Journal of Computers Vol. 32 No. 2, 2021, pp. 33-46
doi:10.3966/199115992021043202004

Wireless Channel Modeling Based on Machine Learning in the High-speed Railway Scenario

Cui-Ran Li*, Yi-Hui Han, Bao-Feng Duan, Jian-Li Xie

School of Electronics and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China
{licr, xiejl}@mail.lzjtu.cn

Received 9 January 2020; Revised 8 May 2020; Accepted 24 June 2020

Abstract. In the high-speed railway (HSR) scenario, there is a problem of poor generalization ability caused by the overfitting phenomenon through wireless channel modeling. In this paper, a precise modeling method based on machine learning is proposed to address this issue under the wireless channel of HSR. According to the K-means clustering algorithm, the typical values of the *K*-factor in a multitude of scenarios are obtained through Rician *K*-factor clustering in the wireless channel fading model. Founded on the theory of model evaluation and selection in machine learning, the measured values of path loss are fitted by least squares regression using the cross-validation method. Then, we obtain the formulas of the path loss models and the expected generalization errors of the models in different scenarios. The mathematical relationship of the ergodic capacity depending on the Rician *K*-factor, the signal-noise ratio of the receiver, and the path loss generalization error are established after the ergodic capacity analysis of the HSR wireless channel. Finally, the lower boundary curves of the ergodic capacity are obtained by simulation experiments in different scenarios.

Keywords: high-speed railway, overfitting, Rician *K*-factor, cross validation, expected generalization errors

1 Introduction

Over the last few years, we have witnessed a sharp increase in the number of high-speed railway (HSR) operating mileages around the world, especially in China. It is estimated that the total HSR mileage will reach 50,000 kilometers by 2020. HSR construction has made remarkable achievements, providing strong support for the development of the economy and the progress of civilization [1]. To meet the demand for high-quality wireless communications in the HSR scenario and provide a reference for the design of the communication system, it is necessary to accurately understand the characteristics of the wireless channel. Owing to the high-speed mobility and multiple propagation scenarios in the HSR environment, it is rather difficult to carry out field measurements to acquire sufficient channel data. To realize reliable communication in an HSR environment, the propagation characteristics of radio frequencies need to be analyzed, which can be achieved by wireless channel modeling [2].

The channel model is an abstract description of the wireless environment and its propagation characteristics. Wireless channel measurement and modeling are important for the optimization, testing, and performance evaluation of communication systems as well as related technologies [3-4], such as network optimization scheduling and wireless resource management.

In an HSR environment, high-speed trains (HSTs) may pass through several scenarios, such as open spaces, hills, cuttings, viaducts and tunnels [5]. It is worth mentioning that viaducts account for over 60% of the HSR mileage [6]. The propagation path loss between transmitters and receivers has been widely predicted in urban or suburban environments [7-8]. This paper focuses on the research of wireless channels in plain, viaduct and cutting scenarios of HSR. Two transmission modes exist in wireless communication systems for HSR: direct transmission and relay-based transmission [9-10]. The use of

* Corresponding Author

mobile relay can improve the performance of a mobile communication system. Studies have demonstrated that the channel characteristic between the antenna of the evolved node base station (eNB) and the relay on the train is in the form of a Rician distribution [11].

Traditional channel modeling is difficult to learn from complex and large quantities of data. Machine learning brings a new idea of extracting data features and dealing with complex problems through models. With the rapid development of hardware technology and in-depth research on machine learning, machine learning algorithms have been widely used in designing and optimizing wireless communication systems.

In this paper, to improve the accuracy and reliability of the model, a new modeling method based on machine learning is proposed in HSR environments. In the wireless channel modeling, the varying Rician factor with a changing T-R distance is considered to overcome the overfitting phenomenon in the least squares fitting. By solving the problem of poor generalization performance in channel modeling, an accurate channel model under different scenarios is obtained. The proposed wireless channel modeling based on the machine learning paper is executed as follows: the K-means clustering algorithm is first used to cluster the Rician *K*-factor in plain, viaduct and cutting scenarios. Then, the path loss is modeled and analyzed, and the generalization error of the path loss is calculated in the plain, viaduct and cutting combined with the methods of cross validation and least squares. Third, the Nakagami-m distribution is used to approximate the Rician distribution to simplify the process of calculation and simulation. Fourth, by comparing the plain, viaduct and cutting scenarios, the influence of the Rician *K*-factor on the ergodic capacity is analyzed, and the lower limit formula of ergodic capacity is deduced.

2 Related Work

There are many research works on channel modeling in HSR scenarios. The cutting scenario is studied under the condition of HSR narrowband communication in [12]. The equation is established, indicating the correlation between the distance and the Rician factor. It is proven that the Rician function is the most suitable for describing the variation in the fading amplitude in the cutting scenario. In addition, the fading of the cutting scenario is more intense than that in other scenarios, and the proposed Rician factor model is more accurate than the classical model. In [13], the Rician *K*-factor of the viaduct scenario is modeled and analyzed under the conditions of HSR narrowband and broadband communication. Then, the model accuracy is verified. The cutting scenario is explored under the condition of HSR broadband communication, and then the expressions of the Rician factor and the path loss with distance are presented [14]. In [15-16], the relationship between the Rician factor and distance is analyzed under the condition of HSR broadband communication. At the same time, the distance is segmented, and the path loss is modeled. In [17], the wireless channel fading characteristics are measured and modeled by broadband communication in an HSR plain scenario. In the existing articles, the modeling error, which is caused by the difference of the Rician factor with the change in distance between the transmitter and the receiver, is hardly considered in the process of wireless channel modeling in the HSR scenario. In addition, there are few studies that theoretically analyze the overfitting phenomenon of path loss in the process of least square fitting. However, these two aspects may lead to poor generalization ability and low modeling accuracy of the wireless channel model.

In the technical specification (TS) 36.104 of the third generation partnership project (3GPP) [18], the fading-free channel model for baseband performance testing is defined in HSR. This model gives the Doppler characteristics of the receiver when the train passes through the base station. The model of WINNER-II [19] shows in detail the fading characteristics of large-scale and small-scale wireless channels measured by mobile relay technology in D2a. The wireless channel consists of two parts: between eNB and the relay and between the relay and the terminal. The HSR tunnel wideband measurements at 2.1 GHz are conducted in [20], and the channel parameters, including path loss, delay and Doppler characteristics, are obtained. In the case of narrowband communication with a bandwidth of 930 MHz, the wireless channel of HSR is measured by the authors of [5, 21], who proposed models of experience path loss for the viaduct height and cutting width and Rician *K*-factor for the first time. In [22], a location-based model of the wireless channel for HSR in the conditions of broadband communication with a bandwidth of 2.35 GHz is proposed. It is based on the channel data collected by the viaduct. Geometry and random propagation graph theory are two applicable methods for wireless channel modeling. In [23-24], a geometry-based stochastic model (GBSM) is established, and the space-time frequency correlation and nonstationary characteristics of channels in HSR are analyzed. In [25], a

simulation model of a time-varying channel based on graphics in HSR is proposed.

The above channel modeling has some limitations. It is difficult to learn quickly and accurately from a large quantity of complex data. Machine learning can track data fluctuation, which meets the requirements of channel modeling in HSR. In [26], the opportunities and challenges of cluster-based wireless channel modeling are discussed, and some common clustering algorithms for channel modeling are introduced. In [27], a clustering algorithm based on statistics is proposed, which uses the model of Gaussian mixture modeling of the multipath components of the channel and then optimizes its parameters using the expectation-maximization algorithm. However, the studies do not consider the modeling error caused by the varying Rician factor with different T-R distances under different scenarios, which is investigated in the paper. Additionally, we analyze the effect of overfitting on ergodic capacity performance theoretically.

3 HSR Scenarios and Relay Communication Mode

As shown in Fig. 1, the ground fluctuates very little in the plain scenario, similar to the surface of the viaduct. The flat surface makes the propagation environment of the electromagnetic wave more open. There are discontinuous and different types of trees around the viaduct. Cutting is a long, narrow and semiclosed structure with symmetrical escarpments on both sides. Its surface is usually covered with vegetation. When the height of the receiving antenna on the top of the train is lower than the height of the escarpments on both sides, the cutting is deep. In this case, the receiver receives more reflected and scattered waves.

(a) plain (b) viaduct (c) cutting

Fig. 1. Three scenarios for HSR operation

In an HSR wireless communication system, there are two communication modes between eNB and HST: direct mode with a single hop and relay mode with a double hop, as shown in Fig. 2. Direct mode indicates that a transmitting antenna installed on the eNB is directly connected to mobile terminals in the HST, thereby realizing signal transmission and reception. However, in this case, the severe penetration loss caused by HST is not conducive to the reception of wireless signals. The relay mode indicates that a transmitting antenna installed on the eNB first transmits the signal to a relay antenna deployed on the roof of the HST, and then the relay antenna retransmits the signal to mobile terminals in the HST. This mode can avoid penetration loss when the signal passes through the train body. Relay-based signal transmission is a better choice when strong penetration loss exists.

(a) direct mode (b) relay mode

Fig. 2. Two transmission modes of the HSR wireless communication system

The direct mode has better throughput performance in the high signal-noise ratio (SNR) region when penetration loss is given. However, in the low and medium SNR regions, the throughput performance based on the relay mode performs better [11]. Founded on the analysis, this paper studies the modeling of HSR wireless channels with relay modes under three types of scenarios: plain, viaduct and cutting.

4 Wireless Channel Modeling of HSR Based on Machine Learning

4.1 Value of Rician *K*-Factor

The Rician *K*-factor, which is used to measure the severity of wireless signal fading, plays a key role in wireless link budgeting, channel modeling, and system simulation [28-29]. The Rician *K*-factor is defined as [30]:

$$K(\mathrm{dB}) = 10\lg\frac{A^2}{2\sigma^2}(\mathrm{dB}), \tag{1}$$

where A refers to the peak value of the amplitude of the main signal and σ represents the variance in a multipath component. The probability density function is expressed as follows:

$$p(r) = \begin{cases} \dfrac{r}{\sigma^2} e^{-\frac{r^2+A^2}{2\sigma^2}} I_0\left(\dfrac{Ar}{\sigma^2}\right), A \geq 0, r \geq 0, \\ 0, \qquad\qquad r < 0 \end{cases} \tag{2}$$

where r is the envelope of the received signal and $I_0(\bullet)$, which indicates the zeroth-order modified Bessel function of the first class, is defined as:

$$I_0(x) = \sum_{m=0}^{\infty} \frac{1}{m!\Gamma(m+1)} \left(\frac{x}{2}\right)^{2m}. \tag{3}$$

The K-means clustering algorithm is a distance-based clustering algorithm. Distance is used as an evaluation index for similarity measurement, that is, the closer the distance between two objects is, the more similar they are. This kind of algorithm usually forms clusters of objects that are close to each other to obtain different clusters that are compact and independent. The algorithm for clustering the Rician *K*-factor using K-means clustering is as follows [31]:

In [3, 17], the Rician *K*-factor was measured with a frequency of 1.89 GHz, 2.35 GHz and 2.35 GHz, respectively, in the plain, viaduct and cutting scenarios. The antenna is configured as single-input single-output (SISO). The K-means clustering algorithm is used to cluster the Rician *K*-factor following the change in distance, as shown in Fig. 3, Fig. 4 and Fig. 5, and the typical values of the Rician *K*-factor within different distances are shown in Table 1.

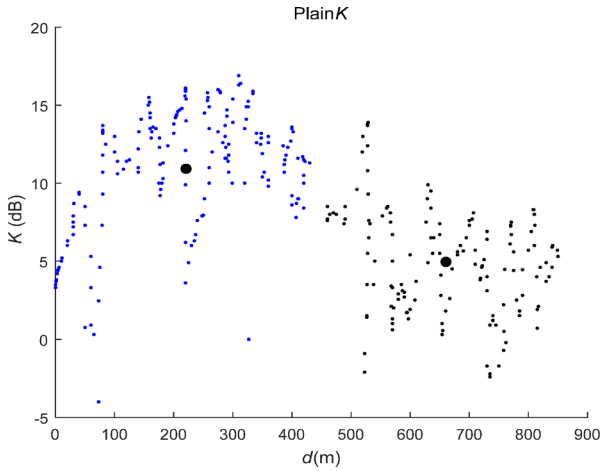

Fig. 3. Rician *K*-factor and clustering in the plain scenario

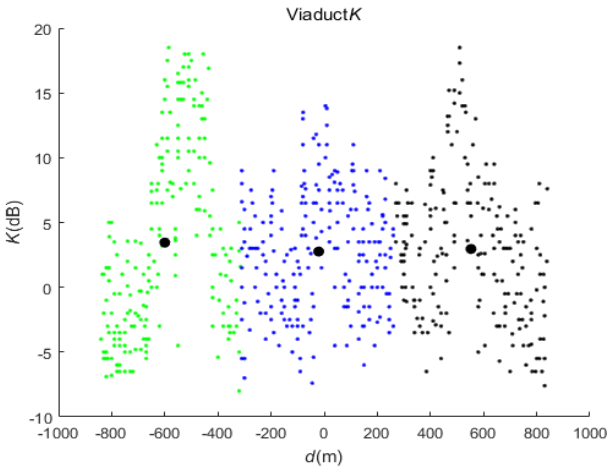

Fig. 4. Rician *K*-factor and clustering in the viaduct scenario

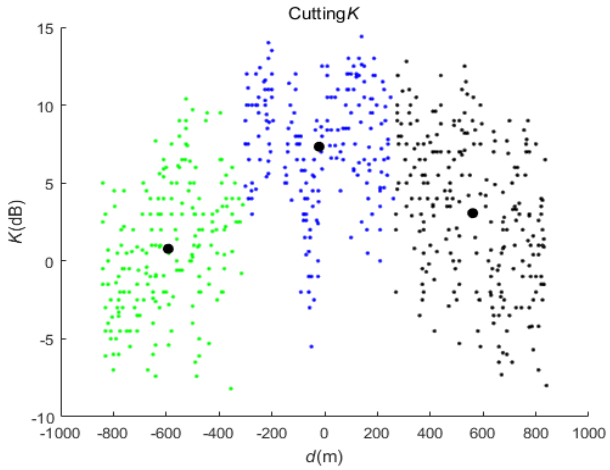

Fig. 5. Rician *K*-factor and clustering in the cutting scenario

Table 1. Algorithm process

Input: Sample set $D=\{\boldsymbol{h}_1,\boldsymbol{h}_2,...,\boldsymbol{h}_m\}$ composed of Rician *K*-factor and the number *k* of Rician factor cluster heads.
Process:
1. Randomly select *k* Rician factor samples from the Rician factor sample set *D* as the initial mean vectors $\{\boldsymbol{u}_1,\boldsymbol{u}_2,...,\boldsymbol{u}_k\}$;
2. repeat;
3. Let $U_i=\varnothing$ $(1\le i\le k)$;
4. for *j*=1, 2, ..., *m* do;
5. Calculate the distance $d_{ji}=\left\|\boldsymbol{h}_j-\boldsymbol{u}_i\right\|_2$ between the Rician factor sample $\boldsymbol{h}_j$ and each mean vector $\boldsymbol{u}_i(1\le i\le k)$;
6. Determine the cluster mark of the Rician factor based on the nearest mean vector: $\lambda_j=\text{argmin}_{i\in\{1,2,...,k\}}d_{ji}$;
7. Divide the Rician factor sample $\boldsymbol{h}_j$ into the corresponding cluster: $U_{\lambda_j}=U_{\lambda_j}\cup\{\boldsymbol{h}_j\}$;
8. end for;
9. for *i*=1,2,...,*k* do;
10. Calculate the new mean vector of the Rician factor: $\boldsymbol{u}_i'=\frac{1}{\left\|U_i\right\|}\sum_{\boldsymbol{h}\in U_i}\boldsymbol{h}$;

11. if $\boldsymbol{u}_i \neq \boldsymbol{u}_i^{'}$ then;
12. Update the current mean vector $\boldsymbol{u}_i$ to $\boldsymbol{u}_i^{'}$;
13. else;
14. Keep the current mean vector unchanged;
15. end if;
16. end for;
17. until the mean vector of the current Rician factor has not been updated.

Output: Cluster division $U = \{U_1, U_2, ..., U_k\}$.

In Fig. 4 and Fig. 5, d represents the distance between the transmitting antenna and the receiving antenna. When a train moves towards the source of the electromagnetic signal, d takes a negative value. When a train runs away from the source of the electromagnetic signal, the value of d is positive. In Table 2, the typical Rician factor values corresponding to the mean vectors of cluster U_1, cluster U_2, and cluster U_3 are represented by K_1, K_2, and K_3, respectively.

Table 2. Typical value of the Rician K-factor

Scenarios	Rician K-factor
Plain	K_1=10.91($0 \le d \le 430$)
	K_2=4.96($430 < d \le 850$)
Viaduct	K_1=3.45($-850 \le d \le -320$)
	K_2=2.76($-320 < d \le 260$)
	K_3=2.95($260 < d \le 850$)
Cutting	K_1=1.27($-850 \le d \le -275$)
	K_2=7.30($-275 < d \le 295$)
	K_3=2.83($295 < d \le 850$)

4.2 Cross-validation Theory and Data Fitting

In the procedure of fitting the path loss data, an overfitting phenomenon easily occurs. Overfitting indicates that the learner regards some characteristics of the training sample as the general property suitable to all samples, which may worsen the generalization ability of the model. To overcome this drawback, the cross-validation method is used to determine whether the path loss model is overfitting and to obtain generalization error. The principle of the cross-validation method is first to divide the dataset $P = \{p_1, p_2, ..., p_n\}$ that is composed of path loss into a reciprocal subset of similar size and to keep the data distribution consistent in each subset. Next, the union of a-1 subsets is used as the training set, and the remaining subsets are used as the test set for training and testing [31].

The fitting values of cross validation, which are similar to the normal fitting values that represent the fitness of the model to the data, can indicate the accuracy of the model prediction data. To measure the fitted value of the cross validation of the observation, this observation must be ignored from the data used to calculate the model, and then the fitting value is calculated by using the unrelated coefficient vector. The cross validation fitting of the ignored path loss observations is defined as follows:

$$\hat{p}_{(i),n} = b_{o(i)} + \sum_j X_{i,j} \times B_{(i)(j,n)}, \quad \textbf{(4)}$$

where (i) represents that the value of observation p_i is ignored when calculating the model. $b_{o(i)}$ represents the model intercept when the ith observed response value is not included in the path loss dataset. X represents the dataset of distance d. $B_{(i)(j,n)}$ represents a model coefficient when the ith observed response value is not included in the path loss dataset. The residual is the difference between the actual response variables and the cross-validation fitted values, which is defined as follows:

$$e(i) = p_i - \hat{p}_{(i),n}, \tag{5}$$

where p_i represents the ith observed response value in the path loss dataset. The expected generalization error for cross validation is defined as:

$$\mu = \frac{\sum_{i=1}^{n} (p_i - \hat{p}_{(i),n})^2}{n}. \tag{6}$$

Deviation-variance decomposition can be used to explain the data fitting generalization ability based on cross-validation theory [31]. This method can decompose the expected generalization error rate. For the test sample x, p is the mark of x in the dataset P, and y is the true mark of x. The expected prediction for the least squares fitting algorithm is expressed as follows:

$$\bar{f}(x) = E_P[\hat{p}_{(i),n}]. \tag{7}$$

The variance generated by using different training sets with the same number of samples is expressed as follows:

$$var(x) = E_P[(\hat{p}_{(i),n} - \bar{f}(x))^2]. \tag{8}$$

The noise is expressed as follows:

$$\varepsilon^2 = E_P[(p - y)^2]. \tag{9}$$

The difference between the expected output and the true mark is called the bias, which is defined as:

$$bias^2(x) = (\bar{f}(x) - y)^2. \tag{10}$$

In summary, using the polynomial expansion and combination to decompose the expected generalization error of the algorithm can obtain:

$$\begin{aligned}
\mu &= E_P[(\hat{p}_{(i),n} - p)^2] \\
&= E_P[(\hat{p}_{(i),n} - \bar{f}(x) + \bar{f}(x) - p)^2] \\
&= E_P[(\hat{p}_{(i),n} - \bar{f}(x))^2] + E_P[(\bar{f}(x) - p)^2] + E_P[2(\hat{p}_{(i),n} - \bar{f}(x))(\bar{f}(x) - p)] \\
&= E_P[(\hat{p}_{(i),n} - \bar{f}(x))^2] + E_P[(\bar{f}(x) - p)^2] \\
&= E_P[(\hat{p}_{(i),n} - \bar{f}(x))^2] + E_P[(\bar{f}(x) - y + y - p)^2] \\
&= E_P[(\hat{p}_{(i),n} - \bar{f}(x))^2] + E_P[(\bar{f}(x) - y)^2] + E_P[(y - p)^2] + 2E_P[(\bar{f}(x) - y)(y - p)] \\
&= E_P[(\hat{p}_{(i),n} - \bar{f}(x))^2] + (\bar{f}(x) - y)^2 + E_P[(p - y)^2]
\end{aligned} \tag{11}$$

For the sake of discussion, it is assumed that the noise expectation is 0, that is, $E_P[p - y] = 0$. Then, the generalization error can be expressed as follows:

$$\mu = bias^2(x) + \mathrm{var}(x) + \varepsilon^2. \tag{12}$$

The above analysis shows that the generalization error can be decomposed into the sum of deviation, variance and noise. Deviation-variance decomposition shows that generalization ability is determined by the ability of the data fitting algorithm, the adequacy of the data, and the difficulty of the fitting task. In addition, the generalization error is more representative of generalization ability than simple deviation or variance.

4.3 Regression-fitting Path Loss Model Based on Cross-validation

The steps of the cross validation based on the regression-fitting algorithm are as follows:
Step 1. According to the cross-validation method, a set of measured path loss observation data, that is, a test set, is ignored;
Step 2. The path loss model is fitted without considering the test set;
Step 3. The cross-validation fitted values of the ignored path loss data are predicted based on the fitting path loss model (ignoring a set of test sets), and the residual values are calculated.
Step 4. Repeat the above steps until all path loss observations are ignored and fitted. Finally, the generalization error value and value of R_{sq_pred} are obtained.

The path loss model formula (16~18) and its generalization error in three scenarios are obtained by cross-validation regression fitting on the path loss measurement values in [3, 17], as shown in Table 3.

Table 3. Generalization error and reliability verification results in different scenarios

Scenarios / Parameters	Distance(m)	μ	R_{sq}	R_{sq_pred}
Plain	$0 \le d \le 430$ $430 < d \le 850$	2.71 4.53	87.88%	87.68%
Viaduct	$90 \le d \le 260$ $260 < d \le 850$	16.09 14.75	86.88%	86.80%
Cutting	$20 \le d \le 295$ $295 < d \le 850$	3.79 13.90	92.49%	92.45%

In Table 3, R_{sq} represents the percentage of variation in the response explained by the model, which determines the fitness between the model and the data. It is defined as follows [32]:

$$R_{sq} = 1 - \frac{\sum_{i=1}^{n} (p_i - \hat{p}_{i,n})^2}{\sum_{i=1}^{n} (p_i - \bar{p})^2}, \tag{13}$$

where $\bar{p}$ represents the average path loss. $\hat{p}_{i,n}$ represents the ith fitting response, which is defined as follows:

$$\hat{p}_{i,n} = b_{oi} + \sum_j X_{i,j} \times B_{i(j,n)}, \tag{14}$$

where b_{oi} represents the model intercept and $B_{i(j,n)}$ represents a model coefficient. R_{sq_pred} represents the prediction goodness of the path loss model. The calculation process of R_{sq_pred} is to delete each observation from the dataset, estimate the regression equation, and then obtain the prediction goodness of the model for the deleted observations, which is defined as:

$$R_{sq\_pred} = 1 - \frac{\sum_{i=1}^{n} (p_i - \hat{p}_{(i),n})^2}{\sum_{i=1}^{n} (p_i - \bar{p})^2}. \tag{15}$$

It can be seen from Table 3 that $\left| R_{sq\_pred} - R_{sq} \right| < 1\%$, which means that there is no overfitting in the models of the three scenarios. Equations (16)-(18) show the path loss models of the plain, viaduct and cutting scenarios, respectively.

In the plain scenario, the relationship between path loss and distance is expressed as follows:

$$PL_P(d) = 14.53 + 33.44 \log d \cdot \tag{16}$$

In the viaduct scenario, the relationship between path loss and distance is expressed as follows:

$$PL_V(d) = 14.30 + 40.98 \log d \cdot \tag{17}$$

In the cutting scenario, the relationship between path loss and distance is expressed as follows:

$$PL_C(d) = 19.95 + 34.90 \log d \cdot \tag{18}$$

5 Ergodic Capacity Analysis in Different Scenarios

5.1 Ergodic Capacity and Lower Bound Value

Assuming that channel state information (CSI) between the eNB and the relay node (train) is known in the relay node, the expression of the ergodic capacity can be written as follows [9]:

$$\begin{aligned} C &= E[\log(1+\gamma)] \\ &= \frac{1}{\ln 2} \int_0^\infty \frac{1}{\bar{\gamma}} \exp(-\frac{\gamma}{\bar{\gamma}}) \ln(1+\gamma) d\gamma \end{aligned}, \tag{19}$$

where γ represents the SNR, and $\bar{\gamma}$ represents the average SNR. The probability density function of the SNR is expressed as follows:

$$p(\gamma) = \frac{(1+K)e^{-K}}{\bar{\gamma}} \exp(-\frac{(1+K)\gamma}{\bar{\gamma}}) \times I_0(2\sqrt{\frac{K(1+K^2)\gamma}{\bar{\gamma}}}) . \tag{20}$$

Then, the iterative analysis expression for the ergodic capacity is expressed as follows:

$$C = \frac{e^{-K}(1+K)}{\bar{\gamma}} \sum_{n=0}^{\infty} \frac{1}{(\Gamma(n+1))^2} \left[\frac{K(1+K)}{\bar{\gamma}}\right]^n \times G_{2,3}^{3,1}\left[\frac{K+1}{\bar{\gamma}} \middle| \begin{matrix} -1-n,-n \\ 0,-1-n,-1-n \end{matrix}\right] . \tag{21}$$

Because equation (21) contains an infinite term, it is necessary to cut off its upper bound during simulation, which is inconvenient in the simulation calculation. Therefore, Nakagami-m fading is used to approximate the Rician fading [11]. In this case, the probability density function of SNR per symbol is expressed as follows [33]:

$$p'(\gamma) = \frac{m^m \gamma^{m-1}}{\bar{\gamma}^m \Gamma(m)} . \tag{22}$$

The relationship between the shape factor m and the Rician K-factor can be expressed as follows:

$$m = \frac{(1+K)^2}{1+2K} . \tag{23}$$

The value of the shape factor m in the three scenarios can be obtained from Table 3 and equation (23) that and is shown in Table 4.

Table 4. The value of the shape factor m in three scenarios

Scenarios	Shape factor m
Plain	m_1=6.22(0<d≤430) m_2=3.25(430<d≤850)
Viaduct	m_1=2.51(-850<d≤-320) m_2=2.17(-320<d≤260) m_3=2.26(260<d≤850)
Cutting	m_1=0.93(-850<d≤-275) m_2=4.42(-275<d≤295) m_3=2.30(295<d≤850)

Therefore, we can obtain an approximate expression of the ergodic capacity as follows:

$$C^* = \frac{1}{\Gamma((1+K)^2/(1+2K))} G_{2,3}^{3,1}\left[\frac{(1+K)^2}{1+2K}\middle|\frac{0,1}{\bar{\gamma}}\middle|\frac{(1+K)^2}{1+2K},0,0\right] = \frac{1}{\Gamma(m)} G_{2,3}^{3,1}\left[\frac{m}{\bar{\gamma}}\middle|\frac{0,1}{m,0,0}\right], \tag{24}$$

where $G(\bullet)$ represents the Meijer G function and is defined as follows:

$$G_{p,q}^{m,n}\left(\begin{matrix}a_1, & \ldots & ,a_p\\ b_1, & \ldots & ,b_q\end{matrix}\middle| z\right) = \frac{1}{2\pi i}\int_L \frac{\Pi_{j=1}^{m}\Gamma(b_j - s)\Pi_{j=1}^{n}\Gamma(1-a_j+s)}{\Pi_{j=m+1}^{q}\Gamma(1-b_j+s)\Pi_{j=m+1}^{p}\Gamma(a_j-s)} z^s ds, \tag{25}$$

$\Gamma(\bullet)$ represents the gamma function, which is defined as follows:

$$\Gamma(x) = 2\int_0^{+\infty} t^{2x-1}e^{-t^2}dt \cdot \tag{26}$$

Assuming that the generalization error of the path loss is μ dB, the lower bound formula of the ergodic capacity can be expressed as follows:

$$C_L^* = \frac{1}{\Gamma(m)} G_{2,3}^{3,1}\left[\frac{m}{\bar{\gamma}\text{-}\mu}\middle|\frac{0,1}{m,0,0}\right]. \tag{27}$$

5.2 Simulation Results

The value of the Rician K-factor varies by the HSR scenario, so the channel ergodic capacity performance in different scenarios is also different. Fig. 6 shows the ergodic capacity to average the SNR curve for different scenarios. In the figure, the Rician K-factor in the plain, viaduct and cutting scenarios takes the mean values of 7.94 dB, 3.05 dB and 3.80 dB, respectively.

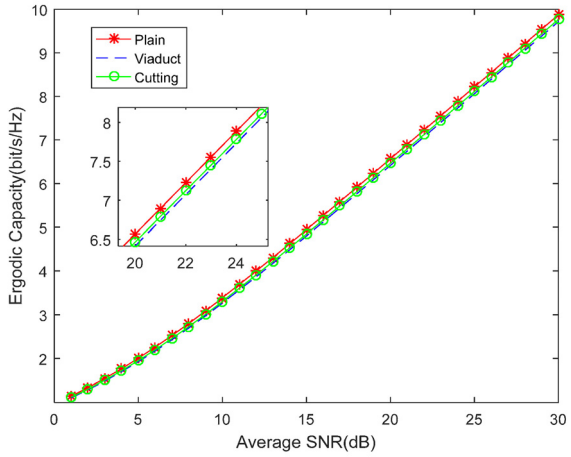

Fig. 6. Ergodic capacity curve in different scenarios

As seen in Fig. 6, the ergodic capacity under the same SNR ratio condition increases with the Rician K-factor. That is, under the same SNR conditions, the ergodic capacity of the HSR wireless channel in the plain scenario is the largest, the ergodic capacity in the cutting scenario is second, and the ergodic capacity in the viaduct scenario is the smallest.

Fig. 7, Fig. 8 and Fig. 9 show the ergodic capacity lower bound curves for the plain, viaduct and cutting scenarios, respectively. In these figures, $C^*(K_i)$ and $C_L^*(K_i)$ represent the HSR wireless channel ergodic capacity and its lower bound, respectively. Because of the path loss generalization error, the ergodic capacity of the HSR wireless channel is drastically reduced.

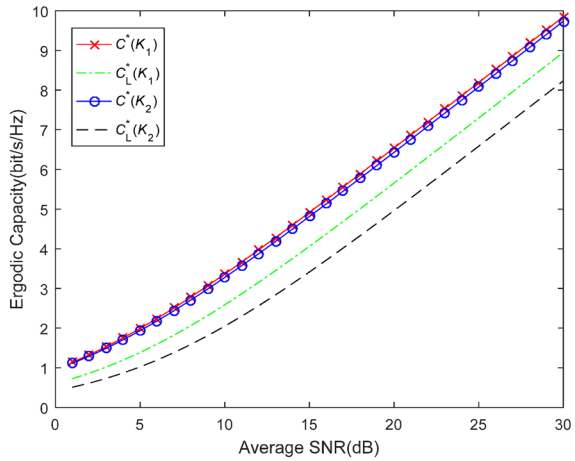

Fig. 7. Ergodic capacity lower bound value in the plain scenario

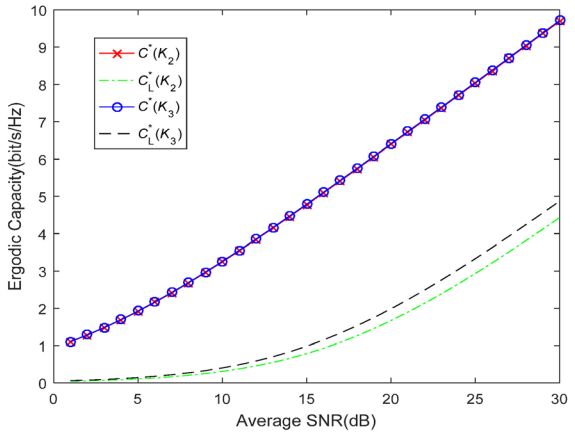

Fig. 8. Ergodic capacity lower bound value in the viaduct scenario

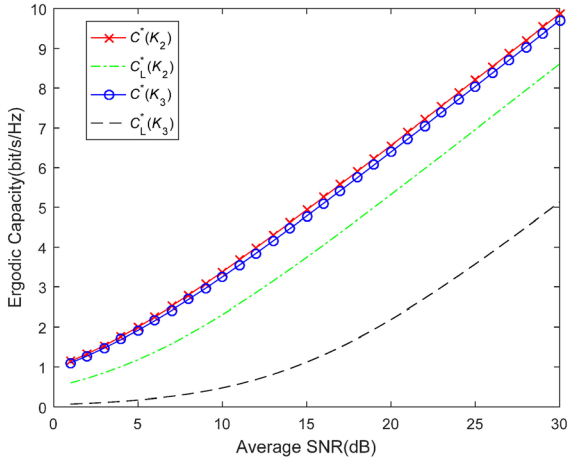

Fig. 9. Ergodic capacity lower bound value in the cutting scenario

In some regions, the generalization error reduces the traversal capacity by more than 50%. With the increase in the average SNR, the influence of generalization error on the ergodic capacity is increasingly obvious. By comparing Fig. 6-9, it can be seen that the influence of generalization error on the ergodic capacity is far greater than that of the difference of scenario, that is, the difference in the Rician factor.

6 Conclusion

This paper studies the wireless channel modeling of HSR based on the relay communication mode in three scenarios: plain, viaduct and cutting. The K-means clustering algorithm is used to cluster the Rician *K*-factor and to obtain the typical value of the Rician *K*-factor in the corresponding distance range. The modeling error caused by ignoring the variation in the Rician *K*-factor value in the modeling process is avoided. The influence of the Rician *K*-factor on ergodic capacity performance is analyzed with a given SNR. It is proven that when the average SNR is the same, the wireless channel ergodic capacity increases with the Rician *K*-factor. Through regression fitting based on cross-validation theory, we can avoid overfitting and obtain path loss models and generalization errors in different scenarios. Lower bounds of ergodic capacity in three scenarios are obtained, which prove that the existence of the generalization error will lead to a sharp decrease in the ergodic capacity.

Our research focused on the SISO wireless system in three typical HSR scenarios. In our future work, channel modeling for the multiantenna system will be carried out. In addition, it is worth exploring the generalization error-related algorithms in machine learning to improve the accuracy in wireless channeling and the ergodic capacity in HSR scenarios.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (61661025, 61661026), and Science and Technology Plan of Gansu Province (Grant No. 20JR10RA 273).

References

[1] Z.-D. Zhong, B. Ai, G. Zhu, H. Wu, L. Xiong, F.-G. Wang, L. Lei, J.-W. Ding, K. Guan, R.-S. He, Dedicated Mobile Communications for High-speed Railway, Springer, Berlin, 2017 (Chapter 1).

[2] Y.-F. Chen, Z.-R. Zhang, T.-F. Qin, Geometrically Based Channel Model for Indoor Radio Propagation, Progress In Electromagnetics Research B 20(2010) 109-124.

[3] T. Zhou, Research on the Propagation Characteristics, Modeling and Measurement of the Wireless Channel of High-speed Railway, Beijing Jiaotong University, 2016.

[4] T. Zhou, C. Tao, L. Liu, Z.-H. Tan, A semiempirical MIMO channel model in obstructed viaduct scenarios on high-speed railway, in: Proc. 2014 IEEE International Conference on Communications, 2014.

[5] R.-S. He, Z.-D. Zhong, B. Ai, J.-W. Ding, Propagation measurements and analysis for high-speed railway cutting scenario, Electronics Letters 47(21)(2011) 1167-1168.

[6] W. Qian, C.-M. Xu, M.-Q. Wu, Zhao M., D.-S. Yu, Propagation characteristics of high speed railway radio channel based on broadband measurements at 2.6 GHz, in: Proc. 2014 IEEE Wireless Communications and Networking Conference, 2014.

[7] B. Ai, R.-S. He, Z.-D. Zhong, K. Guan, B.-C. Chen, P.-Y. Liu, Y. Li, Radio Wave Propagation Scene Partitioning for High-Speed Rails, International Journal of Antennas and Propagation 2012(21)(2012) 1072-1075.

[8] B. Ai, Interview, Electronics Letters 47(21)(2011) 1158.

[9] Y. Zhang, Z. He, W. Zhang, L. Xiao, S. Zhou, Measurement-based delay and doppler characterizations for high-speed railway hilly scenario, International Journal of Antennas and Propagation 2014(2014) 1-8.

[10] J. C. Sanchez, M. M. Garcia, D. D. A. Monte, Performance evaluation of in-band LTE mobile relays in high speed railway environments, in: Proc. 2014 44th European Microwave Conference, 2014.

[11] J. You, Z. Zhong, R. Xu, G. Wang, Transmission schemes for high-speed railway: Direct or relay?, in: Proc. 2012 8th International Wireless Communications and Mobile Computing Conference, 2012.

[12] R.-S. He, Z.-D. Zhong, B. Ai, J.-W. Ding, Y.-Q. Yang, A. F. Molisch, Short-Term Fading Behavior in High-Speed Railway Cutting Scenario: Measurements, Analysis, and Statistical Models, IEEE Transactions on Antennas and Propagation 61(4)(2013) 2209-2222.

[13] T. Zhou, C. Tao, L. Liu, Z.-H. Tan, R.-C. Sun, Research on the rice K factor in measurement-based high-speed railway broadband wireless channel, Tiedao Xuebao 35(9)(2013) 72-78.

[14] R.-C. Sun, C. Tao, L. Liu, Z.-H. Tan, Channel Measurement and Characterization for HSR U-Shape Groove Scenarios at 2.35 GHz, in: Proc. 2013 IEEE 78th Vehicular Technology Conference, 2013.

[15] Y.-L. Guo, J. Zhang, C. Zhang, L. Tian, Correlation analysis of high-speed railway channel parameters based on channel measurement, in: Proc. International Workshop on High Mobility Wireless Communications, 2013.

[16] H. Zhao, Q. Lyu, Y.-C. Liu, J.-X. Chen, S.-J. Zhang, Wireless Channel Measurements and Modeling of LTE Broadband System for High-Speed Railway Scenarios, Chinese Journal of Electronics 27(05)(2018) 208-213.

[17] H. Wen, Research on fading and nonstationary characteristics of Wireless Channel in High Speed Railway scene. Beijing Jiaotong University, 2018.

[18] 3GPP TS 36.104 V9.3.0, Technical specification group radio access network; evolved universal terrestrial radio access; base station radio transmission and reception, 2010.

[19] P. Kyosti, WINNER II channel models part II radio channel measurement and analysis results, 2007.

[20] P. Aikio, R. Gruber, P. Vainikainen, Wideband radio channel measurements for train tunnels, in: Proc. 1998 48th IEEE Vehicular Technology Conference, 1998.

[21] R.-S. He, Z.-D. Zhong, B. Ai, J. Ding, An empirical path loss model and fading analysis for high-speed railway viaduct scenarios, IEEE Antennas and Wireless Propagation Letters 10(2011) 808-812.

[22] L. Liu, C. Tao, J.-H. Qiu, H.-J. Chen, Position-based modeling for wireless channel on high-speed railway under a viaduct at 2.35 GHz, IEEE Journal on Selected Areas in Communications 30(4)(2012) 834-845.

[23] A. Ghazal, C.-X. Wang, H. Haas, M. Beach, A non-stationary MIMO channel model for high-speed train communication systems, in: Proc. 2012 75th IEEE Vehicular Technology Conference, 2012.

[24] B.-H. Chen, Z.-D. Zhong, Geometry-based stochastic modeling for MIMO channel in high-speed mobile scenario, International Journal of Antennas and Propagation 2012(2012).

[25] L. Tian, X.-F. Yin, Q. Zuo, J.-H. Zhou, Z.-M. Zhong, S.-X. Lu, Channel modeling based on random propagation graphs for high speed railway scenarios, in: Proc. 23rd IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2012.

[26] R.-S. He, B. Ai, A. F. Molisch, G. L. Stuber, Clustering enabled wireless channel modeling using big data algorithms, IEEE Communications Magazine 56(5)(2018) 177-183.

[27] Y.-P. Li, J.-H. Zhang J, Z.-Y. Ma, Clustering in wireless propagation channel with a statistics-based framework, in: Proc. IEEE Wireless Communications and Networking Conference, 2018.

[28] X. Cheng, C.-X. Wang, H.-M. Wang, X.-Q. Gao, X.-H. You, D.-F. Yuan, B. Ai, Q. Huo, L. Song, B.-L. Jiao, Cooperative MIMO Channel Modeling and Multi-link Spatial Correlation Properties, IEEE Journal on Selected Areas in Communications 30(2)(2012) 388-396.

[29] X. Cheng, C.-X. Wang, B. Ai, H. Aggoune, Envelope Level Crossing Rate and Average Fade Duration of Nonisotropic Vehicle-to-Vehicle Ricean Fading Channels, IEEE transactions on intelligent transportation systems 15(1)(2014) 62-72.

[30] T.S. Rappaport, Wireless Communications: Principles and Practice 2nd Edition, Prentice Hall Press, New Jersey, 2001.

[31] Z.-H. Zhou, Machine Learning, Tsinghua University Press, Beijing, 2016 (Chapter 2).

[32] H.P. Piepho, A Coefficient of Determination (R2) for Linear Mixed Models, Biometrical Journal 61(4)(2019) 860-872.

[33] X. Yu, G. Bi, Performance of multiband complex wavelet based multicarrier Ds-CDMA system with multi-antenna receiver over Nakagami-M fading channel, Progress In Electromagnetics Research 98(2009) 251-266.