period. To verify whether the composite index conforms to a normal distribution, the computed results of the training set were plotted as a frequency distribution histogram and a Q–Q plot. Q–Q plot (Quantile–Quantile plot) refers to a scatter plot based on the quartiles of the actual data distribution and the quartiles of the theoretical normal distribution. It visually shows whether the data obeys the specified theoretical distribution. If the data points in the Q–Q plot are distributed on the diagonal, the data are proven to conform to the normal distribution.

Figure 3 shows that the composite index basically obeys normal distribution. The skewness coefficient is at a level of 0.097, and the kurtosis coefficient is at 0.611, both of which are less than 1. Other than that, the data in Figure 4 is evenly distributed on the diagonal. So the composite index is considered to be approximately normal in terms of the distribution. Therefore, the division can be done by using the idea of subdivision. In this paper, three bounds of 15%, 50%, and 85% were chosen. Referring to the classification method of Shanghai traffic index, we likewise classify the health levels into four levels: very healthy, healthy, sub-healthy and unhealthy. Figure 5 shows the composite index value of the three lines that have been taken into consideration.

As can be seen from Figure 5, most of the trips pertaining to line 391 are in a very healthy or healthy state. This is primarily because it is located in the suburbs and possesses a high operation efficiency. However, line 123 and 125 are in a poor operative state. Most of them are in the unhealthy operation state especially in the morning peak period. The operation states of line 123 and 125 are greatly affected by the morning peak in the urban area. Besides, the operation state of line 125 is the worst in comparison to the other lines.

4.2 | Machine learning classification algorithm

According to the value range of the composite index, each bus operation sample in the training set had been assigned a value. The value of the unhealthy operation state is represented by 0, the sub-healthy state is represented by 1, the healthy state is represented by 2, and the very healthy state is represented by 3. It is worth noting that there is not a strict functional relationship between the value of the composite index and the level of health. This is because the results of the composite index are related to the line that was chosen for the calculation. When classifying levels by the quantile method, the composite index has value only within the group. A direct comparison of this number between different experimental groups is not meaningful. In this study, the machine learning approach understands how the segmentation is performed by learning the results of segmentation of different groups of data. Then the diagnosis problem was transformed into a multi-classification problem.

The data set in section 4.1 was divided into training set and test set. The machine learning algorithm was used to learn the classification model on the training set. Then, the performance of the classification model was tested on the test set. Finally, the health level of the bus operation could be judged based on the machine learning classification algorithm. In this paper, the k-Nearest Neighbor (KNN) Categorization algorithm and the Decision Tree algorithm were used to establish the classification model. Moreover, the K-Fold Cross Validation was used to verify the generalization ability of the model. According to the evaluation index, the model with a better classification effect was selected to classify and diagnose the health level of the bus operation in Foshan.

4.2.1 | KNN categorization algorithm

The KNN Categorization Algorithm is a simple, yet intuitive classification algorithm. It is often used to solve binary classification problems and multi-classification problems, primarily because of its high accuracy and insensitivity to outliers. Table 3 shows the implementation flow for classification using this algorithm.

In the KNN algorithm, the model parameters need to be optimized, and mainly include the k value and distance function.

Selection of the k *value*

The effects of the classification depend largely on the selection of the relevant k value. When the value of k is too small, the model will become complex, and will therefore be prone to be overfitting. Whereas, when the value of k is too large, it is equivalent to using training examples in a larger neighbourhood

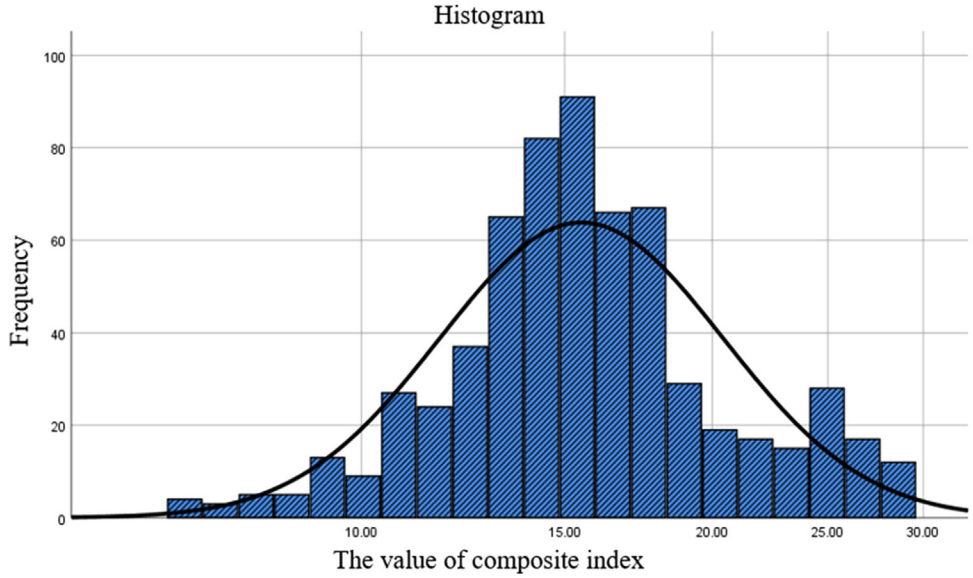

FIGURE 3 Histogram of composite index distribution

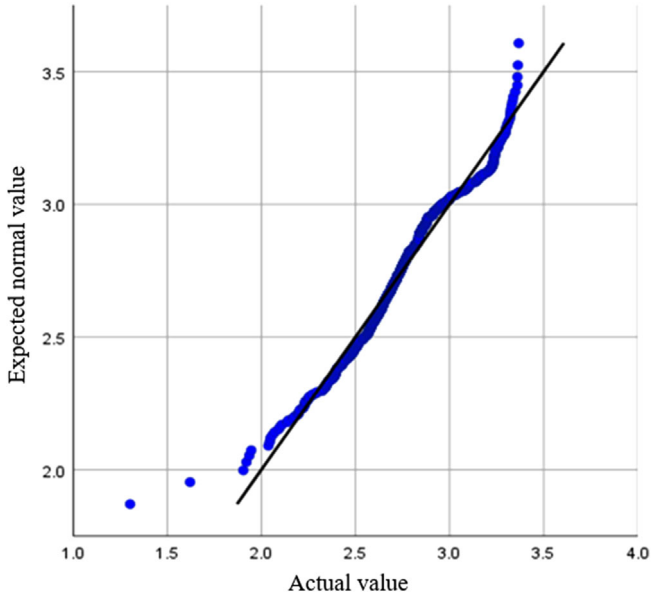

FIGURE 4 *Q–Q* normal distribution of composite index

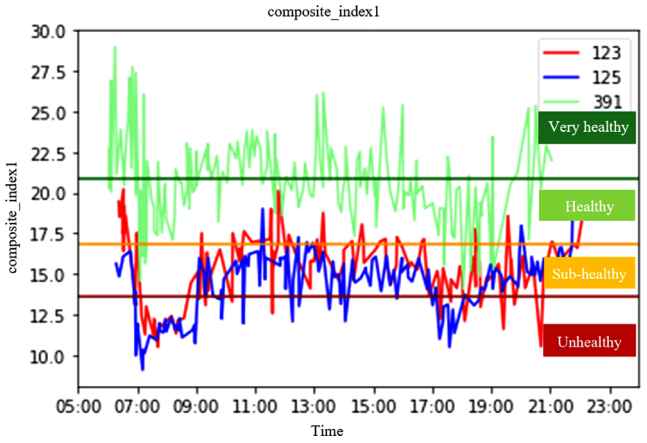

FIGURE 5 Statistics of composite index of bus operation

for prediction, so the learning approximation error will tend to increase. Accuracy is used as the score to reflect the effect of parameter taking on the model. Figure 6 shows the influence of different k values on the accuracy of the health diagnosis, after model training. The distance function adopts the Euler distance (which will be explained in detail in the next point), and the accuracy is calculated by taking the average value of the fivefold cross validation.

As can be seen from Figure 6, for the training set constructed in this study, when $k = 15$, the classification effect is at its best position, and level of accuracy reaches a value of 0.925. With the increase in the value of k, the accuracy experiences a downward trend.

TABLE 3 KNN algorithm process

KNN classification algorithm
1: Input:Dataset D $= \{(x_1, y_1,), \ldots, (x_m, y_m,)\}$, x is the characteristic matrix, y is health grading
2: Randomly take 70% of the data as the training set D and 30% of the data as the test set T
3: **for** test set $\boldsymbol{T}$ Sample values per row $\boldsymbol{t}$
4: Calculate the distance between $\boldsymbol{t}$ and the training set data according to the distance function *Distance* (D,T) respectively
5: Order by increasing distance
6: Get the k closest points
7: Get the class with the most classification among the k points return
8: **end for**
9: Output: Final classification result of the test set

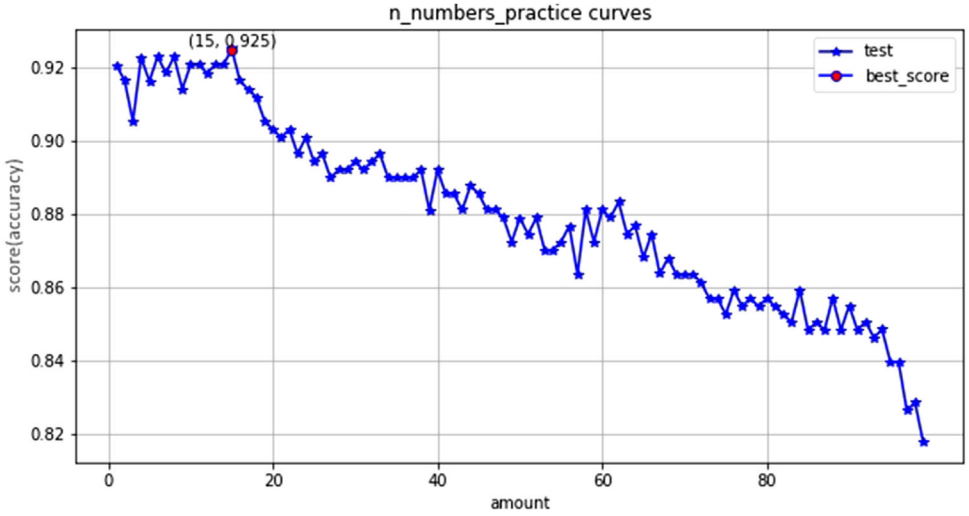

FIGURE 6 Practice curve of k value

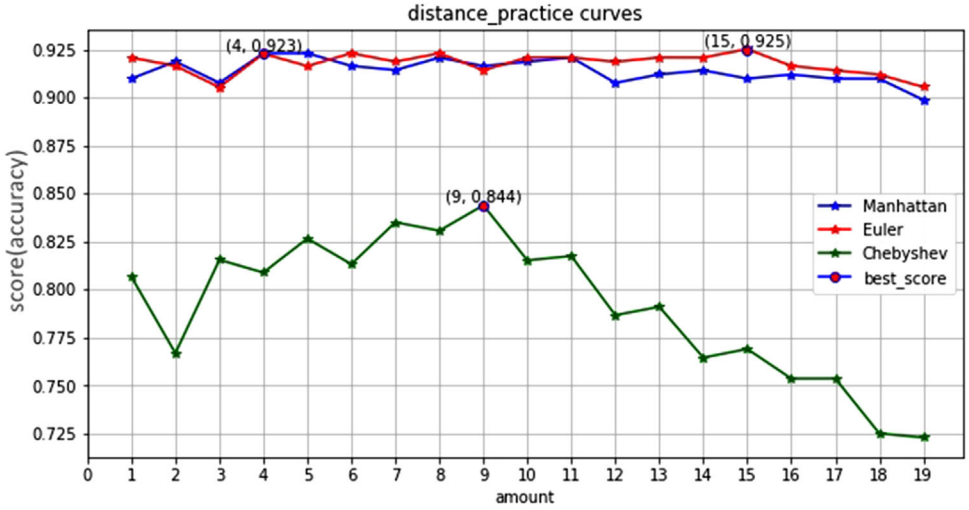

FIGURE 7 Practice curves of distance function

Selection of the distance function

Common distance functions include the Euler Distance, Manhattan Distance, and the Chebyshev Distance. Figure 7 shows the accuracy of the model under different distance functions, when k is in the range of 1–20.

It can be seen from the Figure 7, that when $k = 15$, and the distance function takes into account the Euler distance, the model classification effect is at its best level, with the level of accuracy at 0.925. The evaluation results after the optimizing model parameter are shown in the Table 4 below.

TABLE 4 Results of model evaluation

Labels of health classification	Precision	Recall	F1_Score
0 (Unhealthy)	0.91	0.85	0.84
1 (Sub-healthy)	0.84	0.93	0.88
2 (Healthy)	0.96	0.96	0.96
3 (Very healthy)	1.00	0.94	0.97
Overall accuracy	0.925		

4.2.2 | Decision tree classification algorithm

The Decision Tree is a common machine learning algorithm. The decision process of the decision tree can be represented by the tree structure. Table 5 shows the implementation flow for classification using the Decision Tree algorithm.

The parameters that need to be optimized in the decision tree can be divided into two categories: purity parameters, and pruning parameters.

TABLE 5 The Decision Tree algorithm process

The Decision Tree classification algorithm

Input:Dataset D $= \{(x_1, y_1,), \ldots, (x_m, y_m,)\}$, Feature set T $= \{t_1, \ldots, t_d\}$
Construct function Tree Generate(D,T)
Calculate the impurity of the feature set
Select the feature with optimal impurity and generate node n
if all samples in the training set belong to the same class C
Mark n as a leaf node of category C, return
end if
if the differences in all eigenvalues of the samples in the training set D are within the threshold
Mark n as a leaf node whose represented class is the class with the largest number of samples in the training set D, return
end if
if the difference in the eigenvalues of the samples in the training set D is not within the threshold
Select optimal segmentation attribute t_* from feature sets
for each value t_*^v
Generate a branch for n. Let D_v denote the subset of samples in D that take the value t_*^v at t_*
if D_v is an empty set
Branch nodes are labelled as leaf nodes and classes are labelled as the class with the highest number of samples in the training set, return
else
Use Tree Generate $(D_v, T\{t_*\})$ as a branch node
end if
end for
Output: Decision tree with n as the root node

Purity parameters

The key step of the decision tree learning is to select the optimal partition feature. That is to say that there are three decision tree algorithms that can accurately, and efficiently select the appropriate feature. These include the Iterative Dichotomiser 3 (ID3), C4.5, and the Classification and Regression Tree (CART). The ID3 uses information gain, in order to select the optimal partition features, while the C4.5 uses gain ratio, and the CART uses the Gini index. The information entropy method is very easy to appear as an overfitting phenomenon for small samples, but the effect of the Gini coefficient is often better when analyzed under smaller samples. In this context, there are a few relevant features in this study, and the data dimension is not high as well, therefore, the CART has been deemed to be more appropriate.

Pruning parameters

The decision tree algorithm is prone to appear as an overfitting problem. Pruning is generally used to solve the overfitting problem, in order to improve the generalization performance. Moreover, the pruning strategy tends to extend great amount of influence on the decision tree. Applying the correct pruning strategy is the core to effectively optimizing the decision tree algorithm. The common pruning parameters mainly include maximum features (f_max), maximum depth (d_max), minimum samples leaf (sl_min), and minimum impurity decrease (id_min). When performing optimization, the most important thing is to ensure the accuracy of the model. Therefore, accuracy is used as the score to reflect the effect of parameter taking on the model.

TABLE 6 Optimal parameter value

Parameter	Value
d_max	6
sl_min	2
id_min	0

TABLE 7 Evaluation results of Decision Tree Model

Labels of health classification	Precision	Recall	F1_Score
0 (Unhealthy)	0.98	0.85	0.91
1 (Sub-healthy)	0.88	0.98	0.93
2 (Healthy)	0.97	0.93	0.98
3 (Very healthy)	1.00	0.96	0.98
Overall accuracy	0.941		

a. f_max: This is a pruning parameter used in order to limit the overfitting of the high-dimensional data. There are only four features in this study, so the parameter optimization process is not needed to consider it.
b. d_max: This measure is used to limit the maximum depth of the decision tree. It cuts the tree beyond the set depth in a direct manner. In this regard, Figure 8 shows the practice curve of the maximum depth, in which the d_max dwells in the range of 1–20. Moreover, Figure 8 shows that when the range of d_max is [3, 6], the classification effect tends to be enhanced. When the d_max is greater than 6, the overfitting phenomenon is likely to become more obvious.
c. sl_min: Each leaf node in the decision tree must contain at the sl_min samples after branching. Otherwise, the branch will not continue ahead. Figure 9 shows the practice curve of the sl_min, where the range lies in between 1 and 50. Moreover, Figure 9 shows that the effect tends to be better when the sl_min is [5, 17], and with the increase of the sl_min, the classification effect of the model gradually tends to become worse.
d. id_min: If the information gain is less than the set value, the branch will not continue ahead. As can be seen in Figure 10, the accuracy decreases with the increase of information gain. Therefore, the model effect is better when the id_min is [0–0.08].

According to the optimal range of the three parameters that have been defined, the grid search method is used to obtain the optimal parameter value. The optimal parameter values is shown in Table 6, where the rate of accuracy reaches a value of 0.941.

After optimizing the parameters in the Decision Tree model, the results of the model evaluation are shown in Table 7.

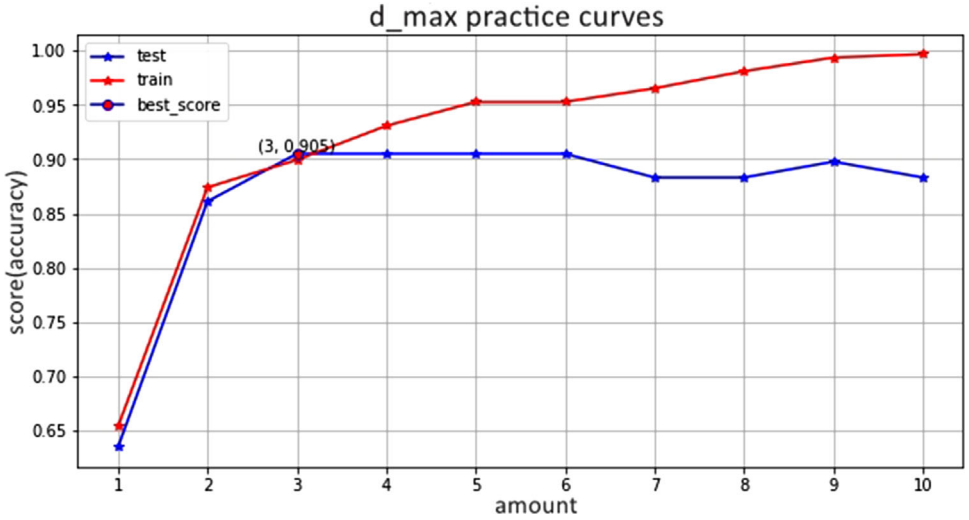

FIGURE 8 Practice curves of d_max

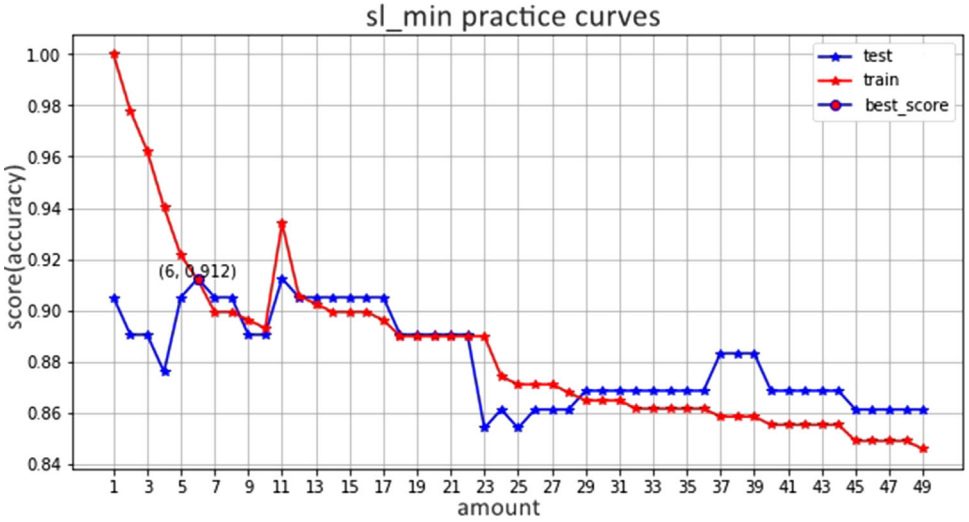

FIGURE 9 Practice curves of sl_min

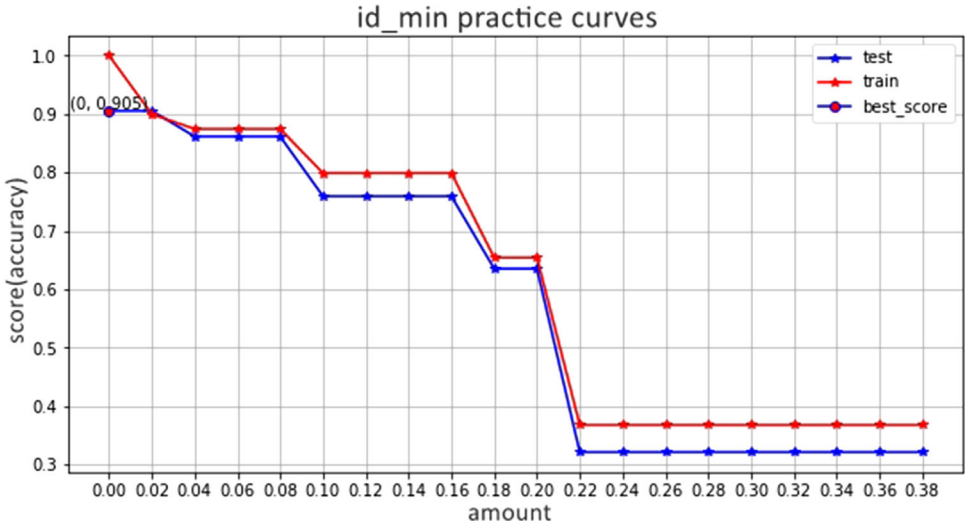

FIGURE 10 Practice curves of id_min

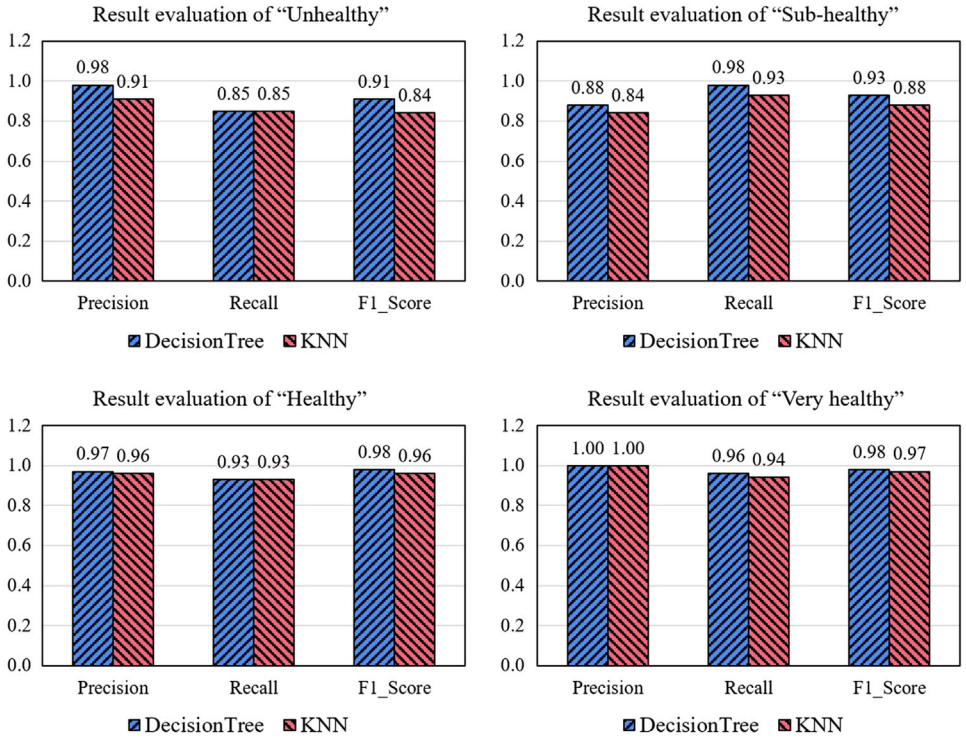

FIGURE 11 Comparison of evaluation indexes

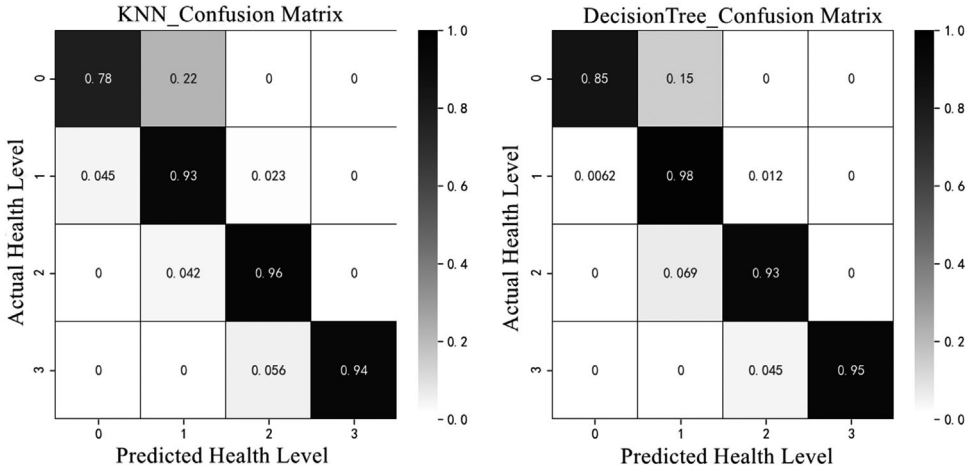

FIGURE 12 Comparison of confusion matrix

4.3 | Effect evaluation of the algorithm

In terms of the overall accuracy, the average accuracy of the KNN classification algorithm is 0.923 after considering the 5-Fold Cross Validation. The Decision Tree algorithm is 0.94. In terms of the evaluation index of each of the classification labels, Figure 11 shows that the evaluation index of the Decision Tree algorithm is higher than that of the KNN algorithm in each health category.

Figure 12 shows the normalized confusion matrix of the Decision Tree model and the KNN model. It can be observed that both the effect that is exerted by both the models is favourable. This is primarily because the maximum value is at the diagonal position. However, the Decision Tree model has a higher level of accuracy in the aspect of overall accuracy when it comes to judging the different health levels. Moreover, its values on the diagonal are higher than those observed in the KNN algorithm. Therefore, the Decision Tree algorithm tends to have a more enhanced effect.

When considering a comprehensive comparison of the various evaluation indexes that have been evaluated for the purpose of this study, it is evident that the classification effect

TABLE 8 Overall health level of Line 101

	Nov. 21	Nov. 22	Nov. 23	Nov. 24	Nov. 25
Morning peak	Unhealthy	Unhealthy	Sub-healthy	Unhealthy	Unhealthy
Off-peak	Healthy	Sub-healthy	Healthy	Healthy	Sub-healthy
Evening peak	Sub-healthy	Sub-healthy	Unhealthy	Sub-healthy	Unhealthy
Overall health level	Sub-healthy	Sub-healthy	Sub-healthy	Sub-healthy	Unhealthy

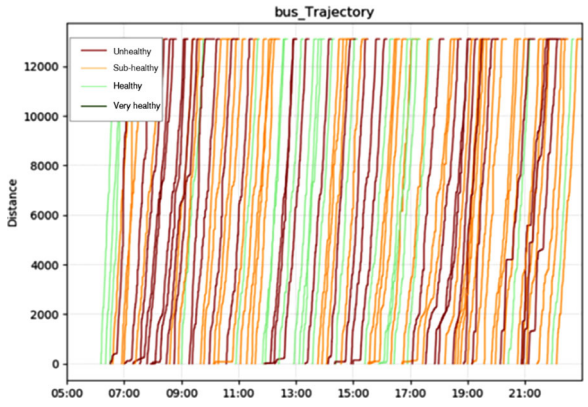

FIGURE 13 Health level judgment of Line 101

of Decision Tree algorithm is better than that of the KNN algorithm. Therefore, the Decision Tree algorithm exerts a better performance for the bus operation data of health classification.

4.4 | Case study: Health diagnosis of Line 101 in Foshan

Figure 13 shows the space-time diagram of bus operation, which reflects the health status of Line 101 in November 21, 2016. It can be observed that the operation health seems to be poor between 7:00-9:00 in the morning peak and 17:00-19:00 in the evening peak. Due to the unreasonable settings of the departure interval, this unhealthy operation also appears evident during the off-peak period. Moreover, the unreasonable setting thus leads to the small station passenger flow. It eventually leads to a situation where the bus does not stop at some stations, which affects the stability of the arrival time at the station. This then results in the large coefficient of variation of time headway and the coefficient of variation of the running speed between the stations.

In order to analyse the overall health level of Line 101, the average values of each evaluation index during the morning peak, off-peak and evening peak timings were calculated. Furthermore, the Decision Tree algorithm was used to judge its health level. The results are reflected in the Table 8.

If the bus operation is in a very healthy, or healthy level, it is not necessary to adjust the bus operation. But if the bus operation is at an unhealthy level, some adjustments have to be adapted. Similarly, if it is at a sub-healthy level, the adjustment can be postponed. In this regard, Table 8 shows that Line 101 is in sub-healthy or unhealthy level during the five days, especially during the morning peak and evening peak periods. It is evident that the operation levels are poor. As a result, the operations of the bus need to be adjusted and optimized accordingly.

5 | CONCLUSION

This paper constructed a health diagnosis system for single-line bus operation systems from efficiency and stability. The Python, ArcGIS and other data processing methods were applied to analyses the GPS and other data of bus operations. The KNN and the Decision Tree algorithm were also used in order to build the health diagnosis model. Firstly, the standard training set had been constructed by the Entropy method based on the selected typical lines. After that, the health diagnosis standard was also determined using the quantile method. By comparing the classification effects of the two algorithms, the Decision Tree algorithm came with a higher classification accuracy. So it was selected to diagnose the health level of each running trip of Line 101 in Foshan. Finally, we had demonstrated that the developed model can accurately and quickly determine the health level with data simulation experiments.

Applying in the thought of human health diagnosis, and bridge or tunnel health diagnosis, this paper introduces the artificial intelligence and machine learning algorithm into the bus operation diagnosis. It can effectively diagnose the health level of single-line bus operation systems. This study is conducted only on the basis of a large amount of easily available data, reducing the workload of the actual investigation. The method can reduce the cost of analysis and improve the efficiency of diagnosis. The results of this study have theoretical significance for enriching the existing theories of bus system operation analysis and diagnose.

ACKNOWLEDGEMENTS

The research of X.Z. and X.J. are supported by the National Natural Science Foundation of China (No. 61873190 and 71801138).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

DATA AVAILABILITY STATEMENT

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

ORCID

Xuemei Zhou https://orcid.org/0000-0001-6472-7557
Yusong Pang https://orcid.org/0000-0001-8094-3436

REFERENCES

1. Chi, L., Lei, L.: Fuzzy comprehensive evaluation for development status of urban bus rapid transit (brt) in china. App. Mech. Mater. 253–255, 1930–1935 (2013) https://www.scientific.net/AMM.253-255.1930
2. Shi, Q., Zhang, K., Weng, J., Dong, Y., Zhang, M.: Evaluation model of bus routes optimization scheme based on multi-source bus data. Transp. Res. Interdiscip. Perspect. 10(1), 100342 (2021) https://doi.org/10.1016/j.trip.2021.100342
3. Wang, C., Ma, C.: Multi-objective optimization of customized bus routes based on full operation process. Mod. Phys. Lett. B 34(2), 2050266 (2020) https://doi.org/10.1142/S0217984920502668
4. Annual report of Shanghai Comprehensive Transportation. Traffic Transp. 34(06), 10–12 (2018) http://doi.org/CNKI:SUN:YSJT.0.2018-06-008
5. Seaborn, C., Attanucci, J. & Wilson, N.H.M.: Analyzing multimodal public transport journeys in London with smart card fare payment data. Transportation Research Record 55–62 (2009)
6. Lu, B.C., Deng, J., Ma, Q.L., Liu, Q.F., Zhang, K.: A short-term public transit volume forecasting model based on IC Card and RBF neural network. J. Chongqing Jiaotong Univ. (Nat. Sci.) 34(06), 106–110 (2015) http://doi.org/10.3969/j.issn.1674-0696.2015.06.20
7. Wang, J.X.: The analysis of bus line OD extrapolation based on big data of bus IC card. Construct. Design Project 403(05), 167–8+71 (2019) http://doi.org/10.13616/j.cnki.gcjsysj.2019.03.054
8. Zhou, X., Cui, Y.F., Liu, Y.X.: Analysis and diagnosis of bus network based on big data. Commun. Shipping 6(04), 25–29 (2019) http://doi.org/10.16487/j.cnki.issn2095-7491.2019.04.006
9. Faouzi, N.E.E., Klein, L.A.: Data fusion for its: Techniques and research needs. Transp. Res. Procedia 15, 495–512 (2016) http://doi.org/10.1016/j.trpro.2016.06.042
10. Qu, X.H., An, G.: Summary and prospect of data fusion methods. Ship Electron. Eng. 2, 2–4+9 (2003) http://doi.org/10.3969/j.issn.1627-9730.2003.02.001
11. Zhao, Y.: Study on interval fusion algorithm of urban expressway based on multi-source detectors. Master's Thesis, Beijing Jiaotong University (2016) http://doi.org/10.7666/d.Y3126502
12. Ren, Y.C., Lin, H., Wang, Q.S., Li, F., Zhou, X., Liu, Y.X.: Optimization of Anyang public transport network based on multi-source big data analysis. Urban Public Transp. 261(03), 62–66+72 (2020)
13. Fonzone, A., Schmoecker, J.D., Liu, R.: A model of bus bunching under reliability-based passenger arrival patterns. Transp. Res. Part C Emerg. Technol. 59, 164–182 (2015) http://doi.org/10.1016/j.trc.2015.05.020
14. Sorratini, J.A., Liu, R., Sinha, S.: Assessing bus transport reliability using micro-simulation. Transp. Planning Technol. 31(3), 303–324 (2008) https://doi.org/10.1080/03081060802086512.
15. Chen, W.Y., Chen, Z.Y.: A simulation model for transit service unreliability prevention based on AVL-APC data. In: *2009 International Conference on Measuring Technology and Mechatronics Automation*, Zhangjiajie, China, pp. 184–188 (2009)http://doi.org/10.1109/ICMTMA.2009.77
16. Chen, X., Yu, L., Zhang, Y., Guo, J.: Analyzing urban bus service reliability at the stop, route, and network levels. Transp. Res. A 43(8), 722–734 (2009) http://doi.org/10.1016/j.tra.2009.07.006
17. Qu, X., Oh, E., Weng, J., Jin, S.: Bus travel time reliability analysis: A case study. Proc. Inst. Civ. Eng. Transp. 167(3), 178–184 (2014) http://doi.org/10.1680/tran.13.00009
18. Sun, C., Zhang, H.M., Chen, X.H.: Road traffic operation assessment based on multi-souce floating car data fusion. J. Tongji Univ. (Nat. Sci.) 46(01), 46–52 (2018) http://doi.org/10.11908/j.issn.0253-374x.2018.01.007
19. Camus, R., Longo, G., Macorini, C.: Estimation of transit reliability level-of-service based on automatic vehicle location data. Transp. Res. Rec. 1927(1), 277–286 (2005) http://doi.org/10.3141/1927-31
20. Hsu, S.C.: Determinants of passenger transfer waiting time at multi-modal connecting stations. Transp. Res. Part E Logist. Transp. Rev. 46(3), 404–413 (2010) http://doi.org/10.1016/j.tre.2009.12.002
21. Liu, J., Wang, H.L., Zhang, Z.G., Wu, L.P.: Dynamic fingerprint and identification algorithm for damage diagnosis of cable stayed bridge based on pattern recognition. J. Civil Archit. Environ. Eng. 38(04), 115–123 (2016) http://doi.org/10.11835/j.issn.1674-4764.2016.04.017
22. Ouyang, J.H., Liu, G.W.: Research progress on damage diagnosis methods of bridge structures. Guangdong Build. Mater. 25(08), 29–31 (2009) http://doi.org/10.3969/j.issn.1009-4806.2009.08.012
23. Kim, H.M., Bartkowicz, T.J.: An experimental study for damage detection using a hexagonal truss. Comput. Struct. 79(2), 173–182 (2001) http://doi.org/10.1016/S0045-7949(00)00126-7
24. Zhang, Z.P., Ma, Y., Yin, X.J.: Research on damage identification of health monitoring for bridge. Railway Comput. Appl. 19(06), 36–38 (2010) http://doi.org/10.3969/j.issn.1005-8451.2010.06.012
25. Zong, Z.H., Chu, F.P., Niu, J.: Damage identification methods of bridge structures using response surface based on finite element model updating. China Civil Eng. J. 46(02), 115–122 (2013) http://doi.org/CNKI:SUN:TMGC.0.2013-02-014
26. Hera, A., Hou, Z., Noori, M.: Wavelet-based approach for asce structural health monitoring benchmark studies. In: Proceedings of the 3rd International Workshop on Structural Health Monitoring, Stanford, p. 400 (2001)
27. Gao, L.: Application of genetic algorithm in optimal placement of arch bridge health monitoring sensors. China Comput. Commun. 14, 75–76 (2016)
28. Guo, H.Y., Li, Z.L.: Structural damage identification based on Bayesian theory and improved immune genetic algorithm. Expert Syst. Appl. 39(7), 6426–6434 (2012) http://doi.org/10.1016/j.eswa.2011.12.023
29. Beena, P., Ganguli, R.: Structural damage detection using fuzzy cognitive maps and Hebbian learning. Appl. Soft Comput. 11(1), 1014–1020 (2011) http://doi.org/10.1016/j.asoc.2010.01.023
30. Sun, Q., Du, B.Q., Jiang, J.J.: Study of health evaluation method for beam bridge based on multi-level fuzzy comprehensive evaluation method. World Bridge 45(2), 47–51 (2017) http://doi.org/CNKI:SUN:GWQL.0.2017-02-012
31. Yuan, S.F., Liang, D., Gao, N., Zhan, Y.F.: The bridge data diagnosis research based on structural health monitoring system. J. Univ. Electron. Sci. Technol. China 42(01), 69–74 (2013) http://doi.org/10.3969/j.issn.1001-0548.2013.01.016
32. Li, L., Liu, M.J.: Fault diagnosis study of bridge based on BP neural network. J. Univ. Jinan (Sci. Tech.) 23(04), 334–337 (2009) http://doi.org/10.3969/j.issn.1671-3559.2009.04.002
33. Zhang, C., Song, G.Q., Wang, M.H.: Structure damage identification by finite element model updating with Tikhonov regularization. J. Nanchang Univ. (Eng. Technol.) 32(04), 394–398 (2010) http://doi.org/10.3969/j.issn.1006-0456.2010.04.020
34. Pan, H., Azimi, M., Yan, F., Lin, Z.: Time-frequency based data-driven structural diagnosis and damage detection for cable-stayed bridges. J. Bridge Eng. 23(6), 04018033.1–04018033.22 (2018) http://doi.org/10.1061/(ASCE)BE.1943-5592.0001199
35. Sun, L.M., Shang, Z.Q., Xia, Y.: Development and prospect of bridge structure health monitoring in the context of big data. China J. Highw. Transp. 32(11), 1–20 (2019) http://doi.org/10.19721/j.cnki.1001-7372.2019.11.001
36. Buckley, J.J., Eslami, E.: Neural net solutions to fuzzy problems: The quadratic equation. Fuzzy Sets Syst. 86(3), 289–298 (1997) http://doi.org/10.1016/S0165-0114(95)00412-2
37. Lin, Y., Yang, X., Zou, N., Jia, L.: Real-time bus arrival time prediction: Case study for Jinan, China. J. Transp. Eng. 139(11), 1133–1140 (2013) http://doi.org/10.1061/(ASCE)TE.1943-5436.0000589
38. Chen, Y.Y., Cai, Y.W., Hou, Y.M., An, J., Li, X., Wei, P.Y.: "Healthy Index" evaluation method of metropolitan bus lines. J. Chang'an Univ. (Nat. Sci. Ed.) 35(S1), 1–6+16 (2015) http://doi.org/10.19721/j.cnki.1671-8879.2015.s1.001

39. Wang, C.: Study on evaluation system of road traffic health status in old urban area. Urban Roads Bridges Flood Control 12, 143–7+18 (2016). http://doi.org/10.16799/j.cnki.csdqyfh.2016.12.043
40. Zheng, C.J., Zhang, Y.H., Feng, X.J.: Improved iterative prediction for multiple stop arrival time using a support vector machine. Transport 27(2), 158–164 (2012) http://doi.org/10.3846/16484142.2012.692710
41. Ke, J.Y.: Transportation efficiency evaluation of public transportation lines based on improved DEA. In: *Proceedings of the 2018 China Urban Transport Planning Annual Conference*, Beijing, p. 14 (2018)
42. Lu, H.P., Yang, M., Zhang, Y.B.: On evaluation and counter measures of the integrated transportation hubs. Compr. Transp. 41(04), 25–30 (2019)
43. Deng, H.X., Zhao, Z.H., Wang, W.Q.: Research on bus OD estimation based on bus IC card and GPS data. J. Chongqing Univ. Technol. (Nat. Sci.) 33(6), 220–226 (2019) http://doi.org/10.3969/j.issn.1674-8425(z).2019.06.033
44. Tao, T.: Bus arrival time prediction based on GPS data. Master's Thesis, Changsha University of Science and Technology (2018)
45. Yuan, H.: Research on modeling spatiotemporal trajectories and spatiotemporal proximity analysis for time geographic analysis. Doctoral Thesis, Wuhan University (2019)
46. Yang, D.K.: Influence of different interpolation methods on GPS time series. GNSS World China 44(5), 66–69 (2019) http://doi.org/10.13442/j.gnss.1008-9268.2019.05.010

How to cite this article: Zhou, X., Guan, Z., Pang, Y., Ji, X., Lin, X.: Health diagnosis of bus operation based on multi-source data. IET Intell. Transp. Syst. 16, 754–768 (2022). https://doi.org/10.1049/itr2.12169