ISSN: 2688-8653

Evaluation and Extension of Text Reading Difficulty

Yinghui Liu, Changyun Lu, Hongxia Ye, Qingbo Ru

Anhui University of Finance and Economics undergraduate, Anhui, China

Abstract

In recent years, with the expansion of the influence of English, people's demand for reading English texts has increased day by day. However, with the rapid development of the Internet, information resources are becoming more and more complicated. Faced with such a situation, many people cannot find English texts that are suitable for them to read. If the levels of different English texts can be determined, finding English texts that match the individual's level will greatly improve efficiency and reading ability. Therefore, the evaluation of the readability of English texts is one of the important and indispensable research directions. Readability, also known as legibility, refers to the degree or nature of the text that is easy to read and understand. It is mainly used to evaluate the ease of reading and understanding by readers. The main factors affecting readability are the average length of sentences, the number of unfamiliar words and the complexity of the grammar used. Therefore, we measure the reading difficulty of English texts. First, this paper selects 8 evaluation indicators such as Word length, Sentence length and Total text syllables to construct the evaluation index system of English text reading difficulty; then, this paper collects 319 typical English texts. Relevant index data is divided into four categories: A, B, C, and D according to the degree of difficulty, and the correlation analysis of the indicators is carried out. The entropy weight-TOPSIS model and the gray comprehensive evaluation model are used to evaluate the reading difficulty of the sample and calculate the average score, and classify the reading difficulty of 319 English texts according to the results; finally, use these 319 as the original database to measure the reading difficulty of any new English text. In addition, This article extends the difficulty of reading English text to the analysis and evaluation of reading difficulty of Chinese text. First, take the score of Chinese text reading difficulty as the explained variable, select the factors that affect the reading difficulty of Chinese text as the explanatory variable, and use BP neural network to predict the expected reading difficulty of the text; secondly, consider that the reading difficulty of Chinese text is generally more difficult, this article uses multi-objective programming to construct the objective function, and uses the simulated annealing algorithm to solve this function, thereby reducing the difficulty of reading Chinese text.

Keywords

Reading Difficulty; Entropy Weight-TOPSIS; Grey Comprehensive Evaluation; BP Neural Network; Simulated Annealing.

1. Introduction

Readability, also known as legibility or legibility, refers to the degree or nature of text that is easy to read and understand. It is mainly used to evaluate how easy the reading material can be read and understood by readers. Developing reading ability is an important component of language learning, and the importance of reading text to developing reading ability is self-evident. In recent years, in order to ensure that the reading text meets the reader's language proficiency, researchers have a certain degree of subjectivity in the data processing of the difficulty of English text. With the development of technologies such as natural language processing, data mining, and machine learning, we can determine the objective accuracy model through the construction of parameters to measure the reading level of an English article.

1.1. Problem Background

As we all know, English is the most widely used world language, and its importance is self-evident. With the development of China's comprehensive national strength and the advancement of its internationalization, more and more people in China are devoting themselves to the study of English. Among the five parts of English listening, speaking, reading, writing, and translation, reading is the most important aspect of improving English learning. Reading can not only enrich and increase vocabulary and grammar knowledge, and improve writing skills, but also broaden one's horizons and broaden access to information. The provision of appropriate reading materials is the prerequisite and key to effective reading activities. In view of the importance of the readability of English texts in meeting people's information needs, and the explosive growth of modern information, the need for measuring the readability of English texts continues to increase. Therefore, it is current to achieve effective readability evaluation of English texts which is needed by the trend.

The earliest research on the readability of English text can be traced back to the 1920s. The Flesch-Kincaid formula and the Gunning Fog formula are two classic readability formulas. The Flesh-Kincaid formula is a built-in readability formula of Microsoft office word.

In such an Internet age and an era when the demand for English reading is soaring, it is particularly important to provide readers with the most relevant and suitable English texts. Research on the readability of English texts came into being. This article analyzes the data by using indicators such as word length, sentence length, and syllables, and uses machine learning methods to promote it.

1.2. Restatement of the Problem

Considering the background information and restricted conditions identified in the problem statement, we need to solve the following problems:

- Problem 1

Comprehensively consider the full text word length, sentence length and the total number of text syllables and other factors to construct an evaluation index system for the difficulty of reading English text.

- Problem 2

By collecting relevant data, an evaluation model of the reading difficulty of English texts is constructed, and the reading difficulty of English texts is quantitatively analyzed, so that the text can be used as a reading material for language tests of moderate difficulty levels.

- Problem 3

Research how to extend this evaluation model of English text reading difficulty to other model languages such as Chinese.

1.3. Literature Review

The research on the difficulty and readability of English text has a long history. The first formula for English legibility was created by the Americans [1] (1923), and then scholars from all over the world began their research for more than a century. At present, there are a large number of legibility formulas. In 1921, Thorndikel [2] (1921) released a table of the frequency of 10,000 commonly used words, which laid the foundation for subsequent research; in 1928, Vogel & Washburne acted as the first legibility formula created using linear regression method was established, which was of original significance; subsequently, the text readability research system was gradually established. In 1949, the Flesh Reading Ease [3] (1948) model can score the difficulty of measuring text on a 100-point scale; in 1952, the Fog Index formula was created to provide a method for assessing the legibility of reading text in the upper and middle grades of elementary schools in the United States. Although the above formulas have different measurement methods and application ranges, they all focus on the measurement of vocabulary and sentences, which are somewhat restrictive. Zheng Sha [4] (2019) pointed out that the indicators created by traditional readability formulas depend too much on the two factors of the number of semantic units and the grammatical complexity, while ignoring features such as fluency between sentences and text topics; but with the advancement of computer technology, the research on text readability has also been further developed. Zhiwei Jiang [5] (2018) pointed out that new machine learning technologies such as Naive Bayes and Contrast Regression will advance the field of text legibility research's innovation and development.

In short, the existing literature has some shortcomings more or less, and the related research is mostly based on theory and idealization, and lacks experience and data support. The factors that affect the difficulty of the text are not comprehensive and need to be modified and improved. At the same time, with the application of new technologies, research in this field will continue to innovate and deepen.

1.4. Our Work

The work we have done in this problem is mainly shown in the following Figure 1.

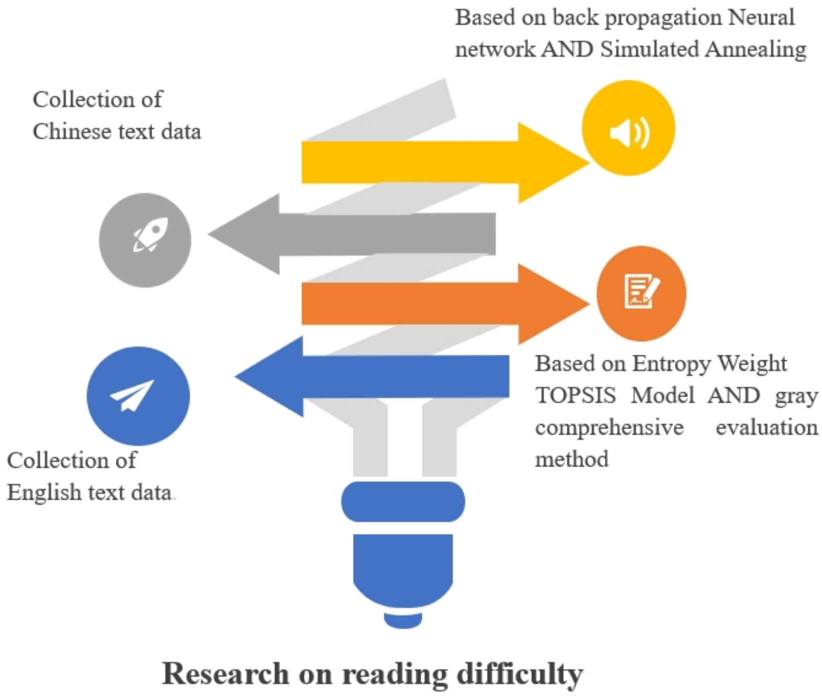

Figure 1. Ideas in this article

2. Assumptions and Justifications

Hypothesis 1: Different groups of people have the same preference for English text.

Reason: We know that different people have different preferences, and there is a greater possibility of accumulating related content that we like. For example, for articles about aircraft scientific research, readers who like airplanes usually involve some related article vocabulary more or less. Such English texts are easier to understand, but this does not represent the difficulty of this English text.

Hypothesis 2: The layout design of the English text, the paragraph design is consistent.

ISSN: 2688-8653

Reason: The purpose of layout design is to effectively combine the main visual elements of the plane: text, graphics, and colors in a certain plane space to convey accurate information in an eye-catching, organized, hierarchical, rhythmic, and rhythmic manner. Therefore, a good layout design will be easier for readers to understand and reduce the difficulty of English text. The influence of paragraph design is roughly the same, so our research needs to be carried out on this basis.

3. Notations

The key mathematical notations used in this paper are listed in Table 1.

Table 1. Notations used in this paper

Symbol	Description
i	Numbering of 319 English texts
$A_1, A_2, ..., A_8$	Each indicator
S_i	Entropy weight-TOPSIS model score
r'_{0i}	Grey relational degree
H_i	Hidden output layer of neural network
e_k	Neural network prediction error
O_k	Neural network predictive output
ω_{ij}, ω_{jk}	Neural network connection weight
Y_i	Neural network output expectations
$\alpha_1, \alpha_2, ..., \alpha_n$	Multi-objective planning dimension elimination parameter
$f'(x_1, x_2, ..., x_n)$	New solutions generated by simulated annealing

(Note: Symbols not appearing in the table will be explained when they are used.)

4. Evaluation Model Establishment and Solution

In the evaluation of the reading difficulty of English texts, this article focuses on the characteristics of the text length, total number of syllables, and proportion of syllables of the English text, Select the total number of words, sentence length, total number of text syllables, average sentence length, total number of single syllables, total number of multiple syllables, the proportion of single syllables in the full text, and the proportion of multiple syllables in the full text. And through the entropy weight-TOPSIS model and the grey comprehensive evaluation model, the reading difficulty of English text is comprehensively evaluated, and the evaluation is scientific and authentic.

4.1. Data Description

The eight indicators of the indicator system constructed in this article are shown in the following Table 2:

Among them, it contains 319 English texts related to the index data of reading difficulty. For detailed data, see the supporting materials. Some of them are shown in the following Table 3:

Table 2. Build an index system

Target layer	Dimension layer	index
The Measurement of the Reading Difficulty of English Text	Text length	Word length
		Sentence length
		Total text syllables
		Average sentence length
	Total number of syllables	Total number of syllables
		Total number of polysyllables
	Syllable ratio	Proportion of monosyllable in full text
		Proportion of multi-syllables in full text

Table 3. Raw data

Serial number	Word length	Sentence length	Total text syllables	Average sentence length	Total number of syllables	Total number of polysyllables	Proportion of monosyllable in full text	Proportion of multi-syllables in full text
1	604	37	1030	16.32	264	295	0.4371	0.4884
2	580	28	834	20.71	366	196	0.6310	0.3379
3	643	23	894	27.96	457	181	0.7107	0.2815
……								
318	308	19	432	16.21	205	92	0.6656	0.2987
319	681	41	997	16.61	453	221	0.6652	0.3245

In order to more intuitively analyze the internal characteristics and manifestations of the selected data, this article first standardizes the above data, and then visualizes the indicator data. The results are as follows:

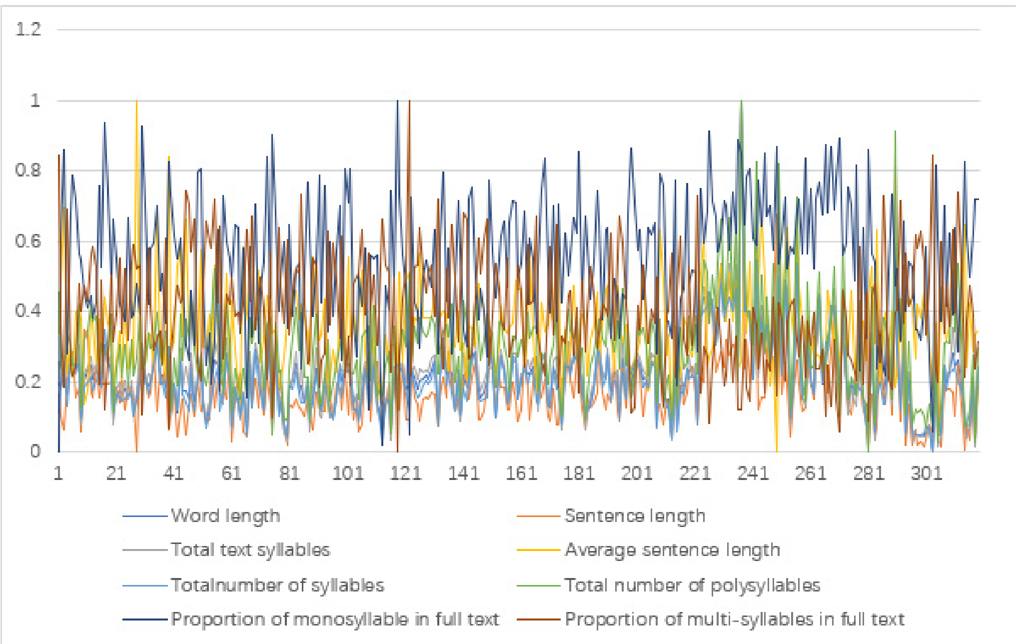

Figure 2. Raw data normalization

Secondly, in order to verify the rationality of the indicators selected in this article, we conducted a correlation analysis on the indicator data, so as to measure the closeness of the correlation

ISSN: 2688-8653

between multiple indicators and study the relationship between the indicators. The results of the correlation data for each indicator are as follows:

Table 4. Correlation Analysis

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
A_1	1.000	0.734	0.979	0.211	0.974	0.888	0.345	-0.320
A_2	0.734	1.000	0.704	-0.438	0.717	0.621	0.247	-0.288
A_3	0.979	0.704	1.000	0.229	0.924	0.953	0.224	-0.152
A_4	0.211	-0.438	0.229	1.000	0.206	0.226	0.101	-0.001
A_5	0.974	0.717	0.924	0.206	1.000	0.773	0.536	-0.491
A_6	0.888	0.621	0.953	0.226	0.773	1.000	-0.045	0.127
A_7	0.345	0.247	0.224	0.101	0.536	-0.045	1.000	-0.888
A_8	-0.320	-0.288	-0.152	-0.001	-0.491	0.127	-0.888	1.000

4.2. The Establishment of Model

4.2.1. Entropy Weight-TOPSIS Model

The entropy weight-TOPSIS model combines the entropy weight method and the comprehensive evaluation model of TIOPSIS, which can make full use of the information in the original data. Its basic idea is to use the cosine method to find the optimal solution and the worst solution from the finite solution on the basis of the normalized original data, and calculate the relative distance of each data object according to the optimal and worst solution. And calculate the relative distance of each data object according to the best and worst solution, so as to get the degree of similarity between the evaluation object and the optimal solution, and evaluate the pros and cons of each object according to the relative closeness. The basic steps of the entropy weight-TOPSIS model are as follows:

$Step\ 1$: Positive indicators

Since the 8 indicators used in this article are all extremely large indicators, the larger the indicator data, the more complicated it is to read English text, so there is no need for normalization processing.

$Step\ 2$: Standardized processing

Standardization can eliminate the difference of different index dimensions. The calculation formula is:

$$z_{ij} = x_{ij} \Big/ \sqrt{\sum_{i=1}^{n} x_{ij}^2} \tag{1}$$

If there are still negative numbers in the z matrix, you need to re-use a standardization method for the matrix. The standardization formula is:

$$\bar{z}_{ij} = \frac{x_{ij} - \min\{x_{1j}, x_{2j}, \ldots, x_{nj}\}}{\max\{x_{1j}, x_{2j}, \ldots, x_{nj}\} - \min\{x_{1j}, x_{2j}, \ldots, x_{nj}\}} \tag{2}$$

Since we did not have negative numbers in the matrix when we used the first standardized formula, there is no need to re-use another formula.

$Step\ 3$: Calculate the probability of each data and divide the individual data value by the sum of the matrix. Since there are 319 evaluation objects and 8 evaluation indicators in the data, the processed non-negative matrix is:

$$\bar{Z}=\begin{bmatrix} \bar{z}_{1,1},\bar{z}_{1,2},\bar{z}_{1,3},\bar{z}_{1,4},\cdots,\bar{z}_{1,8} \\ \bar{z}_{2,1},\bar{z}_{22},\bar{z}_{23},\bar{z}_{24},\cdots,\bar{z}_{2,8} \\ \cdots \\ \bar{z}_{319,2},\bar{z}_{319,2},\bar{z}_{319,3},\bar{z}_{319,4},\cdots,\bar{z}_{319,8} \end{bmatrix}_{319\times 8} \tag{3}$$

Calculate the probability p matrix, the calculation formula of each element p_{ij} in the matrix is as follows:

$$p_{ij}=\frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} \tag{4}$$

$Step\ 4$: Calculate the size of the information entropy e_j and information utility value d_j of each indicator, and get the entropy weight of each indicator after normalization. The calculation formula of its information entropy is:

$$e_j=-\frac{1}{\ln n}\sum_{i=1}^{n} p_{ij}\ln(p_{ij}),j=1,2,...,m \tag{5}$$

The calculation formula of the information utility value is:

$$d_j=1-e_j \tag{6}$$

$Step\ 5$: Determine the weight of each indicator

The weight of each indicator can be calculated through the results of information entropy. The calculation formula is as follows:

$$\omega_j=\frac{d_j}{n-\sum_{j=1}^{8} e_j} \tag{7}$$

$Step\ 6$: Establish the optimal vector z^+ and the worst vector z^- of index, where:

$$\begin{aligned} z^+ &= \max_{nj}(z_1^+,z_2^+,...,z_p^+) \\ z^- &= \min_{nj}(z_1^-,z_2^-,...,z_p^-) \end{aligned} \tag{8}$$

Then calculate the distance between each evaluation object and the optimal value and the worst value:

$$\begin{aligned} D_i^+ &= \sqrt{\sum_{j=1}^{8}\omega_j(Z_j^+-z_{ij})^2} \\ D_i^- &= \sqrt{\sum_{j=1}^{8}w_{ij}(Z_j^--z_{ij})^2} \end{aligned} \tag{9}$$

ISSN: 2688-8653

$Step\ 7$: Then, we can calculate the relative closeness between the $i(i=1,2,...,319)$ evaluation object and the optimal value. The calculation formula is:

$$S_i = \frac{D_i^-}{D_i^+ + D_i^-} \tag{10}$$

This paper analyzes and solves 319 English text data through MATLAB software, and obtains the information entropy value e_j, information utility value d_j and weight of 8 indicators, as shown in the Table:

Table 5. Information entropy value and information utility value

item	Information entropy e	Information utility value d	Weights
Word length	0.979	0.021	14.55%
Sentence length	0.974	0.026	18.02%
Total text syllables	0.976	0.024	16.16%
Average sentence length	0.989	0.011	7.32%
Total number of syllables	0.979	0.021	14.26%
Total number of polysyllables	0.982	0.018	12.28%
Proportion of monosyllable to full text	0.989	0.011	7.21%
Proportion of multi-syllables in full text	0.985	0.015	10.21%

Among them, the sentence length information utility value is the largest, indicating that the indicator contains the most information utility value; the information utility value of the proportion of a single syllable in the full text is the smallest, then the indicator contains the least information. The higher the information utility value, the higher the weight. It seems that this indicator has a greater impact on the difficulty of reading English text.

The weight calculated by the entropy method is substituted into the TOPSIS model, and the reading difficulty of 319 English reading texts is measured by MATLAB software. For detailed data, see the supporting materials. Part of the data is shown in the following Table:

Table 6. Entropy weight-TOPSIS score

serial number	Positive ideal solution distance (D+)	Negative ideal distance (D)	Composite score index	Rank
1	0.165650999	0.058084304	0.259611707	107
2	0.172501104	0.044348225	0.204511701	223
3	0.171026689	0.052500985	0.234874652	164
......				
318	0.197778827	0.026303126	0.117381724	319
319	0.159558322	0.05391975	0.252577463	124

4.2.2. Grey Comprehensive Evaluation Model

The grey comprehensive evaluation model is based on the grey system, aiming at the established index, and judging the influence of the corresponding index according to the grey correlation degree, so as to satisfy the evaluation of the complex system. Its basic idea is to compare the trend of changes between the sequence and the reference sequence through the degree of gray correlation, thereby obtaining the ranking of the evaluation object. The specific steps are as follows:

$Step\ 1$: In the evaluation index system, construct the following data matrix:

$$(X_1^{'},X_2^{'},...,X_n^{'})=\begin{pmatrix} x_1^{'}(1) & x_2^{'}(1) & \cdots & x_{319}^{'}(1) \\ x_1^{'}(2) & x_2^{'}(2) & \cdots & x_{319}^{'}(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{'}(8) & x_2^{'}(8) & \cdots & x_{319}^{'}(8) \end{pmatrix}_{8\times 319} \tag{11}$$

Therefore, the following equation can be obtained:

$$X_i^{'}=(x_i^{'}(1),x_i^{'}(2),...,x_i^{'}(8))^T,i=1,2,...,319 \tag{12}$$

$Step\ 2$: Determine the reference data column

$$X_0^{'}=(x_0^{'}(1),x_0^{'}(2),...,x_0^{'}(m))^T \tag{13}$$

$Step\ 3$: Dimensionless of indicator data

$$x_i(k)=\frac{x_i^{'}(k)}{\frac{1}{8}\sum_{k=1}^{8}x_i^{'}(k)} \tag{14}$$

$Step\ 4$: Form a dimensionless matrix of the data sequence

$$(X_0,X_1,...,X_n)=\begin{pmatrix} x_0(1) & x_1(1) & \cdots & x_{319}(1) \\ x_0(2) & x_1(2) & \cdots & x_{319}(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_0(8) & x_1(8) & \cdots & x_{319}(8) \end{pmatrix}_{8\times 319} \tag{15}$$

Determine the minimum difference between the two levels according to the above matrix:

$$\min_{i=1}^{319}\min_{m=1}^{8}|x_0(k)-x_i(k)| \tag{16}$$

Determine the maximum difference between the two levels:

$$\max_{i=1}^{319}\max_{m=1}^{8}|x_0(k)-x_i(k)| \tag{17}$$

$Step\ 5$: Calculate the correlation coefficient

Calculate the gray correlation coefficient, the formula is as follows:

$$\xi_i(k)=\frac{\min_i\min_k|x_0(k)-x_i(k)|+\rho\max_i\max_k|x_0(k)-x_i(k)|}{|x_0(k)-x_i(k)|+\rho\max_i\max_k|x_0(k)-x_i(k)|},k=1,...,8 \tag{18}$$

Among them, $\rho = 0.5$ is the resolution coefficient and $\{x_0(k)\}$ is the correlation coefficient.

$Step\ 6$: Calculate the degree of relevance

After the correlation coefficient of each trait is calculated, the weighted average can be used to calculate the correlation degree. The formula is as follows:

$$r_{0i}^{'} = \frac{1}{8}\sum_{k=1}^{8}\xi_i(k),(k=1,...,8) \tag{19}$$

Solve the gray comprehensive evaluation model scores by MATLAB software. See the supporting materials for detailed data. Some data are shown in the following Table:

Table 7. Gray comprehensive evaluation score

Serial number	Score	Rank
1	0.003157195	143
2	0.002969111	208
3	0.003194891	131
...		
318	0.002143155	318
319	0.003233872	121

4.3. The Conclusion of the Model

In order to more intuitively reflect the scores of 319 English texts measured by the entropy weight-TOPSIS model and the gray comprehensive evaluation model and the average comprehensive score of the two, we draw a line graph and visualize it. See appendix 2 for detailed data, and the visualization results are shown in the figure below:

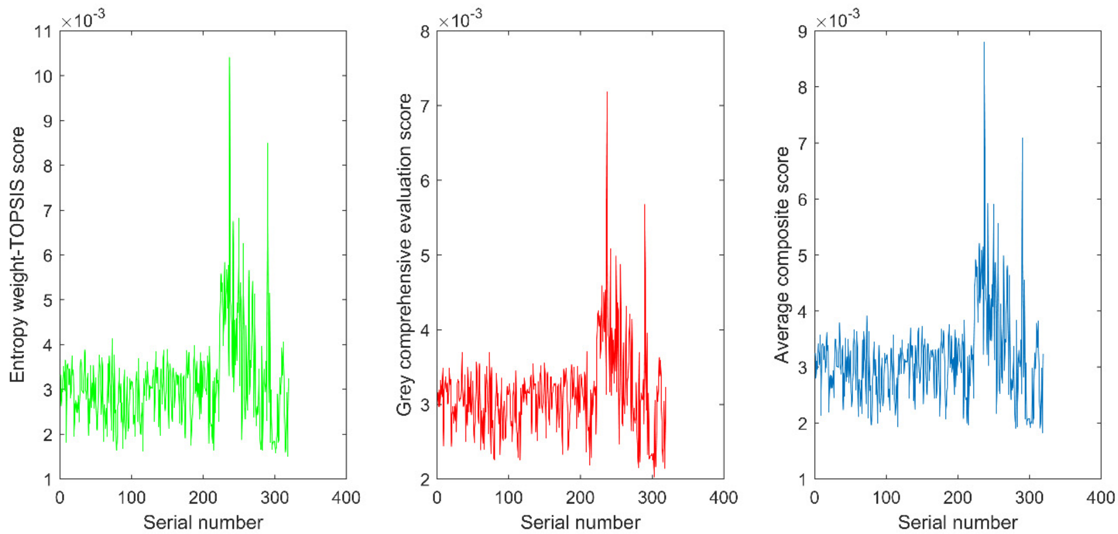

Figure 3. Comprehensive score

4.4. Classification of Model Difficulty Levels

In this paper, the normalized score results of the entropy weight-TOPSIS model and the gray comprehensive evaluation model are fitted and classified into categories A, B, C, and D. The difficulty level of each category is: A (EASY), B(MODERATE), C(DIFFICULT), D(EXTREME),

ISSN: 2688-8653

among which the English texts of category A are the least difficult to read, and the difficulty of category B, C, and D increases in turn, as shown in the following figure:

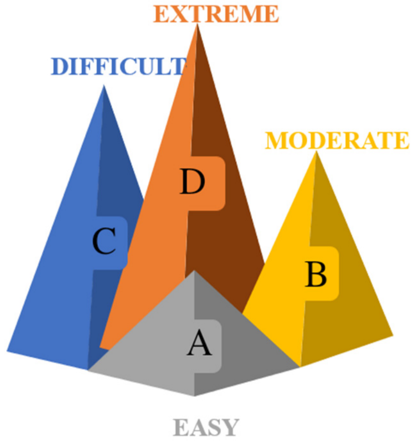

Figure 4. Criteria for scoring classification

The 319 English texts are classified according to the above rules, and the results of categories A, B, C, and D are as follows:

Table 8. Type A result

	217	58	210	167	152	113	298	296
	78	276	195	274	95	306	294	215
	156	77	86	279	64	316	183	79
	197	53	157	211	52	9	297	116
	171	258	42	138	140	175	280	283
$A[0, 0.0027)$	25	111	103	91	40	61	302	301
	63	51	27	184	105	75	80	315
	23	259	212	314	92	299	88	281
	114	289	170	96	20	300	213	318
	185	109	222	72	254	208	303	66
	104	271	132	112				

Table 9. Type B result

	214	149	54	313	189	146	118	89
	123	200	128	102	33	15	199	275
	155	126	203	161	317	192	22	62
	45	124	3	204	304	205	84	97
	119	85	127	7	71	148	24	160
	168	108	129	133	82	41	87	65
	164	13	286	243	106	57	46	81
$B[0.0027, 0.0032)$	166	220	218	16	308	121	93	44
	135	28	99	36	49	125	292	174
	194	142	295	273	4	284	143	21
	100	251	190	32	261	37	115	48
	309	193	255	31	265	182	2	262
	94	38	139	287	26	186	68	277
	147							

Table 10. Type C result

$C[0.0032, 0.004)$	73	134	8	122	56	34	165	172
	270	145	191	288	12	137	130	1
	209	177	187	245	90	181	285	70
	282	55	198	247	136	179	162	319
	312	196	43	30	202	151	163	98
	228	180	201	267	19	311	18	117
	257	76	29	83	141	10	50	101
	35	59	154	207	219	153	6	221
	310	69	266	158	39	107	307	236
	188	17	144	131	159	120	60	241
	150	206	47	5	278	173	178	216
	263	110	67	11	14	169	74	176

Table 11. Type D result

$D[0.004, +\infty)$	237	250	235	225	227	253	293	234
	290	256	233	252	272	238	232	239
	242	230	264	269	226	224	248	229
	244	268	246	291	231	260	240	223
	249							

4.5. The Improvement of Model

For any English text, we can get the above 8 index data, and add these 8 groups of index data to the original data set, that is, this text will be the 320th evaluation object. Since the original data set has 319 evaluation objects, and we have calculated the normalized scores of 319 evaluation objects, adding a set of data has little effect on the normalized score. Therefore, we use this score to classify the reading difficulty of the English text to be tested, and use MATLAB software to solve it to get the reading difficulty level of this English text.

5. BP Neural Network and Simulated Annealing

We have already measured the reading difficulty of English articles above, and extended it to any English article, how can the reading difficulty of other languages be measured? This article takes Chinese as an example, considering that the characters, words, semantics, and grammar of the text will all have an impact on the difficulty of reading Chinese. In this regard, we use a neural network model to measure, and we can determine the reading difficulty of the Chinese text to be measured. And without changing the content of the Chinese text, the model annealing algorithm based on multi-objective programming is used to reduce the difficulty of the Chinese text.

5.1. BP Neural Network Model

BP neural network is a kind of multi-layer feedforward neural network based on error back propagation. The basic idea is gradient descent method. The main feature of this network is signal forward transmission and error back propagation. In the process of forward propagation, the input signal is processed by the hidden layer from the input layer to the output layer. The neuron state of each layer only affects the neuron state of the next layer. If the output layer cannot get the expected output, it will switch to back propagation, adjust the network weight and threshold according to the prediction error, so that the predicted output of the BP neural network is constantly approaching the expected output [6].

Using BP neural network can predict the initial reading difficulty of Chinese text, and the prediction error is small. Here, the meaning, semantics, and grammar are used as explanatory variables, and the reading difficulty of the Chinese text is taken as the explained variables. The topological structure of BP neural network is as follows:

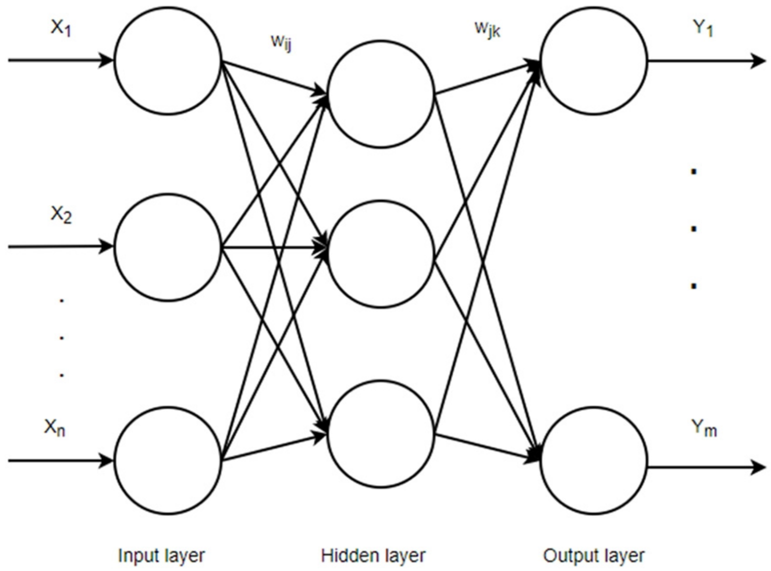

Figure 5. BP neural network structure

Among them, $X_1, X_2, \ldots, X_n$ is the input value of the BP network, and $Y_1, Y_2, \ldots, Y_n$ is the predicted value of the BP neural network. ω_{ij} and ω_{jk} are the weights of the BP neural network. Before the BP neural network predicts, the network must be trained first, and the network has associative memory and prediction capabilities through training [7]. The training process of BP neural network includes the following steps:

$Step\ 1$: Network initialization. According to the system input and output sequence (X, Y) to determine the number of nodes in the input layer of the network n, the number of hidden layer nodes l, the number of nodes in the output layer m, initialize the connection weights ω_{ij} and between the input layer, hidden layer and output layer neurons, initialize the hidden layer threshold a, the output layer threshold b, given the learning rate and neuron activation function [8].

$Step\ 2$: Hidden layer output calculation. According to input variables X, input layer and hidden layer connection weights ω_{ij} and hidden layer threshold a. Calculate hidden layer output H.

$$H_i = f\left(\sum_{i=1}^{n} \omega_{ij} - a_j\right) j = 1, 2, \ldots, l \tag{20}$$

In the above formula, l is hidden layers, f is hidden layer activation function, this function has a variety of expressions. We use the sigmoid function, namely:

$$f(x) = \frac{1}{1 + \exp(-x)} \tag{21}$$

ISSN: 2688-8653

$Step\ 3$: Output layer calculation. According to the hidden layer output H , connect the weight ω_{jk} and the threshold to calculate the predicted output O of the BP neural network.

$$O_k = \sum_{j=1}^{l} H_{j\omega_{jk}} - b_k, k = 1, 2, \dots, m \tag{22}$$

$Step\ 4$: Error calculation. According to the hidden layer output O and expectation Y , calculate the network prediction error e .

$$e_k = Y_k - O_k, k = 1, 2, \dots, m \tag{23}$$

$Step\ 5$: The weight is updated. Update the network connection weight ω_{ij}, ω_{jk} according to the network prediction error e .

$$\begin{gathered} \omega_{ij} = \omega_{ij} + \eta H_j (1 - H_j) x(i) \sum_{k=1}^{m} \omega_{jk} e_k \\ \omega_{jk} = \omega_{jk} + \eta H_j e_k \end{gathered} \tag{24}$$

In the above formula, $i,j = 1, 2, \dots, l$, $k = 1, 2, \dots, m$, η is learning rate.

$Step\ 6$: The threshold is updated. Update the node threshold a, b according to the network prediction error e .

$$\begin{gathered} a_j = a_j + \eta H_j (1 - H_j) \sum_{k=1}^{m} \omega_{jk} e_k \\ b_k = b_k + e_k \\ j = 1, 2, \dots, l; \quad k = 1, 2, \dots, m \end{gathered} \tag{25}$$

$Step\ 7$: Judge whether the iterative algorithm is over, if it is not over, return $Step\ 2$.

For any Chinese text, we can use the above model to solve the expectation Y_k of the Chinese text. Secondly, the initial difficulty score of each Chinese text can be calculated through MATLAB software.

5.2. Simulated Annealing Algorithm

The simulated annealing algorithm is similar to the solid annealing process. It adopts the Metropolis criterion and is a probability-based algorithm. The solid is heated to a sufficiently high level and then slowly cooled. When the temperature is heated, the internal particles of the solid become disordered with the increase in temperature. The internal energy increases, and the particles gradually become orderly when cooled slowly, reaching an equilibrium state at each temperature, and finally reaching the ground state at room temperature, and the internal energy is reduced to a minimum [9]. The search process of the simulated annealing method introduces random factors and accepts a solution that is worse than the current solution with a certain probability. Therefore, it is possible to jump out of this local optimal solution and find the global optimal solution.

In the difficulty measurement of Chinese text, this paper firstly transforms it into a multi-objective planning problem. Since people always hope that the basic content of each Chinese text remains unchanged and the reading difficulty is reduced, the use of simulated annealing algorithm will reduce the difficulty of reading Chinese text without changing the content of the Chinese article.

ISSN: 2688-8653

Since the reading difficulty of Chinese text is affected by multiple indicators, this problem is a multi-objective planning problem. Secondly, considering the different dimensions of each indicator, the logarithmic function is introduced to construct the objective function of the multi-objective programming, namely:

$$f(x_1, x_2, \ldots, x_n) = \ln^{\alpha_1} f(x_1) + \ln^{\alpha_2} f(x_2) + \ldots + \ln^{\alpha_n} f(x_n) \tag{26}$$

$Step\ 1$: Set the current optimal solution

Let $T = T_0$, that is, the initial temperature at which annealing is started, the initial solution of the i -th area $x_i(0)$, and calculate the corresponding objective function value $f(x_i(0))$

$Step\ 2$: The difference between the new solution and the current solution

Perturb the current solution $x_i(n)$ to produce a new solution $x_i(m)$, calculate the corresponding objective function value $f(x_i(m))$, to get $\Delta f = f(x_i(m)) - f(x_i(n))$.

$Step\ 3$: Determine whether the current solution is accepted

Using Metropolis criterion, if $\Delta f < 0$, the new solution $x_i(m)$ is accepted; if $\Delta f > 0$, the new solution $x_i(m)$ is accepted with probability $\exp\left(-\frac{f(x_i(m)) - f(x_i(n))}{T_i}\right)$, T_i is accepted with probability. That is, the formula is as follows:

$$p = \begin{cases} 1 \\ \exp\left(-\frac{f(x_i(m)) - f(x_i(n))}{T_i}\right) \end{cases} \tag{27}$$

$Step\ 4$: When the new solution is determined to be accepted, the new solution $x_i(m)$ is taken as the current solution.

$Step\ 5$: Loop the above steps

At temperature T_i, repeat the perturbation and acceptance process k times, and then proceed to the next step.

$Step\ 6$: Find the global optimal solution

Judge whether the temperature T reaches the termination temperature T_f , if it reaches, terminate the algorithm; reverse and continue to execute the cycle step.

$Step\ 7$: Set cooling schedule

Determine the initial temperature: Generally speaking, the greater the initial temperature, the greater the probability of obtaining the global optimal solution, but the longer it will take. Only a large enough T_0 can meet the algorithm requirements, but the scale of processing for different problems is different, the selection of the initial temperature is quite different. For this question, after repeated debugging, we choose the initial temperature $T_0 = 200$.

Attenuation coefficient α for controlling temperature T : Different annealing methods have different temperature drops. Exponential cooling is the most commonly used annealing strategy. Its temperature changes are very regular and directly related to parameters. It is the main object of our research.

The attenuation coefficient α is a constant very close to 1, and is generally taken as $0.5 \sim 0.99$. Its value determines the cooling process. A small amount of attenuation may lead to an increase in the number of iterations of the algorithm process, so that the algorithm process accepts more transformations, searches for a larger solution space, and returns a higher quality final solution. Choose $\alpha = 0.95$ in this question.

Markov chain length L_k: The simulated annealing algorithm continues the iterative process of "generating new solutions-judging new solutions-accepting or discarding new solutions" within the length of the Markov chain, which corresponds to the process of regional thermal equilibrium of solids at a certain constant temperature. If an infinite number of iterations is performed at a certain temperature, the corresponding Markov chain can achieve a stable distribution probability [10].

The selection of the Markov chain is also closely related to the decrease of the temperature control parameterT_k , and a slow decrease can avoid an excessively long Markov chain. Under the premise that the attenuation wash of the control parameter has been selected, each value can accurately reach the quasi-balanced state.

According to this principle, $L_k = 100N$ is generally taken, where N is the scale of the problem, combined with the data of this question, here is selected $L = 1000$, and the cooling formula is:

$$\begin{cases} t_{k+1} = \alpha t \\ t_k = \dfrac{L-K}{L} - t \end{cases} \tag{28}$$

The termination temperature t of the control parameter: A reasonable stopping criterion is to ensure that the algorithm converges to a certain approximate solution, and can make the final solution have a certain globality [11]. The termination of the iteration can usually be judged based on the number of iterations or the termination temperature or the solution of the iteration process in several successive chains does not change. For this question, the termination temperature is set as $T_f = 10$.

Input the independent variable data, and use MATLAB software to solve the original reading difficulty and the reduced reading difficulty. Take a 2021 college entrance examination reading question article as an example, the optimization results are as follows:

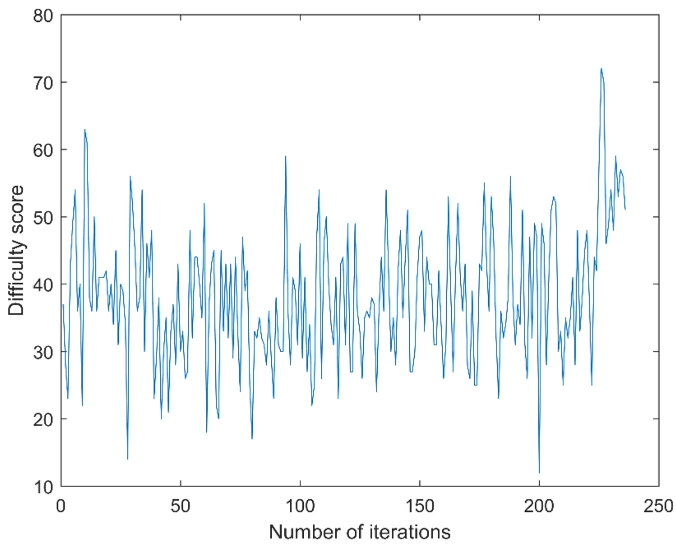

Figure 6. Results of simulated annealing iteration

From the figure, we can see that the initial reading difficulty of the Chinese text is 37, and the difficulty after optimization by the simulated annealing algorithm can be reduced to 13, and the text difficulty reduction effect is greater, which proves that the model can greatly reduce the difficulty of reading Chinese text.

6. Sensitivity Analysis

In the process of using the simulated annealing algorithm, the initial temperature T_0 , termination temperature T_f , attenuation coefficient α , and Markov chain length L will all affect the results of the model.However, considering that the attenuation coefficientα has a greater impact on the optimal solution, we take the step size of 0.02 that$\alpha=0.93$， $\alpha=0.97$ to perform sensitivity analysis, and the results are as follows:

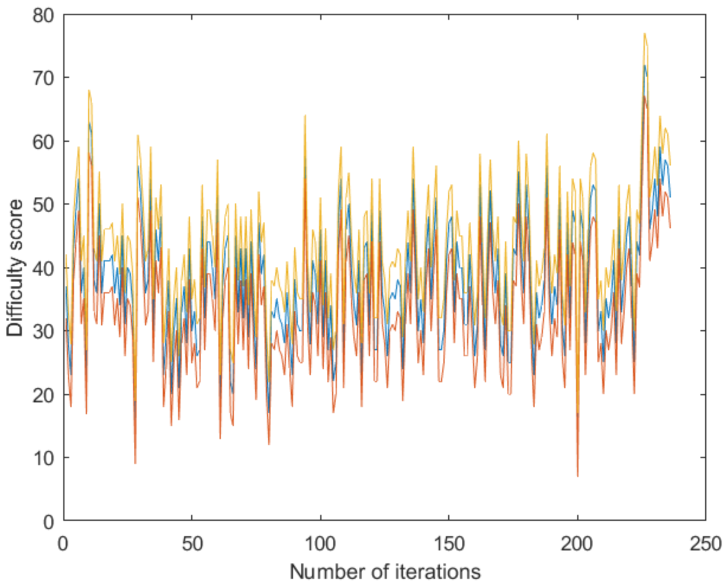

Figure 7. Sensitivity analysis results

According to the sensitivity analysis results, the reading difficulty of Chinese text fluctuates around a center within a certain range. It can be seen intuitively that the sensitivity of the attenuation coefficient α is not obvious. Therefore, the fluctuation of the attenuation coefficient within a certain range has a small impact on reducing the difficulty of reading Chinese text.

7. The Advantages and Disadvantages of the Model

7.1. Strengths

Strength 1: The model used in this article is based on the full mining and processing of data. Through the evaluation and analysis of different text data, the relationship between the difficulty of reading English text and its influencing factors is extracted, and the index system is constructed from multiple perspectives. And it has carried on the correlation analysis to it, which has high rationality.

Strength 2: This paper uses a model combining entropy weight-TOPSIS and gray comprehensive evaluation to evaluate the reading difficulty of English texts. Compared with a single model, it eliminates the interference of subjective factors in the evaluation, and considers

ISSN: 2688-8653

the correlation between multiple factors. The thinking is simple. The loss and error caused by information asymmetry are reduced to a large extent, and it is more objective and scientific.

Strength 3: The BP neural network is used to predict the reading difficulty of Chinese text. This model has the function of realizing any complex nonlinear mapping, and solves complex problems of internal mechanisms through learning. In addition, the network has certain promotion and generalization capabilities.

Strength 4: In reducing the difficulty of reading Chinese text, this paper uses multi-objective programming to construct an objective function, and uses simulated annealing algorithm to solve it. It is robust and can find a stable global optimal solution.

7.2. Weaknesses

Weaknesses 1: In the entropy weight-TOPSIS model, the relevant factors that affect the difficulty of reading English text are still too one-sided. Although the later classification test effect is relatively satisfactory, it still lacks a certain degree of credibility.

Weaknesses 2: The selection of evaluation indicators and the rating of reading difficulty are still inevitably affected by subjective factors.

7.3. Further Discussion

7.3.1. Model Improvement

(1) In the construction of the English text evaluation index system, it is difficult for us to construct a perfect index system, because there are many indicators that affect the difficulty of English text, such as logical structure, local language, etc., and we cannot cover all possible indicators.

(2) In the measurement of Chinese reading difficulty, we use machine learning and choose the BP neural network prediction model, and its learning efficiency is unstable.

7.3.2. Model Extension

Practicality: In the current Internet age, information is too explosive. Through our model, information can be processed digitally, English text evaluation, and text suitable for reading can be found, which is highly efficient and applicable.

Feasibility: Our data combines more than 300 English texts in different books, and has relatively reliable data support in a limited time. At the same time, the TOPSIS model based on the entropy weight method is a commonly used comprehensive evaluation method that can be fully utilized the information of the original data is weighted objectively, and the result can objectively reflect the gap between the evaluation schemes.

8. Conclusion

Reading ability is very important in the process of language learning. With the in-depth development of globalization, the demand for reading and application of English texts is also increasing. It is still time-consuming to find English text materials that match the reading ability of different people. This highlights the importance of studying the legibility of English texts.

This paper first selects eight evaluation indicators from the dimensions of sentence length, word length and total number of syllables to construct an English text legibility evaluation system, and uses entropy weight-TOPSIS and grey comprehensive evaluation to obtain the reading difficulty scores of 319 typical English texts. The evaluation grades can be expanded to measure the difficulty of reading any English text.

Chinese is the most used language. Here, we select multiple explanatory variables and use BP neural network to predict the reading difficulty of Chinese text. Multi-objective planning is used to construct the objective function, and the simulated annealing algorithm is used to reduce the

ISSN: 2688-8653

difficulty of reading Chinese texts and reduce the reading obstacles of readers without changing the content of the text.

Finally, after checking the sensitivity of the relevant model, the advantages and disadvantages of the model are analyzed, and reasonable optimization and expansion methods of the model are proposed.

References

[1] Lively, B., Pressey, S. A method for measuring the vocabulary burden of textbooks[J]. Educational administration and supervision. 1923, 73 (09) : pp 389-398.

[2] Thorndike,E.L.,1921.The teacher's word book.

[3] Flesch,R,1948.A new readability yardstick. Journal of applied psychology 32,221.

[4] Zheng Sha, Based on the deep learning English text readability measurement research[D]. Chongqing university, 2019. The DOI:10.27670/ d.cnki.gcqdu.2019.001561.

[5] Zhiwei Jiang, Research on Text representation Technology for Readability Evaluation[D]. Nanjing University,2018.

[6] Fujin Zhuge. WSN-based wind power busway health monitoring and evaluation system[D]. Harbin Institute of Technology, 2018.

[7] Kang Zhao. Research on Intelligent Monitoring Technology for Structural Damage of Jack-up Offshore Platform Based on Vibration Information[D]. China University of Petroleum (East China), 2016.

[8] Chunyang Liu. Research on Early Warning of Corn Price in Jilin Province [D]. Jilin Agricultural University, 2016.

[9] Boye Jiang. Research on Urban Logistics Distribution Route Optimization Based on Improved Simulated Annealing Algorithm [D]. Shijiazhuang Railway University, 2018.

[10] Changlong Li. Research on Distribution Network Optimization Based on Ant Colony Algorithm [D]. Chongqing University, 2004.

[11] Jiangkun Liu. Research on Node Localization Optimization Algorithm for Wireless Sensor Networks [D]. Kunming University of Science and Technology, 2019. DOI:10. 27200/ d.cnki. gkmlu. 2019. 001138.