IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 30, 2019, accepted August 28, 2019, date of publication September 2, 2019, date of current version September 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2938851

Data-Driven Failure Characteristics and Reliability Analysis for Train Control On-Board Subsystem

BIN CHEN[1], BAIGEN CAI[1,2,3], (Senior Member, IEEE), WEI SHANGGUAN[1,2,3], (Member, IEEE), AND JIAN WANG[1,2,3], (Member, IEEE)

[1]School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
[2]Beijing Engineering Research Center of EMC and GNSS Technology for Rail Transportation, Beijing Jiaotong University, Beijing 100044, China
[3]State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China

Corresponding author: Wei Shangguan (wshg@bjtu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61490705 and Grant 61773049.

ABSTRACT The train control on-board subsystem (TC-OBS) plays an important role in the safety and efficiency of the high-speed train's operation. Therefore, there is an urgent demand for the analysis of failure characteristics and the reliability of TC-OBS. In this paper, a specific data model is built for the TC-OBS operational and failure data based on data cubes. This model analyzed the failure distribution characteristics of TC-OBS from the combined angles of System Identification, Time and Operation Attribute through the operations of data cubes. Thus, the representative units and systems can be the research objects of the reliability evaluation. With these representative units and systems, this paper uses Bayesian estimation combined with Markov Chain Monte Carlo (MCMC) to estimate the parameters of the time between failures (TBF) distribution model and the reliability is analyzed. Simulation results show that the data model based on data cubes can offer an efficient and convenient method to analyze the failure characteristics and reliability of TC-OBS.

INDEX TERMS Train control system, reliability, failure distribution characteristics, data modeling.

I. INTRODUCTION

The rapid development of high-speed railways offers a great development space for the economy and for people's livelihoods. Meanwhile, the operational safety of high-speed railways becomes extremely important. The train control on-board subsystem (TC-OBS) of the high-speed railway (HSR) is the core part for train controlling, and its safe and reliable operation is critical and basal for ensuring the safe and efficient operation of HSRs. As a complex and safety-critical system, the failure feature of the TC-OBS is very important to the maintenance and repair work as well as to the renewal of equipment. During the operation process, the performance degradation of the system caused by environmental interference, component wear, equipment aging, etc., will lead to system reliability degradation, causing failures, and creating hidden dangers to the safety of the HSR. When the TC-OBS is running, it will automatically generate log data of the running status and events of the system. The study of its data characteristics enables maintenance personnel to keep abreast of the system's operation, so they can timely repair and maintain the system to ensure the safety of the HSR's operation.

The processing of a system's data should be efficient and accurate for corresponding purposes. Huang and Zhou [1] proposed the structure, elements, basic calculations and multidimensional reasoning method of the new knowledge model for electric power based on ontology and the semantic web. The work in [2] introduced random matrix theory to model the massive data sets for power equipment monitoring and big data mining analytics. Wang and Bai [3] built a fuzzy spatiotemporal data model by expanding the standard modeling language UML; thus, the model can describe the fuzzy spatiotemporal objects better, and it is dynamic. After data modeling, the operational status of the system can be efficiently acquired.

The system's failure feature is a mathematical description that can represent the processes of fault distribution and evolution. Studies on fault features play an important role in fault diagnosis, reliability analysis, and maintenance strategies.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhigang Liu.

Many scholars have recently carried out research on fault features in the fields of power transmission systems, military science, computer networks, and software engineering. The work in [4] analyzed the time trend of the failure frequency in communication networks using the cumulative number of failures and estimated the relationship between failure frequency and severity by the generalized Pareto probability distribution. Zheng *et al.* [5] collected the fault recording from one converter station when the communication failed due to a disturbance in the AC system and calculated and statistically analyzed the turn-off angle of valves, zero-crossing offset, and landing amplitude of bus voltage. Jager *et al.* [6] proposed an integration step that evaluates the failure model of shared information in relation to an application's fault tolerance and presented a mathematically defined generic failure model as well as a processing chain for automatically extracting failure models from empirical data. The work in [7] proposed a risk index system for the catenary lines of high-speed railways considering the characteristics of time-space differences to represent and quantify the characteristics of risk. Arno *et al.* [8] analyzed the relationship between equipment failure characteristics and reliability-based maintenance to optimize preventative maintenance regime as well as a corrective maintenance regime.

Few researchers focus on the data modeling of train control system operational data. However, the application of systems' operational data has been studied widely by many scholars in the railway field, and they usually focus on the fault diagnosis, failure prediction and reliability evaluation of train control systems.

In the areas of fault diagnosis and failure prediction, researchers mainly focus on the accurate and rapid detection of faults that are about to happen or have happened in each part of the train control system such as the TC-OBS, track circuit and so on. Ding *et al.* [9] proposed a method based on fuzzy rules and a time series analysis for the online failure prediction of the Automatic Train Protection system. Zhao *et al.* studied fault diagnosis methods for track circuit [10], [11]. Bruin *et al.* [12] used the long short-term memory recurrent neural network to accomplish the timely detection and identification of faults in railway track circuits for the safety and availability of the railway network. In [13], Wang *et al.* proposed a bilevel feature extraction-based text mining that integrates features extracted at both the syntax and semantic levels with the aim of improving the fault classification performance of the train control system. Similarly, many researchers have conducted their research on fault diagnosis through text mining and big data analysis [14]–[17].

As for the application of train control system operational data in reliability evaluations, research carried out according to the different forms of equipment or system level of the train control system. Sun *et al.* [18] proposed a life prediction method for analyzing and assessing the reliability of railway safety relays by blending the principal component analysis to extract the key degradation features of relays. Xu *et al.* [19] proposed an online performance degradation monitoring approach for the onboard speed sensors of trains and provided a compensation algorithm for the distorted speed readings resulting from the existence of the performance degradation. Zhu *et al.* [20] modeled the next generation communication-based train control (CBTC) system with deterministic and stochastic Petri nets (DSPNs), and the performance data were converted as DSPN model parameters to evaluate the system reliability. At the system level, Su and Che [21] introduced Bayesian networks for the limitations of the traditional fault tree analysis method to establish the reliability model of train control systems, and the reliability of the train control system and its redundant configuration was assessed. Additionally, Morant *et al.* [22] and Pascale *et al.* [23] conducted their research studies on reliability evaluations and maintenance strategies for railway signaling systems.

Operational data and failure characteristics are important to the reliability as well as the maintenance strategies for the TC-OBS. However, research on the fault diagnosis and reliability evaluation for TC-OBSs are carried out based on a small amount of data. It is necessary to take full advantage of the operational data of TC-OBSs to analyze the failure distribution characteristics as well as to evaluate its reliability.

The rest of this paper is organized as follows. In Section II, we establish a specific model for the operational data of TC-OBSs based on data cubes to analyze the failure characteristics and provide the necessary data for the reliability evaluation. In Section III, the failure distribution characteristics are analyzed using the data model of the TC-OBS, and in Section IV the reliability of representative equipment and systems, which are selected by the failure distribution characteristics is estimated by Bayesian estimation. Finally, conclusions and recommendation for future works are given.

II. DATA MODEL FOR TRAIN CONTROL ON-BOARD SUBSYSTEM

In this section, we build a specific model for the operational data of the TC-OBS based on the data cube to analyze the failure characteristics and reliability of the TC-OBS.

Figure 1 shows the partial structure of the TC-OBS. It includes several units working together to control the high-speed train. While in operation, the working status of the units in the TC-OBS are influenced by various factors that could lead to failures, or even disasters.

The partial operational data of the TC-OBS is shown in Figure 2. Most of the data are recorded by text formatting which is difficult to process because of the misunderstanding of different maintainers and the complexity of the operational data. Thus, an effective model is needed for the analysis of the operational data.

A complete data record contains three key elements: *System Identification*, *Time*, and *Operation Attribute*. *System identification* represents where the data belongs. *Time* represents the time when the system operates, and *Operation Attribute* represents the operational status, including the normal operational status and types of failures. To process and

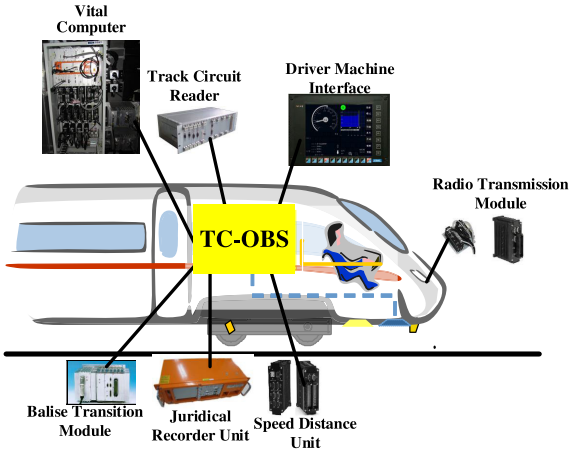

FIGURE 1. Partial structure of TC-OBS.

```
1 17-04-08 03:36:19,220 FID:    tih_a.cpp LID:  2754 TID:SMGM_LogTask
   000042160     ATP A A1C2R000 Vehicle Ready Active (Lifesign 8 OK).

2 17-04-08 03:36:18,228 FID:    tih_a.cpp LID:  2005 TID:SMGM_LogTask
   000041562     ATP A A1B1R720 Alive Lifesign 2.

3 17-04-08 03:36:17,706 FID:    tih_a.cpp LID:  1999 TID:SMGM_LogTask
   000040963     ATP A A1B0R710 Alive Lifesign 1.

4 17-04-08 03:36:17,120 FID:    stmm.cpp LID:  5127 TID:SMGM_LogTask
   000040347     ATP A 4A13R000 STM connection running.

5 17-04-08 03:36:16,620 FID:tih_a_stmdat LID:   461 TID:SMGM_LogTask
   000039751     ATP A A069D0B0 Profibus TI-H to TSG connected.

6 17-04-08 03:36:16,120 FID:bih_a_stmdat LID:   468 TID:SMGM_LogTask
   000039148     ATP A A4C9D0B0 Profibus BI-H to TSG connected.

Hit any key to continue or ESC to stop ...

7 17-04-08 03:36:15,620 FID:wi_a_radioCh LID:  1677 TID:SMGM_LogTask
   000038552     ATP A 5207I000 Remove error: Fewer modems registered than installe

8 17-04-08 03:36:14,623 FID:wi_a_radioCh LID:  1667 TID:SMGM_LogTask
   000037956     ATP A 5207I000 Fewer modems registered than installed 1
```

FIGURE 2. Partial operational data of TC-OBS.

analyze the data efficiently, we build a specific data model based on the data cube.

An n-dimensional data cube [24] is a quadruple shown in (1), where D is the set of dimensions of the data cube, and H is the hierarchical set of dimensions. M is a set of measures of the data cube, and Γ is the aggregate function for the measures.

$$N = \{D, H, M, \Gamma\} \tag{1}$$

The operational data of the TC-OBS has three dimensions as described above: that is, *System Identification*, *Time*, and *Operation Attribute*. Therefore, the set of dimensions D is shown as (2), where S represents *System Identification*, T represents *Time*, and O represents *Operation Attribute*.

$$D = \{S, T, O\} \tag{2}$$

Because H is the hierarchical set of dimensions, in this case, the hierarchical set is shown as (3), where H_S, H_T, H_O represent the hierarchical set of dimension S, T and O, respectively. H_S and H_O represent the coordinating relation in systems and operational status, respectively. T represents the concept hierarchy of time such as *year* $\leftarrow$ *month* $\leftarrow$ *day*.

$$H = \{H_S, H_T, H_O\} \tag{3}$$

TABLE 1. Meanings of abbreviations for fault types.

Abbreviations	Meanings
BTM	Balise Transition Module
DMI	Driver Machine Interface
JRU	Juridical Recorder Unit
TCR	Track Circuit Reader
TSG	Train Signalling Gateway
VC	Vital Computer
VDX	Vital Digital Input/output Unit
SDU	Speed Distance Unit
RE	Relay
COM	Communication
BR	Breaking
others	other types of failure

The set of measures M is the number of the specific operational status, and its element is expressed by $m_{s,t,o}$ where $s \in S, t \in T$ and $o \in O$. Furthermore, the sum of $m_{s,t,o}$ represents the number of operations of systems. The aggregate function Γ has different forms depending on different analytic targets such as the sum function.

The dimension of *Operation Attribute* includes the normal operational status and several types of failures. We use O_n and O_f to represent these two sub-dimensions, respectively. Similarly, H_O includes H_{O_n} and H_{O_f}, and $m_{s,t,o}$ includes m_{s,t,o_n} and m_{s,t,o_f}.

Thus, we can get the failure data cube for the TC-OBS shown in equation (4).

$$N_f = \{D_f, H_f, M_f, \Gamma\} \tag{4}$$

where $D_f = \{S, T, O_f\}$, $H_f = \{H_S, H_T, H_{O_f}\}$, M_f is the set of m_{s,t,o_f} and $o_f \in O_f$.

According to the practical maintenance manual, the sub-dimensions O_f includes 12 types such as the Balise Transition Module type and the Communication type, and they are shown in Table 1 along with their abbreviations.

Here is an illustration of the data cube for the operational data of the TC-OBS shown in Figure 3. There are three dimensions in Figure 3, and the content in each cuboid is the number of the specific operational status in a specific time, system and failure type. The S dimension includes the different train numbers in which the TC-OBS is located. T dimension includes the time of the specific operation. Finally, the O dimension includes different types of failures as well as the normal operational status. All the cuboids represent the failure number and normal operation times in all three dimensions.

There are two important operations of the data cube: slice and dice. Slice selects just one specific dimension such as S or T or A in operational data cubes. Dice selects two or more dimensions from the data cube. Both operations provide a new subcube of the original data cube. We operate three slices from the three dimensions of the failure data cube and operate three dices from the combinations of these three dimensions. Then, we get the hierarchical directed graph structure of the data cube for the TC-OBS as shown in Figure 4.

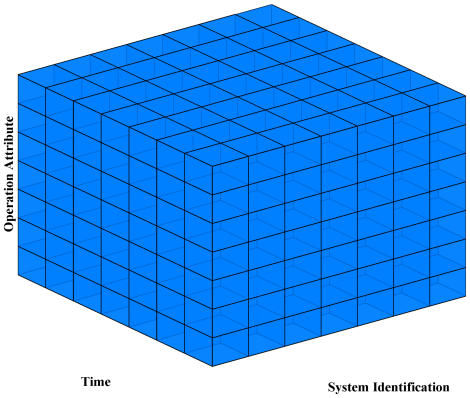

FIGURE 3. Illustration of data cube for operational data of TC-OBS.

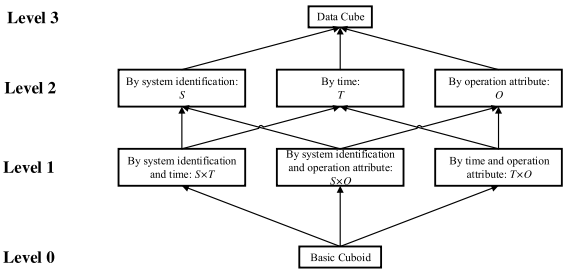

FIGURE 4. Hierarchical directed graph structure.

Along with the hierarchical directed graph structure and different demands on the analysis, the failure distribution characteristics can be extracted from different perspectives. For example, because the operation attribute includes different types of failures, we can get the distribution of different types of failures in the dimension of time from the node by *Time and Operation Attribute*, expressed as $T \times O$.

In this paper, we process the data from a specific type of TC-OBS in a railway bureau of China in 2015, and the data consists of more than 3000 records including 12 types of failures and more than 100 systems. The analysis in this paper is based on these data.

Figure 5 illustrates the flowchart of the processing and analysis of the TC-OBS data in this paper. First, the raw data are preprocessed to delete useless information, and the three key elements are reserved. Then, the data model of the TC-OBS is established based on the data cube. Next, the failure characteristics of the TC-OBS by different dimensions are analyzed by applying the dice operation of the data cube, and the representative units and systems are selected to evaluate the reliability of the TC-OBS. Finally, the reliability is evaluated by Bayesian estimation and the Markov Chain Monte Carlo (MCMC) method. This method improves the comprehensiveness and veracity of the failure distribution characteristics and reliability evaluation in this data-driven way and provides a new effective way for the maintenance strategy of TC-OBSs.

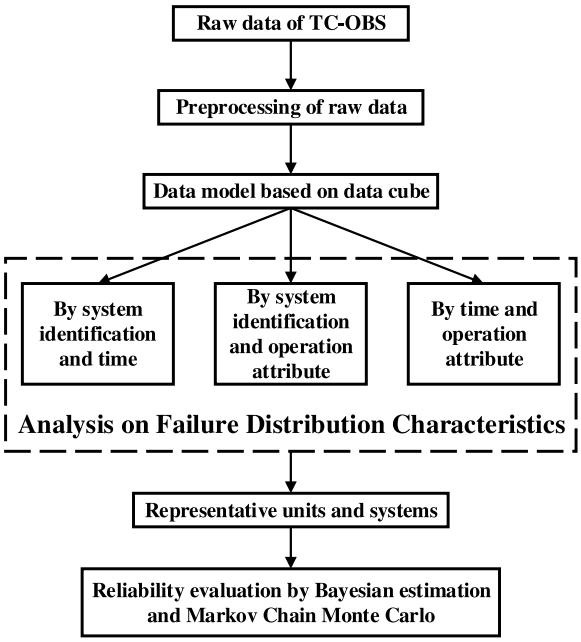

FIGURE 5. Flowchart of the process and analysis of TC-OBS data.

III. DATA-DRIVEN FAILURE DISTRIBUTION CHARACTERISTICS ANALYSIS OF TC-OBS

The failure distribution characteristics are presented in many ways such as the distribution of failure rate by time. It is difficult to cover all types of distribution characteristics unless there is an efficient way to tease these characteristics from all of the dimensions. The operational data cube mentioned above is one efficient way to do that.

As shown in Figure 4, Level 1 of the hierarchical directed graph structure contains the combinations of the three dimensions that could be used as the basis for the classification of the failure characteristics. Additionally, many extensions could be obtained from these combinations.

A. FAILURE DISTRIBUTION CHARACTERISTICS BY SYSTEM IDENTIFICATION AND TIME

The data subcube by *System Identification* and *Time* is a quadruple shown in (5).

$$Sc_{ST} = \{D_{ST}, H_{ST}, M_{ST}, \Gamma_{ST}\} \tag{5}$$

where $D_{ST} = \{S, T\}$, $H_{ST} = \{H_S, H_T\}$, $M_{ST} = \{for\ (s,t) \in (S,T) | m_{s,t} = \sum_{i=1}^{n_O} m_{s,t,o}\}$ and n_O is the sum of dimension O. Γ_{ST} is the same as Γ in (1).

As mentioned above, dimension O includes O_n and O_f. Therefore, Sc_{ST} includes two parts: the normal operation part and the failure part. Based on equation (4), the failure data subcube is a quadruple shown in (6).

$$Sc_{ST_f} = \{D_{ST_f}, H_{ST_f}, M_{ST_f}, \Gamma_{ST_f}\} \tag{6}$$

where $D_{ST_f} = D_{ST}$, $H_{ST_f} = H_{ST}$, $\Gamma_{ST_f} = \Gamma_{ST}$, $M_{ST_f} = \{for\ (s,t) \in (S,T) | m_{s,t_f} = \sum_{i=1}^{n_{O_f}} m_{s,t,o_f}\}$ and n_{O_f} is the sum of dimension O_f.

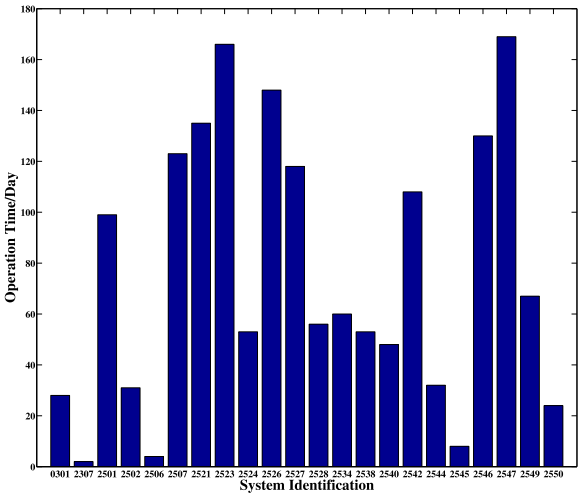

FIGURE 6. Operation time in 2015.

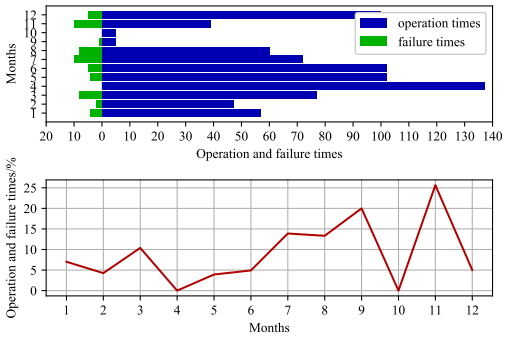

FIGURE 7. Monthly analysis on operation and failure for system 2547 in 2015.

Because M represents the set of specific operational statuses including the failures of systems, the number of operations and the failure of each system can be calculated through equation (7) and equation (8) by making $m_{s,t} \in M_{ST}$.

$$N_s = \{for\ s \in S | \sum_{i=1}^{n_T} m_{s,t}\} \tag{7}$$

$$N_{sf} = \{for\ s \in S | \sum_{i=1}^{n_T} m_{s,t_f}\} \tag{8}$$

Figure 6 shows the operational time of partial systems. It shows that the operational times of different systems differ widely. For example, the operating times of System 2547 and System 2523 are much longer than that of System 2307. Different operational times leads to the different working strengths of different systems. The longer the operating time the system has, the more information it contains.

According to Figure 6, we make $s = 2547$, and the number of operational and failure statuses of the system can be calculated through equation (7) and equation (8) in which the time dimension is measured in months as well as the failure rate, which is shown in Figure 7.

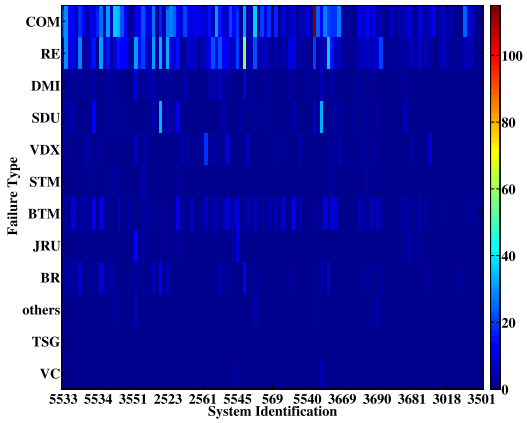

FIGURE 8. Heatmap of failure types.

From Figure 7 we can see that there is no obvious dependence between the number of failures and time, and the failure rate is approximately 10% in months when the systems ran more than 50 times. The failure rate of a few months is higher than others because the numbers of operation times are fewer in these months, and the numbers of failures are at a relatively low level.

B. FAILURE DISTRIBUTION CHARACTERISTICS BY SYSTEM IDENTIFICATION AND OPERATION ATTRIBUTE

In this section, we analyze the failure distribution from the angle of system identification and operation attribute, especially in failure types.

Similarly, the data subcube by *System Identification* and *Operation Attribute* is a quadruple shown in (9).

$$Sc_{SO} = \{D_{SO}, H_{SO}, M_{SO}, \Gamma_{SO}\} \tag{9}$$

where $D_{SO} = \{S, O\}$, $H_{SO} = \{H_S, H_O\}$, $M_{SO} = \{for(s, o) \in (S, O) | m_{s,o} = \sum_{i=1}^{n_T} m_{s,t,o}\}$ and n_T is the sum of dimension T. Γ_{SO} is the same as Γ in (1).

Because $O = O_f \bigcup O_n$, we can get that $M_{SO} = M_{SO_f} \bigcup M_{SO_n}$, where M_{SO_n} is the set of m_{s,o_f} which is the number of failures for a specific system and failure type. Thus, we can draw the heatmap for the failure rate of each type of failure by $m_{s,o_f}/m_{s,o}$ as shown in Figure 8.

Figure 8 shows the types of communications and relays that occur frequently in most trains, and BTM failures are distributed uniformly. These types of failures can be classified as frequent faults. In addition, it can be seen that the numbers of SDU failures occurring in two of the systems (they are actually system 201 and system 2521) are far greater than in other systems, to which more attention should be paid.

In addition, we can calculate the percentage of systems in which the specific type of failures occurred among all the systems through equation (9). The result is shown in Figure 9.

Figure 9 shows that communication, relay, BTM and DMI type failures commonly occurred in most systems, while TSG and VC type failures occurred only in a small number

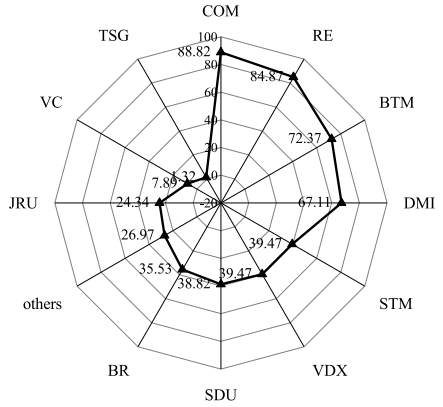

FIGURE 9. Proportions of failure types.

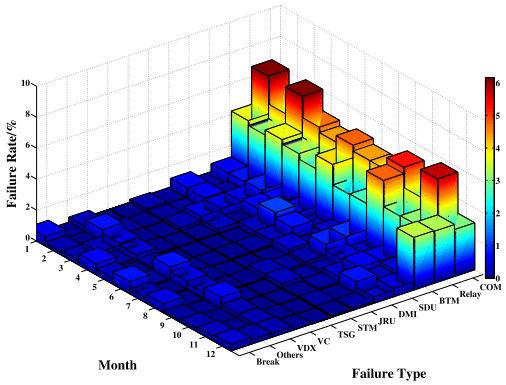

FIGURE 10. Time distribution of failure types.

of systems. This result can provide a reference for system maintenance.

C. FAILURE DISTRIBUTION CHARACTERISTICS BY TIME AND OPERATION ATTRIBUTE

The data subcube by *Time* and *Operation Attribute* is a quadruple shown in (10).

$$Sc_{TO} = \{D_{TO}, H_{TO}, M_{TO}, \Gamma_{TO}\} \tag{10}$$

where $D_{TO} = \{T, O\}$, $H_{TO} = \{H_T, H_O\}$, $M_{TO} = \{for(t, o) \in (T, O) | \sum_{i=1}^{n_S} m_{s,t,o}\}$ and n_S is the sum of dimension S. Γ_{TO} is the same as Γ in (1).

Figure 10 shows the distribution of different types of failures by time based on the subcube by *Time* and *Operation Attribute* in which the normal operating status is not selected. It clearly shows that the failure rates of communication and relay were always at a high level in 2015. The other types of failures have a lower failure rate except for the BTM type in December. It is important to emphasize the equipment for the communication and relays as well as the BTM in December to analyze the reason for the higher failure rate.

D. SUMMARY OF THE FAILURE DISTRIBUTION CHARACTERISTICS

In Subsections III-A, III-B and III-C, we analyzed the failure distribution by taking full advantage of the operational data from TC-OBSs and we obtained different characteristics on the operational and fault status of the TC-OBS. However, it is difficult to analyze the reliability of all the systems as well as the different units in the system in a shorter period of time because of the massive quantity and scale of the operational data. We take the analysis on failure distribution characteristics that uses the full operational data as the basis of the reliability analysis, and it helps us to determine which subsystems have enough data to ensure an accurate reliability analysis.

1) The operating times of different systems differ greatly as do the number of failures. The longer the system operates, the more failure information the system contains, which makes the reliability evaluation more representative.
2) Different types of equipment have different failure times and failure rates, either in different systems or at different times. It is important to put more focus on the equipment with more failure times and higher failure rates.

Based on the above subsections, we select communication and relay equipment in 3 systems (indicated as Sys. 1, Sys. 2 and Sys. 3) to analyze their reliability characteristics.

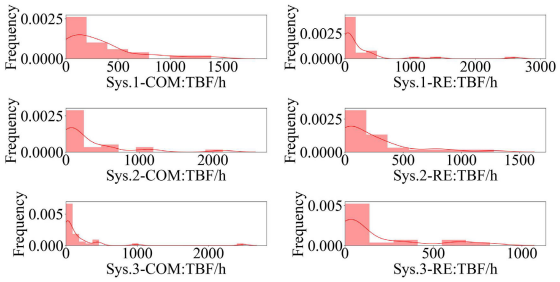

FIGURE 11. Histogram and kernel density estimation for selected data.

IV. RELIABILITY EVALUATION BASED ON FAILURE CHARACTERISTICS

A. BAYESIAN ESTIMATION FOR DISTRIBUTION FITTING

As mentioned above, we select communication and relay equipment in Sys. 1, Sys. 2 and Sys. 3 to analyze their reliability characteristics for the purpose of conducting a reliability analysis on this type of TC-OBS. The data is processed to obtain the time between failures (TBF) for these selected data, and their histogram and kernel density estimation are shown in Figure 11, which indicates that the TBF roughly follows log-normal or Weibull distribution. In the field of reliability, the life data of systems, especially electronic systems, mostly follow an exponential distribution or a Weibull distribution. Therefore, the Weibull distribution is chosen as the distribution for the TBF.

The probability density function (PDF) and cumulative distribution function (CDF) of the Weibull distribution are shown in equations (11) and (12), respectively.

$$f(t) = \lambda\beta t^{\beta-1} \exp(-\lambda t^{\beta}), t \geq 0 \tag{11}$$

$$F(t) = 1 - \exp(-\lambda t^{\beta}), t \geq 0 \tag{12}$$

Although classical statistical techniques such as the maximum likelihood estimation (MLE) have been developed to estimate the parameters of distributions, they do not work well with small-to-moderate sample sizes. In this study, the amount of failure data for each unit is small. In contrast, Bayesian estimation has an advantage in small-to-moderate sample sizes. It combines the sample information and prior information and has the ability to update the prior information through posterior information and observational data. Therefore, we chose the Bayesian estimation to conduct the parameter fitting for the Weibull distribution. The Bayes theorem is shown as equation (13):

$$p(\theta|y) = \frac{f(\theta|y)p(\theta)}{\int f(y|\theta)p(\theta)\mathrm{d}\theta} \tag{13}$$

where $p(\theta|y)$ is the posterior density function, $p(\theta)$ is the prior density function and $f(\theta|y)$ is the sampling density function for y. After the test, the value of y is certain and the sampling density function is the function of unknown parameter θ which is also called likelihood function.

The prior distributions for λ and β in equation (11) are usually set as the gamma distribution and the log-normal distribution shown as equations (14) and (15), and their probability density functions are shown as equations (16) and (17).

$$\lambda \sim Gamma(\alpha, \theta) \tag{14}$$

$$\beta \sim LogNormal(\mu, \sigma^2) \tag{15}$$

$$f_g(\lambda|\alpha, \theta) = \frac{\theta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}\exp(-\theta\lambda) \tag{16}$$

$$f_{ln}(\beta|\mu, \sigma^2) = \frac{1}{\beta\sqrt{2\pi\sigma^2}}\exp[-\frac{1}{2\sigma^2}(\ln(\beta) - \mu)^2] \tag{17}$$

With the independence of these two parameters, the joint prior density function is shown as equation (18).

$$p(\lambda, \beta) = f_g(\lambda|\alpha, \theta)f_{ln}(\beta|\mu, \sigma^2) \tag{18}$$

Considering a random sample consisting of n observations $(x_1 < x_2 < \ldots < x_n)$, when equation (11) is the density function, the likelihood function of this sample is:

$$L(x_1, x_2, \ldots, x_n|\lambda, \beta) = \prod_{i=1}^{n} \lambda\beta {x_i}^{\beta-1}\exp(-\lambda {x_i}^{\beta}) \tag{19}$$

The relationship among the posterior distribution (PO), the prior distribution (PR) and the likelihood function (LF) is as shown:

$$PO \propto PR \times LF \tag{20}$$

Therefore, the joint posterior distribution density function is determined by equation (21).

$$p(\lambda, \beta|x_1, x_2, \ldots, x_n) = p(\lambda, \beta) \cdot L(x_1, x_2, \ldots, x_n|\lambda, \beta) \tag{21}$$

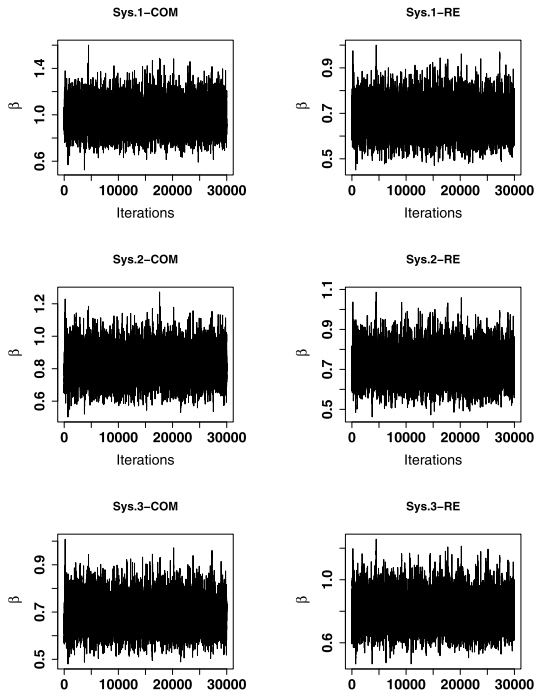

FIGURE 12. Trace of β for 6 groups of data.

Obviously, it is extremely difficult to obtain the joint posterior distribution density function in (21). Because of the difficulty of the analytical calculation for the Bayesian estimation, we used the MCMC to produce samples from the posterior distribution that can be used as an approximation of the probability distribution. In this paper, we use the MCMC method containing Metropolis algorithms to generate samples of model parameters from equation (21) and to carry out the sample-based posterior analysis based on these generated posterior samples.

B. NUMERICAL TEST

1) DATA PREPARATION

The sets of time between failures of the communication and relay equipment in Sys. 1, Sys. 2 and Sys. 3 are the data used to do the numerical test.

2) PARAMETERS FITTING

The MCMC method containing Metropolis algorithms is used to implement the sampling procedure, and 35000 samples are generated from the joint posterior distribution (21) with 5000 samples for burn-in. Figure 12 and Figure 13 are the traces of β and λ. They clearly show that the traces of all the parameters are indistinguishable from each other, which means that the iterative process became stable.

Figure 14 and Figure 15 are the autocorrelation functions (ACF) of each parameter, and it can be seen that the ACF tends to be zero after several iterations and the degree of

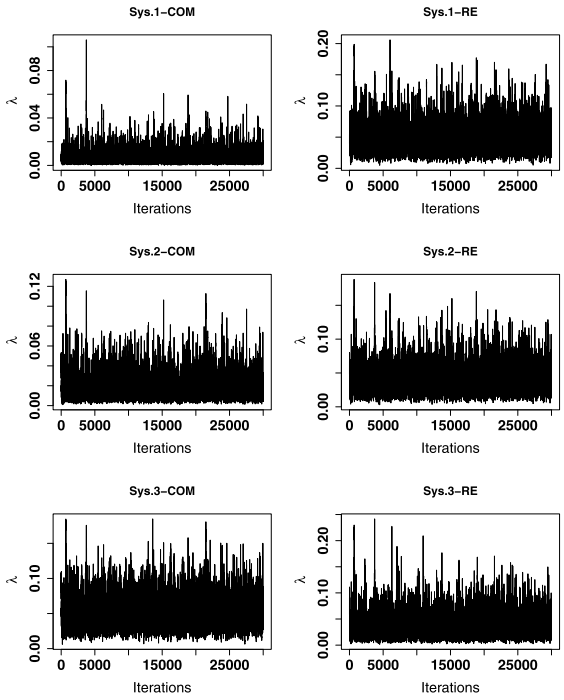

FIGURE 13. Trace of λ for 6 groups of data.

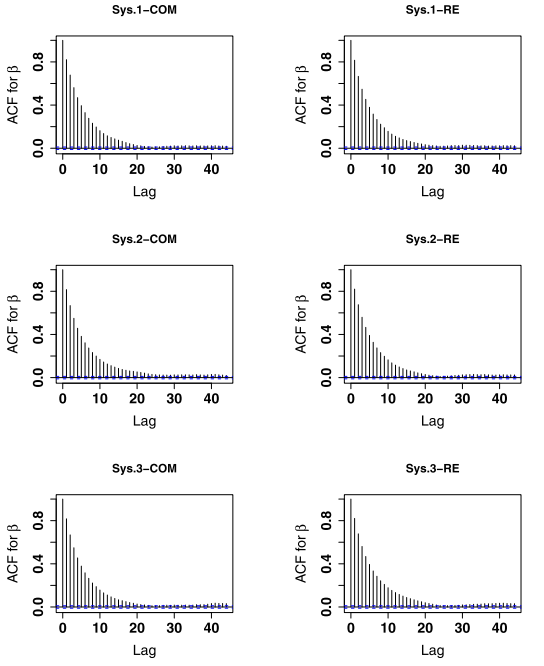

FIGURE 14. ACF of β for 6 groups of data.

autocorrelation is very low which means that the convergence property is excellent.

Table 2 shows the posterior quantities of the parameters as well as the posterior confidence intervals. As we can see from Table 2, the standard deviations of the parameters are at a low level, which means that the precision of the posterior mean values meets the requirements. Combining the analysis on the

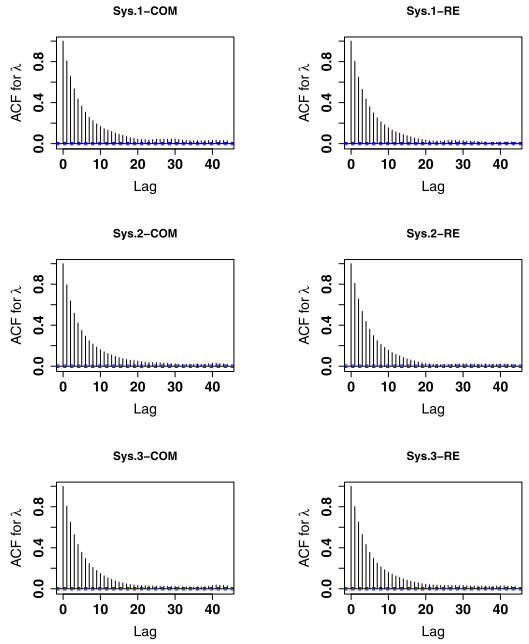

FIGURE 15. ACF of λ for 6 groups of data.

TABLE 2. Results for parameters fitting.

Units	Param	Mean	SD	2.5%	97.5%
Sys.1-COM	λ	0.00633	0.00599	0.00080	0.02251
	β	0.98500	0.12428	0.75161	1.24260
Sys.1-RE	λ	0.04480	0.02232	0.01474	0.10021
	β	0.68393	0.07044	0.55055	0.82860
Sys.2-COM	λ	0.01687	0.01204	0.00332	0.04870
	β	0.82613	0.09711	0.64651	1.02544
Sys.2-RE	λ	0.03610	0.01957	0.01076	0.08321
	β	0.72372	0.08063	0.57798	0.88747
Sys.3-COM	λ	0.04762	0.02128	0.01745	0.09904
	β	0.69599	0.06723	0.57075	0.83456
Sys.3-RE	λ	0.03345	0.02277	0.00691	0.09168
	β	0.80349	0.10373	0.61622	1.02154

trace and the ACF of the parameters, we can conclude that the MCMC method has a strong convergence; in other word, the results of the parameter fitting are sufficiently accurate. The densities of λ and β for the 6 groups of data are shown in Figures 16 and 17, respectively.

After fitting the distribution, the hypothesis test should be done on the distribution model. In this study, mean value of each parameter is used to conduct the hypothesis test with the Kolmogorov-Smirnov (K-S) test method. First, we sort the TBFs from smallest to largest and calculate the value of the cumulative distribution function $F(t_i)$ and the cumulative sample frequency $F_n(t_i)$ on every data element. The test statistics are then calculated by equation (22).

$$D_n = \sup_{t_i} |F_n(t_i) - F(t_i)| \tag{22}$$

Table 3 shows the results of the K-S test of the parameters. D_B and ρ_B are the results of the K-S test method with the threshold values D_ε and $\rho_\varepsilon = 0.05$. If $D_B < D_\varepsilon$ and $\rho_B > \rho_\varepsilon$,

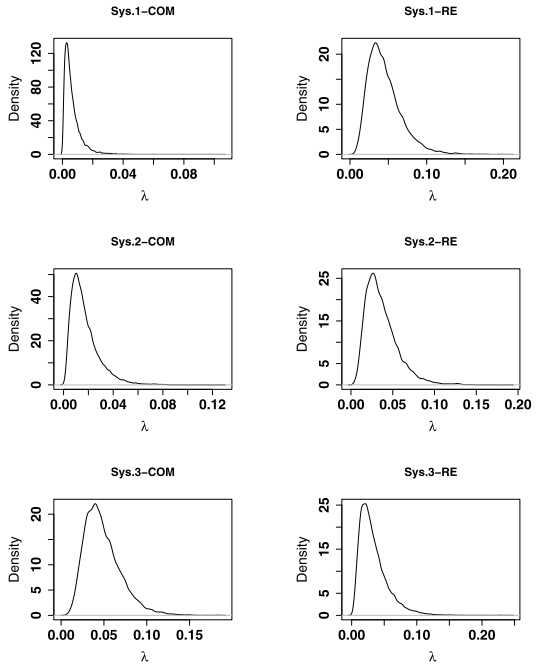

FIGURE 16. Density of λ for 6 groups of data.

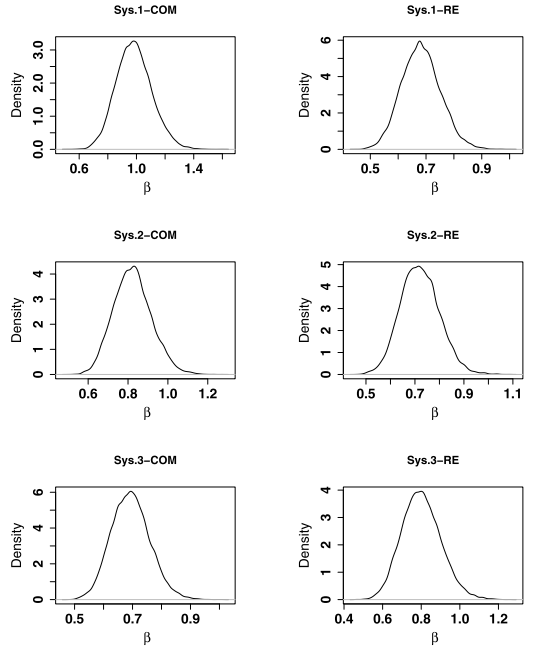

FIGURE 17. Density of β for 6 groups of data.

then the distribution fitting holds. The results clearly show that all the selected data follow the Weibull distribution.

In addition, we applied the MLE to estimate the parameters of the Weibull distribution as a comparison. The parameters estimated by the MLE are $\hat{\lambda}_{MLE}$ and $\hat{\beta}_{MLE}$ as shown in Table 4. Furthermore, D_{MLE} and ρ_{MLE} are the results of the K-S test for the MLE and are also shown in Table 4. Comparing D_B and ρ_B from Table 3 with D_{MLE} and ρ_{MLE}

TABLE 3. Results of hypothesis test for Bayesian estimation.

Units	$\hat{\lambda}$	$\hat{\beta}$	D_B	ρ_B	D_ε
Sys.1-COM	0.00633	0.98500	0.16856	0.4292	0.326
Sys.1-RE	0.04480	0.68393	0.14519	0.4859	0.293
Sys.2-COM	0.01687	0.82613	0.16178	0.4805	0.326
Sys.2-RE	0.03610	0.72372	0.19341	0.1481	0.284
Sys.3-COM	0.04762	0.69599	0.11569	0.6781	0.272
Sys.3-RE	0.03345	0.80349	0.16129	0.5898	0.356

TABLE 4. Results of hypothesis test for MLE.

Units	$\hat{\lambda}_{MLE}$	$\hat{\beta}_{MLE}$	D_{MLE}	ρ_{MLE}
Sys.1-COM	0.01118	0.79373	0.17025	0.4191
Sys.1-RE	0.05759	0.56673	0.1689	0.3041
Sys.2-COM	0.02434	0.67792	0.16245	0.4709
Sys.2-RE	0.05402	0.57827	0.20547	0.1322
Sys.3-COM	0.05821	0.59592	0.13451	0.491
Sys.3-RE	0.04381	0.65297	0.1857	0.4135

from Table 4, we can see that for each unit, D_B is always smaller than D_{MLE}, and ρ_B is always bigger than ρ_{MLE}, which means the Bayesian estimation is better than MLE in this case.

3) RELIABILITY ANALYSIS

In this section, we chose the degree of reliability, failure rate and mean time between failures (MTBF) as the indicators for reliability. The degree of reliability represents the trend for reliability over time. The failure rate is the frequency at which the system or equipment fails per unit of time, and the MTBF is the average of the interval between two adjacent faults of the system, which is also one of the most important indicators for the system's reliability. The degree of reliability, failure rate, and MTBF are calculated by equations (23), (24) and (25), respectively.

$$R(t) = \exp(-\lambda t^{\beta}) \tag{23}$$

$$r(t) = \lambda\beta t^{\beta-1} \tag{24}$$

$$MTBF = \lambda^{-\frac{1}{\beta}}\Gamma\left(1+\frac{1}{\beta}\right) \tag{25}$$

where Γ is the Gamma function.

Figure 18 shows the reliability function (97.5% confidence interval) of the 6 units in Table 3. The reliability declines smoothly, and different units have different descent velocities. The time when the posterior median reliability of Sys. 1-COM and Sys. 2-COM drops to 50% is more than 100 hours, while the same time of the other units is less than 100 hours. This finding indicates that different units in different systems have different reliability characteristics, and thus, the maintenance work should be carried out with each individual.

The failure rate function decreases monotonously over time and levels off, which means that these units are in an early failure period and a random failure period as shown in Figure 19. Additionally, it can be seen that the failure rate of Sys. 1-COM is lower than any of the other units, which indicates that the reliability of Sys. 1-COM is the highest.

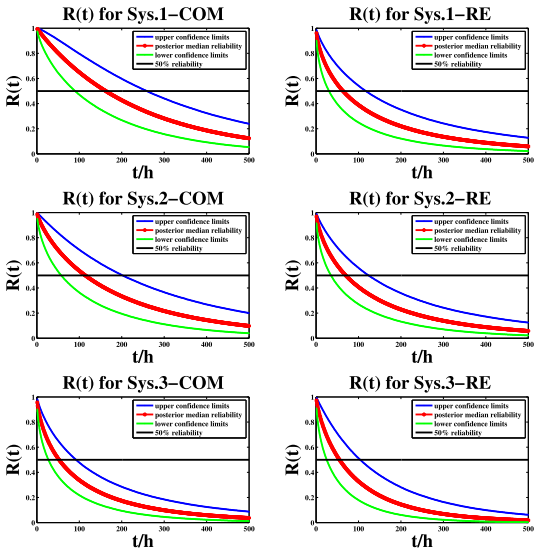

FIGURE 18. Reliability function.

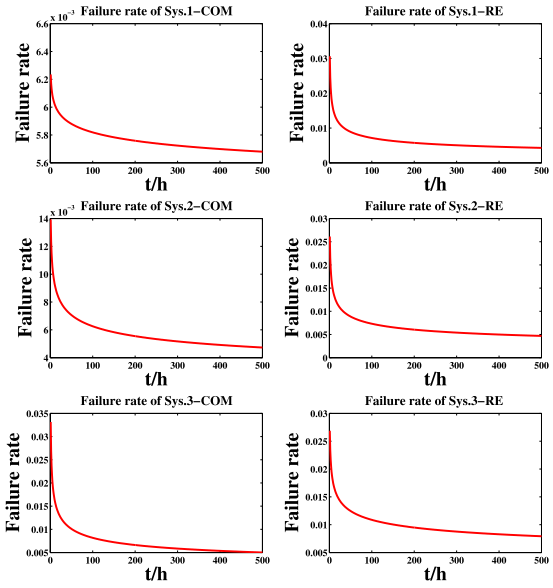

FIGURE 19. Failure rate of 6 units.

TABLE 5. MTBF of units.

Units	$\hat{\lambda}$	$\hat{\beta}$	MTBF/h
Sys.1-COM	0.00633	0.98500	171.7536
Sys.1-RE	0.04480	0.68393	121.4047
Sys.2-COM	0.01687	0.82613	155.0968
Sys.2-RE	0.03610	0.72372	120.8391
Sys.3-COM	0.04762	0.69599	101.0466
Sys.3-RE	0.03345	0.80349	77.5136

This can also be proved by the MTBF of these units in Table 5, which is calculated by equation (25).

V. CONCLUSION

This paper builds a specific model for operational data as well as for the failure data of train control on-board subsystems based on the data cube, and failure distribution characteristics are analyzed based on the slice and dice operations of the data model. After the failure distribution characteristics are analyzed, the representative equipment and systems are selected to analyze the reliability evaluation, which is estimated by Bayesian estimation. Further studies are expected to analyze the influence of different systems and their failure situations on the system's reliability.

REFERENCES

[1] Y. Huang and X. Zhou, "Knowledge model for electric power big data based on ontology and semantic Web," *CSEE J. Power Energy Syst.*, vol. 1, no. 1, pp. 19–27, May 2015.
[2] Y. Yan, G. Sheng, R. C. Qiu, and X. Jiang, "Big data modeling and analysis for power transmission equipment: A novel random matrix theoretical approach," *IEEE Access*, vol. 6, pp. 7148–7156, 2017.
[3] Y. Wang and L. Bai, "Fuzzy spatiotemporal data modeling based on UML," *IEEE Access*, vol. 7, pp. 45405–45416, 2019.
[4] T. Matsukawa and H. Funakoshi, "Analyzing failure frequency and severity in communication networks," in *Proc. Annu. Rel. Maintainability Symp. (RAMS)*, Jan. 2010, pp. 1–6.
[5] H. Zheng, Z. Jingkai, Z. Jian, C. Jia, H. Hua, J. Heng, and Z. Dan-Dan, "Characteristics of commutation failure based on fault recording," *J. Eng.*, vol. 2019, no. 16, pp. 1346–1349, Mar. 2019.
[6] G. Jäger, S. Zug, and A. Casimiro, "Generic sensor failure modeling for cooperative systems," *Sensors*, vol. 18, no. 3, p. 925, Mar. 2018.
[7] D. Feng, Z. He, S. Lin, Z. Wang, and X. Sun, "Risk index system for catenary lines of high-speed railway considering the characteristics of time-space differences," *IEEE Trans. Transp. Electrific.*, vol. 3, no. 3, pp. 739–749, Sep. 2017.
[8] R. Arno, N. Dowling, and R. J. Schuerger, "Equipment failure characteristics and RCM for optimizing maintenance cost," *IEEE Trans. Ind. Appl.*, vol. 52, no. 2, pp. 1257–1264, Mar./Apr. 2016.
[9] Z. Ding, Y. Zhou, G. Pu, and M. Zhou, "Online failure prediction for railway transportation systems based on fuzzy rules and data analysis," *IEEE Trans. Rel.*, vol. 67, no. 3, pp. 1143–1158, Sep. 2018.
[10] Z. Lin-Hai, W. Jian-Ping, and R. Yi-Kui, "Fault diagnosis for track circuit using AOK-TFRs and AGA," *Control Eng. Pract.*, vol. 20, no. 12, pp. 1270–1280, 2012.
[11] L.-H. Zhao, C.-L. Zhang, K.-M. Qiu, and Q.-L. Li, "A fault diagnosis method for the tuning area of jointless track circuits based on a neural network," *Proc. Inst. Mech. Eng., F, J. Rail Rapid Transit*, vol. 227, no. 4, pp. 333–343, Jul. 2013.
[12] T. de Bruin, K. Verbert, and R. Babuska, "Railway track circuit fault diagnosis using recurrent neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 523–533, Mar. 2017.
[13] F. Wang, T. Xu, T. Tang, M. Zhou, and H. Wang, "Bilevel feature extraction-based text mining for fault diagnosis of railway systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 49–58, Jan. 2017.
[14] J. Yin and W. Zhao, "Fault diagnosis network design for vehicle on-board equipments of high-speed railway: A deep learning approach," *Eng. Appl. Artif. Intell.*, vol. 56, pp. 250–259, Nov. 2016.
[15] Y. Zhao, T. Xu, and W. Hai-Feng, "Text mining based fault diagnosis of vehicle on-board equipment for high speed railway," in *Proc. 17th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 900–905.
[16] D. E. Brown, "Text mining the contributors to rail accidents," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 346–355, Feb. 2016.
[17] D. Zhang, "High-speed train control system big data analysis based on fuzzy rdf model and uncertain reasoning," *Int. J. Comput., Commun. Control*, vol. 12, no. 4, pp. 577–591, Aug. 2017.
[18] Y. Sun, Y. Cao, Y. Zhang, and C. Xu, "A novel life prediction method for railway safety relays using degradation parameters," *IEEE Intell. Transp. Syst. Mag.*, vol. 10, no. 3, pp. 48–56, Mar. 2018.
[19] Z. Xu, W. Wang, and Y. Sun, "Performance degradation monitoring for onboard speed sensors of trains," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1287–1297, Sep. 2012.
[20] L. Zhu, D. Yao, and H. Zhao, "Reliability analysis of next-generation CBTC data communication systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2024–2034, Mar. 2019.
[21] H. Su and Y. Che, "Reliability assessment of CTCS-3 using Bayesian networks," in *Proc. Int. Conf. Quality, Rel., Risk, Maintenance, Saf. Eng. (QR2MSE)*, Jul. 2013, pp. 284–288.

[22] A. Morant, P.-O. Larsson-Kråik, and U. Kumar, "Data-driven model for maintenance decision support: A case study of railway signalling systems," *Proc. Inst. Mech. Eng., F, J. Rail Rapid Transit*, vol. 230, no. 1, pp. 220–234, Jan. 2016.
[23] E. Pascale, L. Bouillaut, T. Freneaux, R. Sista, P. Sannino, and P. Marmo, "A weibull approach for enabling safety-oriented decision-making for electronic railway signaling systems," *Safety*, vol. 4, no. 2, p. 17, Apr. 2018.
[24] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh, "Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals," *Data Mining Knowl. Discovery*, vol. 1, no. 1, pp. 29–53, Mar. 1997.

BIN CHEN received the B.S. degree from Beijing Jiaotong University, Beijing, China, in 2012, where he is currently pursuing the Ph.D. degree with the School of Electronics and Information Engineering. His current research interests include fault diagnosis and reliability evaluation for train control systems.

BAIGEN CAI received the B.S., M.S., and Ph.D. degrees from Beijing Jiaotong University, in 1987, 1990, and 2010, respectively, all in traffic information engineering and control. Since 1990, he has been the Faculty Member with the School of Electronics and Information Engineering, Beijing Jiaotong University. He was a Visiting Scholar with The Ohio State University, from 1998 to 1999. He is currently a Professor with Beijing Jiaotong University. His current research interests include train control systems, intelligent transportation systems, GNSS navigation, multi-sensor fusion, and intelligent traffic control.

WEI SHANGGUAN received the B.S., M.S., and Ph.D. degrees from Harbin Engineering University, in 2002, 2005, and 2008, respectively. He was a Lecturer with the School of Electronics and Information Engineering, Beijing Jiaotong University, from 2008 to 2011. He was an Academic Visitor with the University College London, from 2013 to 2014. He is currently a Professor with Beijing Jiaotong University. His current research interests include train control systems (CTCS/ETCS/ERTMS), system modeling, simulation, and testing, GNSS (GPS, Galileo, Glonass, and BDS)/GIS, integrated navigation, intelligent transportation systems, and cooperative vehicle infrastructure system of China (CVIS-C).

JIAN WANG received the B.S., M.S., and Ph.D. degrees from Beijing Jiaotong University, Beijing, China, in 2000, 2003, and 2007, respectively. He was a Lecturer with the School of Electronics and Information Engineering, Beijing Jiaotong University, from 2007 to 2010, where he is currently a Professor. His current research interests include collaborative vehicle-road system research, computerized simulation of train control systems, new GNSS applications in railway, and other ITS technologies.