A study on predicting the results of Wordle guessing game based on time series and linear regression

Chenhao Liu*

Faculty of Materials and Manufacturing, Beijing University of Technology, Beijing China, 100124

*Corresponding author: 18811586515@163.com

Abstract. Wordle is a globally popular guessing game in which players guess a five-letter word in six or fewer attempts. To explain the changes in the number of players on a daily basis and to make predictions about the number of users playing on a future date, this paper uses the contagion model as the basis for an analogy to explain the reasons for the changes in the data. Since the contagion model is not very accurate in predicting the future, considering that the data itself is a set of time series data, this paper uses a time series prediction model to predict the future data and gets a good fit. In order to predict the distribution column of the number of guesses for a word, this paper performs regression fitting of word attributes and data distribution based on different eigenvalues of different words to get the change curve of eigenvalues, and considers the uncertainty of the model based on the obtained functional relationship, and then the model obtained by replacing the words to be predicted results is tested and error analyzed.

Keyword: Time series, Linear Regression, Contagion Model, Guessing Game.

1. Introduction

Wordle is a popular word-guessing game that requires players to guess a five-letter word in six or fewer attempts, and players can share their data after winning. However, the number of daily active users of Wordle has been decreasing in recent years. Therefore, in this paper, based on the data provided on Twitter, which is assumed to be real, reliable, valid and without machine involvement, we build mathematical models to predict and explain the reasons for the change in the number of reported results[1] and create a prediction interval for the number of reported results on March 1, 2023 [2][3]. And consider whether the percentage of reported scores performed in Difficult mode is affected by any of the attributes of the words.

It is also observed that the distribution of word results varies a lot from day to day[4][5], so for the variation in the distribution of word results, this paper will also build a model to predict the distribution of reported results for any future day and make a prediction for the EERIE on March 1, 2023.

Although there is no previous paper for Wordle's prediction, there are many practical cases in some similar fields for reference. Based on the traffic flow in Shenyang in recent years, Yu Denghui[6] constructed a gray prediction model by accumulating, subtracting, and matrix transforming to get the data change pattern, and backpropagated to achieve the logistics park logistics flow data prediction in the next three to ten years. The system layout planning（SLP） is then applied to conduct comprehensive correlation analysis, which solves the logistic relationship between functional areas and the non-logistic relationship between functional areas based on the predicted traffic flow, the conflict between functional areas and the type of goods, and then evaluates the weights to obtain the comprehensive correlation degree. Then, the area of each functional area is derived based on the predicted material flow. And finally, the layout of logistics park is improved with multi-objective migratory bird migration algorithm.

In another paper, Dandan Wang[7] applies the SEIR model to analyze the transmission mechanism and the impact of prevention and control measures of the novel coronavirus pneumonia outbreak, which has spread rapidly and deeply around the world. They also considered the impact of the system by triage of mildly ill patients and birth and natural death of the population. Based on this, the basic regeneration number of the transmission model was solved using the regeneration matrix method,

and the sensitivity of the system threshold to parameters such as the proportion of actively isolated groups and the proportion of follow-up detection was explored using the normalized sensitivity index.

Therefore, based on the relevant research results, this paper focuses on predicting the change pattern of the number of players of Wordle over time and the distribution of the number of attempts per word. And propose a prediction scheme that can work. To make the changes in the data more visible and understandable, we visualize the results by building a time series forecasting model. However, prior to building the model, it is particularly important to pre-process the data in advance as there may be outliers and incorrect values in the dataset, which can lead to inaccuracies in the model. Once the data processing is complete, attention is paid to the trend of the data results over time. Considering that this variation should be continuous, equations are established using time and Number of reported results as the two variables to obtain a function of the data results over time, and subsequent results are predicted and fitted according to the model. We can therefore estimate the approximate range of data outcomes on March 1 2023. In analyzing the effect of word attributes on the percentage of scores reported that were played in Hard Mode, and the correlation with the percentage of scores reported that were played in Hard Model was analyzed by summing the probability of occurrence of letters as word attributes. The relationship between word attributes and the percentage of difficulty was obtained. When analyzing how to build the model and predict the distribution of future outcome data, this paper assumed that all data were obtained from normal players playing, the model can be fitted with a regression of word attributes and data distribution, and the change curve of feature values with data distribution can be obtained according to the different feature values of different words, and the uncertainty of the model can be considered according to the function relationship obtained; Afterwards, the ERRIE is substituted into the obtained model for testing and error analysis. This can be done by substituting the trained model into the data and comparing it with the known results.

2. The variation of reported result

2.1. Prediction of the number of reported results

The number of results reported in Wordle varies from day to day. In order to be able to accurately analyze the changes in reported results on a daily basis, this paper uses the date of the report collected from Twitter as the horizontal coordinate and the reported results as the vertical coordinate.

Looking at the Figure 1. The data grows in a near-exponential model for a month starting on January 7, 2022, and reaches an extreme value of 361,908 on February 2, 2022, just one month later. after that, the number of reported results drops rapidly again, and after another month the value becomes half of the extreme value, with no stopping trend.

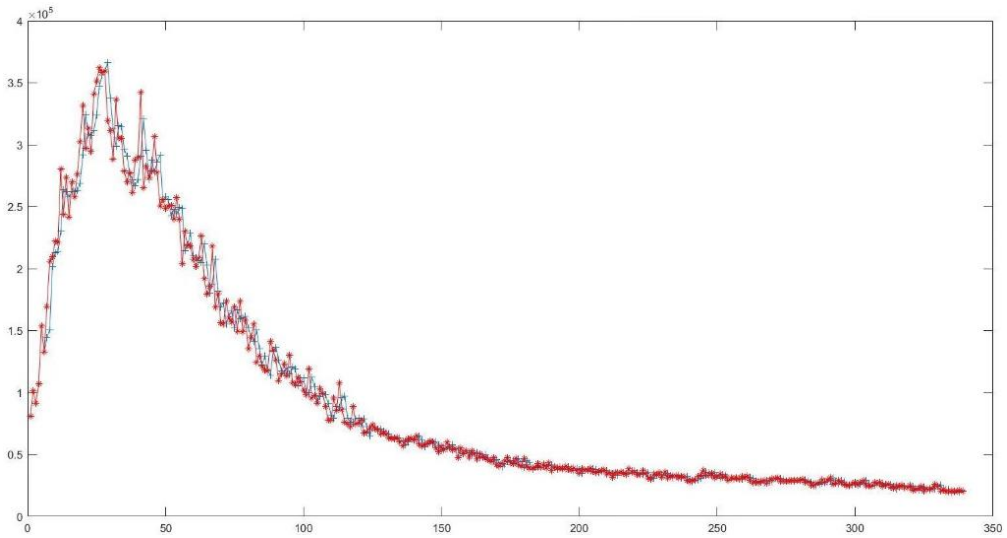

Figure 1. Comparative chart of reported volume projection results

In the third month, April, the values have dropped below 100,000, reaching about a quarter of the peak. Entering the five-digit range, the trend of data changes further back gradually becomes flat, bringing the values mostly concentrated in this area. The time required for each 10,000 reductions in data also becomes longer and longer, reaching the final region of 30,000 versus 20,000, with about 160 given values, accounting for half of the entire sample space. At this point the curve is very flat and the slope becomes progressively smaller, but the overall trend still continues downward. To get the data interval for March 1, 2023, our solution is to use a model to predict the data obtained for March 1 and multiply it by the average error rate of this model to obtain it. In this paper, we use the time series forecasting model to analyze the data, a model that speculates on future changes in the values of the study variables based on the observed values of the variables themselves and their patterns of change. So, the prediction of the number of reported results getting from this model for the March 1,2023 are 20662, and the average error rate of the model per day of data is 0.0621%. This means that the data for March 1, 2023 is calculated as

$$（20662 * 0.9379，20662 * 1.0621）$$

After rounding, the result of the calculation is

$$（19379，21945）$$

2.2. Explanation of the reported result

While time series models can help to accurately show changes in the number of users, they cannot explain the reasons for such changes very well. Therefore, we need to explain it by introducing other models. Here we introduce an infectious disease model[8] called SEIR. In a simple Susceptible-Exposed-Infected-Removed (SEIR)[9] model, susceptible are people who have not been infected with the infectious disease and have been healthy so far, that is, most of the population, exposed are people who are in the incubation period of the infectious disease, infected are people who have been diagnosed, removed are people who have been removed (including those who recovered and those who died).

Figure 2 is the schematic:

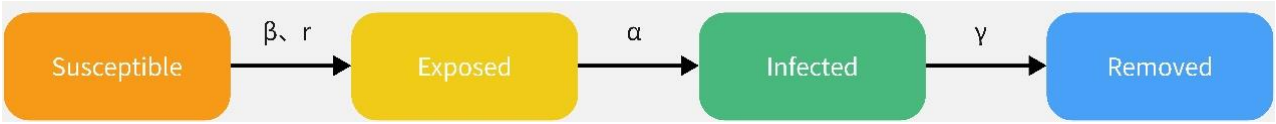

Figure 2. The Principle of SEIR Mode

All the above parameters are greater than 0.

The spread of a virus in a population can then be described by the following nonlinear ordinary differential equation:

$$\frac{dS}{dt} = -\frac{r\beta SI}{N} \tag{1}$$

$$\frac{dE}{dt} = \frac{r\beta SI}{N} - aE \tag{2}$$

$$\frac{dI}{dt} = aE - \gamma I \tag{3}$$

$$\frac{dR}{dt} = \gamma I \tag{4}$$

Above equations represent the rates of change of S, E, I, and R over time, respectively. We can substitute Wordle's data into this model. Susceptible represents those who have not been exposed to Wordle. Exposed are those potential players, they haven't played the game yet, maybe they've seen such data on Twitter, maybe they've heard the name of the game and are ready to play it. Infected are those who are playing Wordle and who are willing to share their attempts daily. In the end, removed are those who quit the game and will not play again.

After given a small initial value, here comes to the reason. In the beginning, Wordle quickly

became a topic of communication and a fashion on the Internet through various advertisements or recommendations of other users or friends. Every day more and more people were exposed to the game and the number was increasing rapidly every day. It is similar to a virus called Wordle that is spreading rapidly on the Internet and making people infected and carrying it. The number of people playing Wordle is rapidly reaching a peak. But with the sudden explosion of Wordle, research on it has become more and more popular. People gradually discovered strategies to get through the game faster by looking for patterns and constant research. People gradually lost interest in the game when they realized that the game was no longer difficult to play. The number of reported results will tend to smooth out and eventually decrease and disappear.

We assume that

$$\boldsymbol{\alpha = 0.125, \beta = 0.6}$$

Figure 3 shows the predictions for infectious diseases by this model.

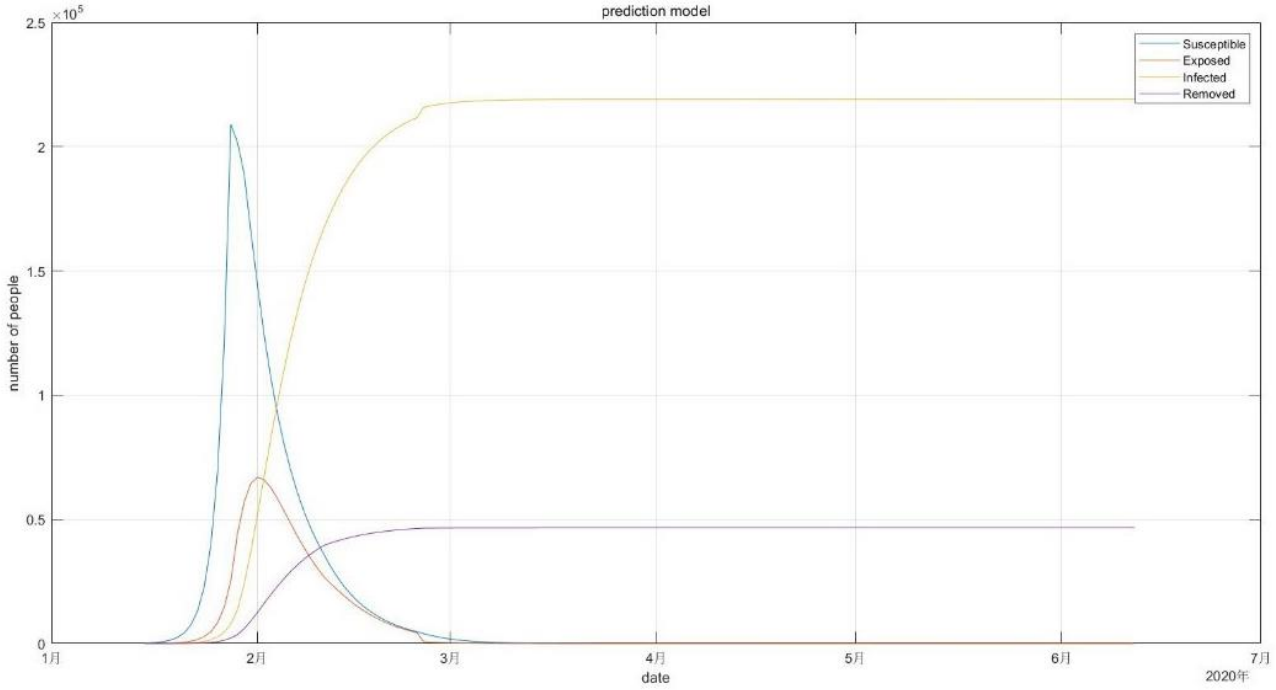

Figure 3. The Variation Trend of SEIR Model

Interestingly, the curve is not smooth, but has many sawtooth-like folds. This paper thinks this phenomenon occurs because there is a difference in the difficulty of the words from day to day. People are more willing to share those attempts where they solve the problem faster, so when the words are more difficult, the number of people sharing their scores will be smaller and vice versa.

Table 1. The Frequency of Letter in the word

Letter	Frequency (%)
A	8.167%
B	1.492%
C	2.782%
D	4.253%
E	12.702%
F	2.228%
G	2.015%
H	6.094%
I	6.966%
J	0.153%
K	0.772%
L	4.025%
M	2.406%
N	6.749%
O	7.507%
P	1.929%
Q	0.095%
R	5.987%
S	6.327%
T	9.056%
U	2.758%
V	0.978%
W	2.360%
X	0.150%
Y	1.974%
Z	0.074%

So, we parsed the attributes of words in the hope of establishing a functional relationship between word attributes and the number of outcomes. In this paper, we reviewed data published by the American Mathematical Society on the frequency of use of each letter in words[10]. The frequencies are shown in Table 1. Each word was split into five letters, and the probability of occurrence of each letter was summed to get the feature value exclusively for that word.

Also, it was found experimentally that if the same letter appears multiple times in a word, it becomes more difficult for the word to be guessed. Therefore, in order to ensure that the feature values can better reflect the properties of words, this paper penalizes letters that appear multiple times by adding the frequency of that letter only once when calculating the feature values if a certain letter appears multiple times, which can reduce the feature values of the words. The modified feature value corresponding to each word can be obtained by the above method. Table 2 below is an example of some word feature values.

Table 2. Excerpt of eigenvalues

Word	Feature Values
Manly	23.321
Molar	28.092
Havoc	25.528
Impel	28.028
Condo	21.291
Judge	21.881
Poise	35.431
Aorta	20.717

Divide the number of people in hard mode on the same day by the total number of people to get the percentage of scores reported that were played in Hard Mode. Use this percentage to do correlation analysis[11] with the eigenvalues and explore the results.

Table 3. Percentage-Weight

	Percentage	Feature Values
Percentage	1(0.000***)	-0.037(0.494)
Feature Values	-0.037(0.494)	1(0.000***)

In Table 3 we can see that the value obtained is 0.494, which is much greater than 0.05, the criterion for judging the correlation between the two variables. **We conclude that the attributes of the word do not affect the percentage of scores reported that were played in Hard Mode.**

3. The distribution of the word

3.1. Prediction of the reported result

In order to be able to successfully predict the distribution of data results, when choosing a prediction model, it is important to consider whether it can meet our needs. The first thing that comes to mind in this paper is to reflect the word by summing up the probability of each letter appearing in the word as an attribute value earlier.

After obtaining the eigenvalues, we use the eigenvalues and 1, 2, … 7 times to obtain the corresponding curves and obtain the functional relationship. The scatter plot is drawn with the percentage as the vertical coordinate and the eigenvalue as the horizontal coordinate. By observing the distribution of the data, we find that the distribution is more in the middle and less on the two sides, proving that most of the data points are clustered next to the fitted straight line rather than symmetrically. This also shows that our model is very accurate and reliable in predicting word outcomes. For the prediction of word outcomes, the eigenvalues of the words can be derived from our model and then substituted into a seven-unit linear model to obtain the distribution of outcomes for any five-letter word. Figure 4 shows the distribution of results we obtained after substituting EERIE into the model.

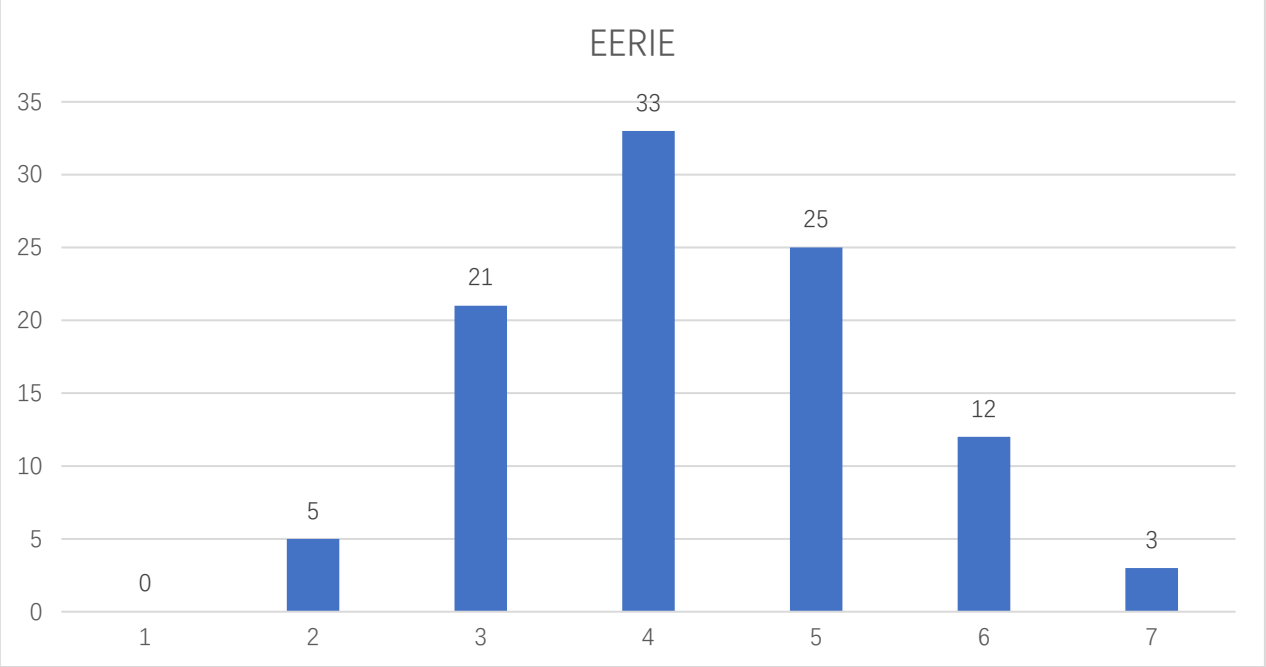

Figure 4. The prediction of EERIE

At the same time, however, there are uncertainties in the model: firstly, because it is a linear regression equation, there is a limit to the interval of the linear equation, and when the eigenvalues of the words are outside this range, the substitution function may result in negative values, giving obvious errors. Secondly, because we assume that the frequency of each letter is equal to the frequency of each letter across all words, but in practice, for words with only five letters, the letter frequencies are not equal, so there is a degree of error.

For the above case we considered the effect of different initial letter beginnings on the eigenvalues by adding the initial letters to the calculation of the special values as well.

In the subsequent modeling, we proposed to fit the regression results with multivariate linear equations, considering that building multiple unit linear regression equations may make the errors in the superposition process larger. For the multiple linear equations, due to the large variety of variables requiring input and output, the need for multiple inputs and outputs at the same time, and the input and output variables are not the same variables, this paper finally chooses to use the regression chain model for prediction and adopts the weighted random forest algorithm based on the particle swarm algorithm for classification, and optimizes the random forest model by selecting the number of decision trees through parameters such as iteration and pruning threshold. Meanwhile, in order to solve the voting weight problem, a part of training samples is used as prediction samples to calculate the weight of each decision tree to ensure its fairness in voting. In this paper, we simplify the calculation of weights by using the correct classification rate of the prediction samples as the weights of each decision tree. The reason is that overly complex weights will greatly increase the training time, while the simpler correct rate weighting method optimized by the particle swarm algorithm can also ensure the classification accuracy and eventually get more accurate results through the neural network[12]. Figure 5 shows the distribution of data results after optimization.

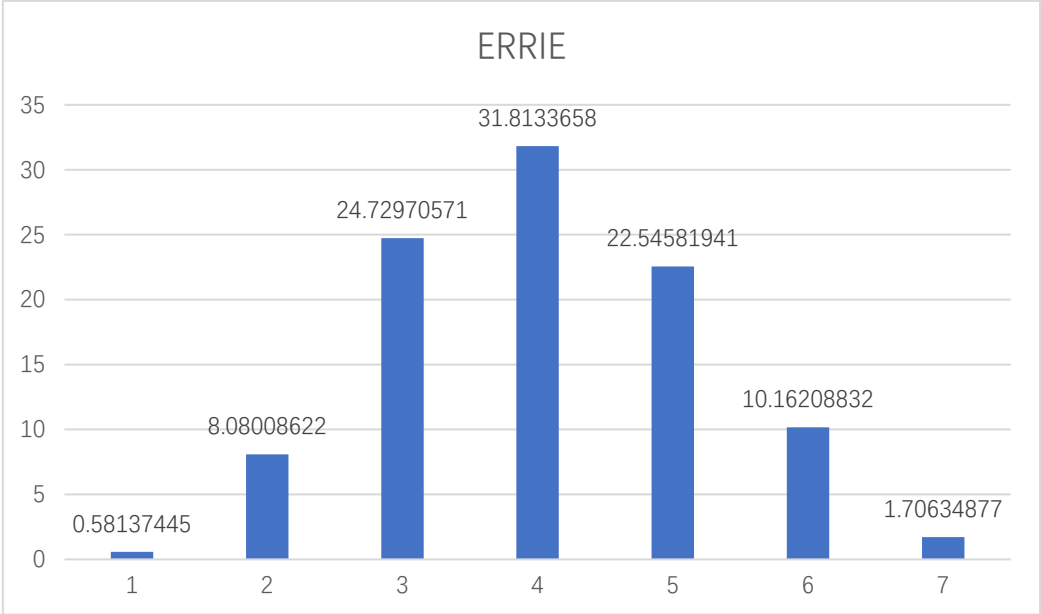

Figure 5. More accurate prediction of EERIE

The steps of the particle swarm optimization weighted random forest algorithm are as follows[13].

$$\omega l = \frac{X_{correct,l}}{X}, l = 1,2,\dots,L, \tag{5}$$

$$f_{WRF}(x) = \arg max\left\{\sum_{f_{tree,l}(x)=i} \omega_l\right\} l \epsilon L, i = 1,2,\dots,c \tag{6}$$

$$f_{RF}(x) = \arg max\{I(f_{tree,l}(x) = i)\} i = 1,2,\dots,c \tag{7}$$

1.Determine the parameters of the algorithm, setting random initial values for the pruning threshold, the number of decision trees L, the number of pretest samples X, and the number of random attributes m.

2. Using Bootstrap algorithm sampling, produce L random training sets and select X prediction trial samples in each training set.

3. Generate decision trees using the remaining samples from each training set separately, a total of L. During the generation process, m attributes are selected from all attributes as the decision attributes of the current node before each attribute selection.

4. When the number of samples contained in the node is less than the threshold, the node is taken as a leaf node and the plurality of its target attributes is returned as the classification result of the decision tree.

5. When all decision trees are generated, pretest each decision tree and calculate its weights using figure 1.

6. Calculate the classification result of the model using equation 6.

7. Use the classification result as the fitness value and iterative optimization of the parameters mentioned in Step1 using the particle swarm algorithm to determine the parameters of the final model.

4. Conclusion

Wordle is a very popular game, and his developers publish daily the response data of the day. In this paper, we investigate the extent and reasons for the variation of the response data, and also predicts the future data of the game through SEIR model, linear regression analysis model and random forest model.

In order to be able to accurately predict the number of Wordle users, the change in the number of reports needed to be studied and the future number of reports needed to be predicted. By pre-processing the data, this paper finds that the change trend of the model matches with the infectious disease model,

so in the process of explaining the reason for the change of data, we use the infectious disease model as the basis for analogy. However, since the infectious disease model is not very accurate in predicting the future, considering that the data itself is a set of time series data, this paper uses a time series prediction model to predict the future data and gets a good fit.

In addition, in order to be able to accurately predict the distribution of word outcomes, this paper uses the sum of the frequencies of the individual letters in the words as the eigenvalues to quantify the words. Given the need to examine the relationship between the eigenvalues and the seven variables, this paper examined the relationship between the eigenvalues and each variable separately using a one-dimensional linear regression method, and obtained seven functions with high confidence. Later, in order to make the prediction results more accurate, we changed the independent variables from one-dimensional to 26-dimensional and performed a multiple linear regression with the number of occurrences of the 26 letters as variables to make the regression function fit better.

This, in turn, makes the prediction results more accurate. For prediction, the percentage distribution of the results for the word is obtained by simply substituting the eigenvalues of the word into each function. The average percentage error obtained was 0.45.

Reference

[1] Hu Z-X, Nie L-F. A study of reaction-diffusion infectious disease model with horizontal and environmental transmission[J]. Journal of Mathematical Physics,2022,42(06):1849-1860.

[2] Li W, Chen JW, Liu RX, et al. Tensor time series prediction T-Transformer model[J/OL]. Computer Engineering and Applications:1-7 [2023-04-07]. http://kns.cnki.net/kcms/detail/11.2127.TP.20230220.1138.012.html.

[3] Xu ZJ, Liu DWS. Time series prediction model for the impact analysis of new crown pneumonia epidemic on inpatient business[J]. China Health Statistics,2022,39(03):435-437.

[4] Xu Xu. Study on the activity characteristics and habitat suitability distribution of Oriental White Stork (Ciconia boyciana) [D]. University of Chinese Academy of Sciences (Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences), 2021. doi:10.27536/d.cnki.gccdy.2021.000035.

[5] Yang Li. Study on the spatial pattern and interaction measurement of urban population mobility network in China based on multimodal migration data [D]. Nanjing Normal University, 2020. DOI:10.27245/d.cnki.gnjsu.2020.000665.

[6] Yu, Deng-Hui. Research on the layout of logistics park based on multi-objective migratory bird migration algorithm[D]. Supervisor: Zhang S. Q. Shenyang University of Architecture, 2022.

[7] Wang Dandan. Research on new coronavirus propagation model and critical period control strategy based on SEIR [D]. Nanjing University, 2021. DOI:10.27235/d.cnki.gnjiu.2021.001430.

[8] November 1984Journal of Theoretical Biology 110(4):665-79 DOI:10.1016/S0022- 5193(84)80150-2 SourcePubMed

[9] HONG Bin,CHEN Jinxiu,WANG Liansheng*,YU Rongshan.Analysis and prediction of the spread trend of COVID-19 based on SEIR-LSTM mixed model[J].Journal of Xiamen University(Natural Science),2020,59(06):1034-1040.doi:10.6043/j.issn.0438-0479.202003040

[10] Mička, Pavel. "Letter frequency (English)". Algoritmy.net. Archived from the original on 4 March 2021. Retrieved 14 June 2022. Source is Leland, Robert. Cryptological mathematics. [s.l.] : The Mathematical Association of America, 2000. 199 p. ISBN 0-88385-719-7”

[11] Liu JY, Du MJ, Wang YX. Research on the prediction of subway vehicle wheelset wear based on multi-factor correlation analysis[J]. Mechatronics Product Development and Innovation,2022,35(06):76-79+87.

[12] Selectman Gao. Multi-objective regression combining target-specific features and target correlation [D]. Chongqing University of Posts and Telecommunications, 2020. DOI:10.27675/d.cnki.gcydx.2020.000492.

[13] WANG Jie, CHENG Xuexin, PENG Jinzhu. A Weighted Random Forest Model Based on Particle Swarm Optimization. Journal of Zhengzhou University (Natural Science Edition), 2018, 50(1): 72-76.