

Research on the development strategy of Worlde based on ARIMA time series and mean expectation model

Zehao Li*, Nianxu Lin, Ruibin Zou, Zhenwen Liang

School of Industrial Automation, Beijing Institute of Technology (Zhuhai), Guangdong, China, 519001

*Corresponding author: zehao_li@139.com

Abstract. This paper mainly studies the important role of the percentage prediction of the number of users and the number of word-filling attempts of Wordle game in the future development strategy. The purpose of this article is to research and develop two mathematical models of Wordle game, one of which is the prediction model of user volume and the other is the percentage prediction model of the number of word-filling attempts. The fitting effect is very good, and the result is $R^2=0.984$. By quantifying specific letters at different positions in a word, it is found that the proportion of difficult modes has a direct impact on specific positions and specific letters. Secondly, for the second model, the proportion of each corresponding variable is obtained through comprehensive evaluation and analysis by using the mean expectation model for the percentage of the predicted number of attempts. The average variance formula is used by calculating the average of the relevant percentage of each characteristic value under the corresponding number of attempts. The relevant percentage distribution results of "EERIE" (1, 2, 3, 4, 5, 6, X) are (1.20%, 16.01%, 28.31%, 27.83%, 18.16%, 5.81%). Through the analysis of variance, the P value is less than 0.05, and the prediction result of the model has high confidence.

Keywords: Worlde, User prediction, Percentage of attempts, ARIMA time series model, Average expectation model.

1. Introduction

Wordle is a free ad-free online crossword puzzle. The game is simple: players must have six chances to guess a five-letter word. The correct word is displayed on the green bottom, the yellow bottom indicates that the word contains the letter in the wrong position, and the gray bottom indicates that the word does not contain the letter. After guessing, you can share the customs clearance time and quantity of the guessed words to the social platform, and compete with your friends for the speed of guessing words. This artificial scarcity increases players' desire for challenges and expectations, and increases their interest in learning English [1]. This research makes Wordle's innovative thinking and word-filling design more and more popular. It can improve the update direction of the game by intuitively predicting the number of players and the percentage of word-filling attempts in the next few quarters.

2. Material department and method

2.1. Data acquisition and preprocessing

The data is obtained from the data table in question C of the 2023 American University Mathematical Modeling Contest [2]. For model 1, the date is first preprocessed to predict the number of report results, that is, from the first number to the last number, there are more than 300 days in total, and then the 60-day forecast is added to reach the data of March 1, 2023. The data table provided by Meisai in 2023 will be analyzed to study whether the part of speech of a word has an impact on the problem. The root of a word is the main standard to reflect a word. The total number of non-repeating words is 359, most of which are composed of 5 letters. Clean up and process the acquired data to correspond to the word root table. Then, the given word attribute is noun (NN), the total number is 212; The attribute of the given word is adjective (JJ), 76 in total; The attribute of a given

word is adverb (RB), with a total of 11; Other types of parts of speech belong to the same category, with a total of 60.

For model 2, we have made statistics on the percentage of each test times in the beauty contest, and classified those percentages corresponding to the number of repetition of letters in the part of speech, the part of speech, and whether the end of letters is the relevant characteristic value of e.

2.2. Method introduction

The purpose of this article is to research and develop two mathematical models of Wordle game, one of which is the prediction model of user volume and the other is the percentage prediction model of the number of word-filling attempts. First of all, for model 1, aiming at predicting the number of players and users on March 1, 2023, the time series ARIMA model is used to determine the reasonable model parameters based on the known user data, and the time series prediction chart of each situation is analyzed. The number of players decreases from the beginning until it is smooth and stable. The model predicts the number of users in the range of (19167~18903). The fitting effect is very good, and the result is $R^2=0.984$. By quantifying specific letters at different positions in a word, it is found that the proportion of difficult modes has a direct impact on specific positions and specific letters. Secondly, for the second model, the proportion of each corresponding variable is obtained through comprehensive evaluation and analysis by using the mean expectation model for the percentage of the predicted number of attempts. The average variance formula is used by calculating the average of the relevant percentage of each characteristic value under the corresponding number of attempts. The relevant percentage distribution results of "EERIE" (1, 2, 3, 4, 5, 6, X) are (1.20%, 16.01%, 28.31%, 27.83%, 18.16%, 5.81%). Through the analysis of variance, the P value is less than 0.05, and the prediction result of the model has high confidence [3]. Our work is shown in Figure 1 below:

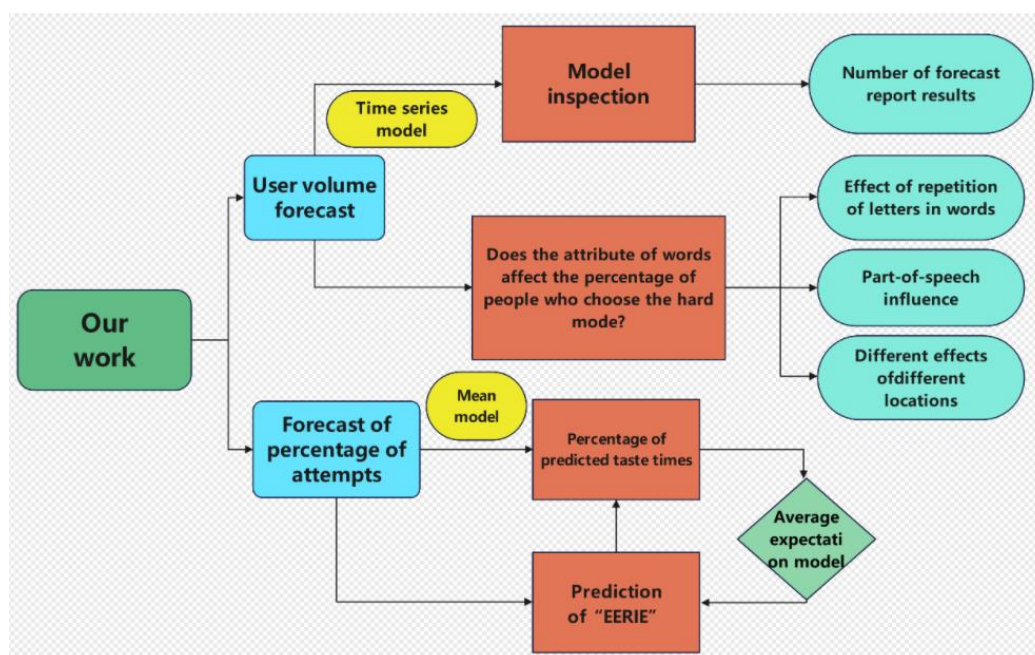


Figure 1. Workflow

2.2.1 Result analysis of ARIMA model

ARIMA model [4] is applicable to the prediction results in this paper. We can get the time series diagram we want only by the variables of date and user volume, without any other external variables. By drawing a chart of the number of difficult players and their time, without considering all players together.

2.2.2 Result analysis of average expectation model

The average expectation model [5] can predict the percentage of attempts with good results. It can reflect problems in real life and produce general results.

3. Model establishment and solution

3.1. The impact of ARIMA model prediction results and word attributes on the proportion of difficulty model

3.1.1 Establishment of ARIMA model

We mainly study the number of results reported. Because we have carried out first-order difference [6] on the data and analyzed the whole difference, the results show that the error is large. The difference result after deviation is relatively stable, so we think it is better to use time series in this model. The difference diagram is shown in Figure 2:

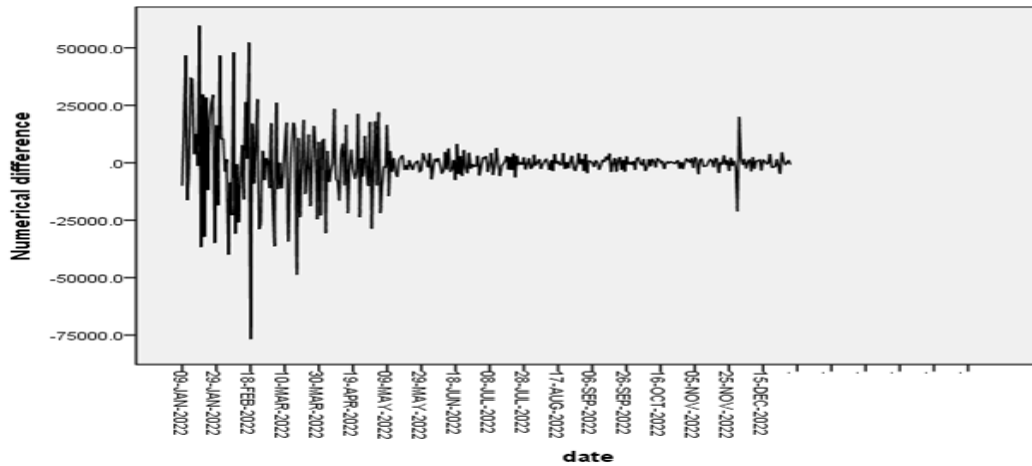


Figure 2. Difference Diagram

Again, in the regression model, we first determine the order P as 60, and consider the data of one year to predict the data of March 1, 2023. We assume that y_t represents the data on March 1, u represents the data on January 1, 2022, r_i represents the autocorrelation coefficient, E_t represents the error, and y_{t-i} represents the value of the previous day.

Therefore, we establish an autoregressive model of p_g , which should conform to the following formula.

$$y_t = \mu + \sum_{i=1}^p r_i y_{t-i} + \varepsilon_t \quad (1)$$

In order to effectively eliminate the random fluctuations in the prediction, we established the moving average model ma , where θ_i is the moving average coefficient, ε_{t-i} is the error of the previous day, and represents the q -order autoregressive coefficient, which should meet the following equation, y_t is equal to

$$y_t = \mu + \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \varepsilon_t \quad (2)$$

Combining the autoregressive model with the moving average model, the formula of the autoregressive moving average model ARMA [2] is obtained, which is rewritten as:

$$y_t = \mu + \varepsilon_t + \sum_{i=1}^p r_i y_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i} \quad (3)$$

In addition, when calculating the correlation coefficients r_i and c_i , we derive the autocorrelation function ACF , which can obtain the following formula

$$A(FCK) = \rho_k = \frac{\text{cov}(y_t, y_{t-1})}{\text{var}(y_t)} \quad (4)$$

When calculating the correlation coefficient p_k of lag k for a stationary ARIMA model, the result is not a simple correlation between y_t and y_{t-k} . y_t is also affected by the intermediate $k-1$ random variables y_{t-1}, y_{t-2} . This $k-1$ random variable has a correlation with y_{t-k} , and the correlation coefficient PK can actually be doped with other variables to eliminate the influence of this degree, We introduce the partial correlation coefficient $PACF$ to enhance the correlation between these two variables.

3.1.2 ARIMA model solutions

A prediction curve is obtained by solving the time series. The first part of the data set is increasing rapidly, while the latter part is decreasing gradually and the trend is stable. Without considering other special external circumstances. The R^2 of our time series result is 0.984, which is quite accurate. We get the prediction curve for the next 60 days from the smooth curve behind.

3.1.3 Index parameters of evaluation model

The results of the report show that the number of people is decreasing, which is consistent with the actual situation. We decisively use time series to predict the model and obtain a prediction interval. The first step of the model is to import data into SPSS, and then analyze the data with time series. In order to verify our model, we also use quadratic parabola to fit the data set, R^2 is 0.984. Then we also used grey prediction, and the prediction fitting effect is also average, with a relative error of 15%

3.1.4 Analysis model results

The main reason is that y_{t-i} direction y_t increases continuously in the previous curve interval, so the curve in the first part increases rapidly. The main reason is that players are full of curiosity about a new game. Gradually, some players no longer want to play this game. Then some y_{t-i} began to decrease y_{t-i} to y_t in the previous period, and then the whole interval began to decrease gradually. Finally, when the number of people increases within the interval of $y_t, -i$ equals the decrease of the number of people for y_t , and the curve reaches the highest value. Then, some old players are not willing to give up the game so soon and join some new players, so the number of decreases gradually decreases, the spread of the interval also gradually decreases, and finally the interval starts to decline steadily, either from the previous difference chart or from the slanted prediction curve.

Finally, we find that the fitting effect of time series is the best. We use the time series model to predict the results of the report, and the prediction range we get is 19167~18903. As shown in Figure 3:

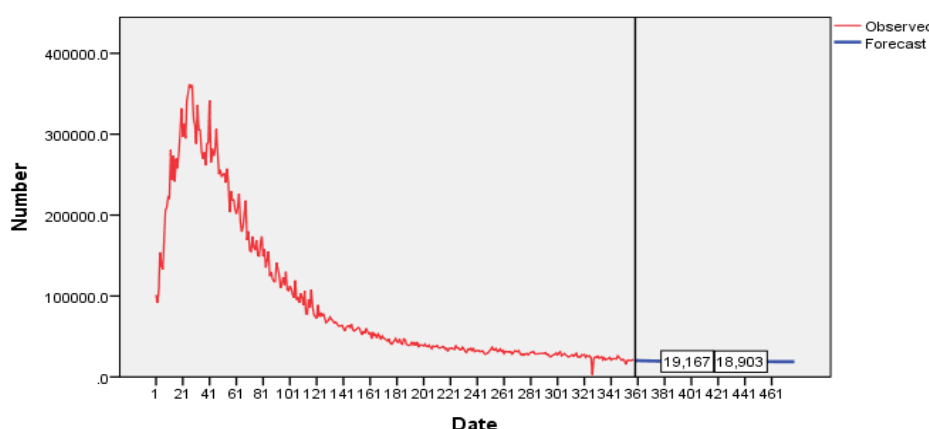


Figure 3. Player Forecast

3.1.5 Further analyze the proportion of the number of repeated letters in a word that affect the difficulty mode

Take the number of letters in a word as a feature, that is, study this feature to study the properties of the word. Therefore, we conduct relevant analysis through SPSS, and then establish the model.

First of all, assign the numerical code "0, 2, 3" to the letters with repeated data as their labels for model analysis. In each word with five bytes, give a definition of the number of letters that repeat.

Define the number of letters that do not repeat as the code "0", the number of letters that repeat twice as the code "2", and the number of letters that repeat three times as the code "3".

Secondly, the number of repeated letters in each time period that has been arranged is sorted, and then the time line of the whole process is used to draw a scatter chart of the proportion of the number of repeated letters and the number of difficult patterns in each word. As shown in Figure 4, there is almost no difference in the proportion of users who are difficult to select by observing the overall trend and analyzing. On the whole, the distribution trend of this curve is increasing.

For words that repeat two letters, it does not prevent users from continuing to play the difficult mode, and the proportion of the number of people is still increasing. Most players are novices at the beginning of the game, so most players are unlikely to play the direct difficulty mode, the specific proportion is 2% or 3%. As Wordle becomes more and more popular, they learn more about the difficult mode and choose more of it.

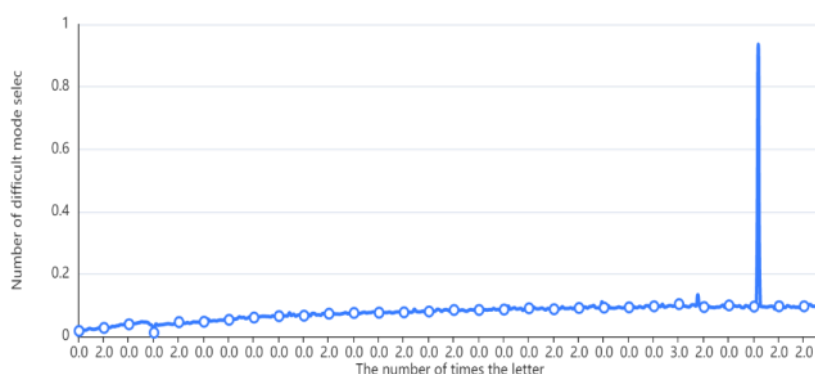


Figure 4. Proportion of difficult members

Finally, the probability of those players choosing the difficult mode is only about 10%. The number of letters repeated in a word is a characteristic attribute of a word. However, the number of repetition of letters in words did not affect the proportion of difficult modes. Therefore, what affects the proportion of players is their familiarity with Wordle.

3.1.5.1 Part of speech will affect the proportion of difficult modes

Discuss the types of letter parts of speech [7] in these five words. We use the numbers "1", "2", "3" and "4" to classify the characteristic assigned nouns, adjectives, adverbs and other parts of speech. After processing the data, the noun is defined as "1", the adjective is defined as "2", the adverb is defined as "3", and other parts of speech are defined as "4". Then we classify the data into one category and draw corresponding graphs.

As can be seen from Figure 5, from nouns to verbs, from verbs to adjectives, this is a growing trend, because more and more players are familiar with this game, and many people are already playing this game. The conclusion is that another characteristic attribute of a word, that is, the part of speech of the word, does not affect the proportion of people who choose the difficult mode.



Figure 5. Scatter plots of people in difficulty modes under four parts of speech

Then we found that the data distribution is between 0.05 and 0.5 and 0.9, so we can determine that the difficulty pattern of these words is related to the part of speech, and the data distribution has almost no difference.

3.1.5.2 Solve the role and influence of different letters in different parts of the word in difficult mode

In word analysis, each position of letters is classified and discussed separately. Considering the difficulty of the rare words with the last letter, such as X, J and Z, which are not common at the end, the existence distribution of the first four letters is different from that of the fifth letter. Considering whether any attribute of these words will affect the percentage of reported scores in the difficult mode, we can see from the analysis of this figure that the frequency density is basically around 0.2, and the distribution is relatively uniform between 0.0 and 0.2.

Through these analyses, it is easy to find that the letter repetition degree of words and the part of speech have little influence on the proportion of people who choose difficult patterns, while the position of letters in some words has an influence on the proportion of people who choose difficult patterns. We need to reduce the special position of letters in these words to increase people's interest.

3.2. Correlation percentage prediction model

3.2.1 Establishment of average expectation model

The percentage of the data set passing the date and the number of attempts, and the percentage of most of the data for each customs clearance has not changed much, and their fluctuation curve is very stable. This percentage can be predicted by investigating the number of repeated letters in the word, the part of speech of the word in the word, and whether the position at the end of the word is "E". As the change of data tends to be stable, the average prediction model is established. Data analysis is shown in Figures 6 and 7:

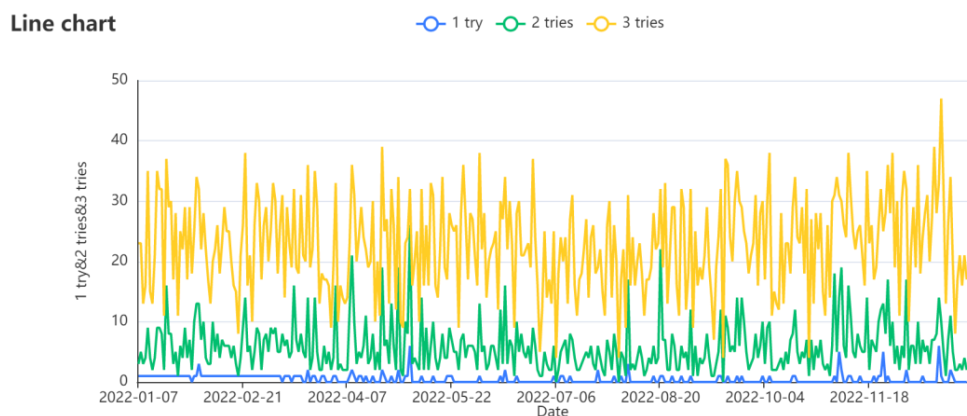


Figure 6. Percentage change of one to three attempts

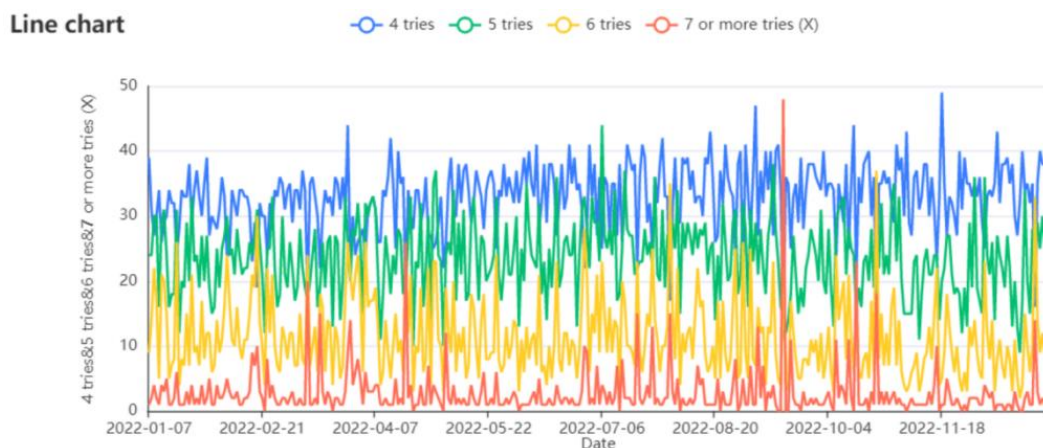


Figure 7. Percentage change of four to x attempts

Define X_1 as the number of times a letter in a word is repeated, the number of times it is not repeated is 0, the number of times it is repeated twice is 2, and the number of times it is repeated three times is 3. X_2 is the classification of part of speech in words. We divide them into four categories, nouns, adjectives, formal adverbs and other parts of speech. Assign the values "1", "2", "3" and "4" respectively. Finally, X_3 is whether the letter at the end of the word is e. According to the second result of model 1, whether there is no "E" at the end of the word [8] has a great impact on the percentage.

3.2.2 Solution of average expectation model

The average prediction model is shown as follows:

$$E(X_i) = Q(x_{1i})P_{x_{1i}} + Q(x_{2i})P_{x_{2i}} + Q(x_{3i})P_{x_{3i}} \quad (5)$$

We have carried out an average statistical survey on how many letters and percentages in X_1 are repeated. The statistical results are as follows Table 1:

Table 1. X_1 is the percent of different degrees

	The percentage of attempts						
X1	tries 1	tries 2	tries 3	tries 4	tries 5	tries 6	x
0	1.558824	6.665441	24.47426	33.08456	22.33456	16.40074	2.415441
2	0.202381	3.309524	17.55422	33.5	30.16071	14.6747	3.869048
3	0	0.5	4	19.5	35.5	30	11

We have made statistics on the relevant percentage of each part of speech and X_{2i} , and averaged the percentage corresponding to the number of times of each part of speech. The statistical results are as follows Table 2:

Table 2. X_2 is the percent of different degrees

	The percentage of attempts						
X2	tries 1	tries 2	tries 3	tries 4	tries 5	tries 6	x
1	0.447619	5.828571	22.64286	32.89048	23.81905	11.72857	2.71564
2	3.202532	5.873418	22.82278	33.05128	23.46154	11.28571	2.833333
3	5.909091	5.909091	22.96491	31.45455	21.72727	11.54545	7.090909
4	0.525424	5.745763	22.86441	33.13559	17.45833	11.50847	2.355932

We have calculated the relative percentage of the end position of X_{3i} and obtained the average value of the corresponding percentage Table 3:

Table 3. X_3 is the percent of different degrees

	The percentage of attempts						
X3	tries 1	tries 2	tries 3	tries 4	tries 5	tries 6	x
0	0.402778	5.194444	21.19444	32.20833	24.54167	13.19444	3.569444
1	0.43671	5.530848	22.52948	33.18668	23.94576	11.58537	2.684519

Then a comprehensive evaluation analysis is given [9]. The percentage of corresponding influencing factors is given. Q_{1i} is equals 0.4, Q_{2i} is equals 0.35, and Q_{3i} equals 0.25. Therefore, we have obtained the calculation formula and model of the final variable:

$$E(X_i) = 0.4P_{x_{1i}} + 0.35P_{x_{2i}} + 0.25P_{x_{3i}} \quad (6)$$

Finally, we conducted variance test on the model and listed the results of variance test, which well tested our model. Take the result of variance test as an example. In real life, some percentage changes greatly, but the overall trend is not large, and some data graphs show relatively stable fluctuations. We use data sets, with only more than 300 data. The amount of data may be a little small, leading to some uncertainty in the test.

3.2.3 Evaluation model predicts EERIE related percentage and index parameters

Use this model to predict the word EERIE. This word has three repeated letters, and then E takes three letters. Query the properties of the word and find that it is an adjective, so X_2 takes 2, and then the end position of the word is e, and X_3 needs 1. Check the figures through models and tables, and we can get the calculation results of each percentage: the EERIE (1,2,3,4,5,6, X) distribution results are (1.20%, 16.01%, 28.31%, 27.83%, 18.16%, 5.81%)

Finally, we test the confidence of the model. For our results, the confidence interval is $p < 0.05$, and the effect is very good. The percentage prediction is instructive for game innovation [10].

4. Conclusion

We use the ARIMA model and an average expectation model commonly used in time series analysis to predict the impact of the number of users and the difficulty of words on the model on March 1, 2023, and then provide conventional development strategies for the innovation direction of Wordle. This article is not only a popular science analysis that can promote the development of the game industry, but also a forward-looking application development that can effectively help promote the optimization and innovation of the game industry.

References

- [1] Deng Shilian Cultivation of interests and hobbies in English autonomous reading [C]//Editorial Committee of China Education and Teaching Series. Summary of basic education theory research results (Volume II). Central University for Nationalities Press, 2007: 365-366.
- [2] <http://www.comapmath.com/MCMICM/index.html>
- [3] Wang Yu, Zhao Xiaoyan, Yang Xingli, Li Jihong. Confidence interval of AUC measurement based on K-fold cross-validation beta distribution [J]. Systems Science and Mathematics, 2020,40 (09): 1564-1577
- [4] Zhang Jinyan, Guo Pengjiang. Application of deterministic time series model and ARIMA model [J]. Journal of Xi'an University of Posts and Telecommunications, 2009,14 (03): 128-132. DOI: 10.13682/j.issn.2095-6533.2009.03.005
- [5] Sun Menglong Portfolio optimization under the criterion of maximizing geometric average expected rate of return [D]. Wuhan University, 2018
- [6] Staley James R, Windmeijer Frank, Suderman Matthew, Lyon Matthew S, Davey Smith George, Tilling Kate. A robust mean and variance test with application to high-dimensional phenotypes. [J]. European journal of epidemiology, 2021 (prepublish).
- [7] Xu Xuehua. Research on the classification of the part of speech of Japanese Chinese words [J]. Northeast Asian Foreign Language Research, 2022,10 (03): 74-84. DOI: 10.16838/j.cnki.21-1587/h.2022.03.012
- [8] <https://youxizhinancn.com/5%E4%B8%AA%E5%AD%97%E6%AF%8D%E5%BC%80%E5%A4%B4%E7%BB%93%E5%B0%BE%E7%9A%84%E5%8D%95%E8%AF%8D-wordle-game-help/>
- [9] Dong Hongqing. Application of comprehensive evaluation in statistical analysis [J]. Journal of Xichang Agricultural College, 2000 (02): 82-85.
- [10] Feng Shiqi, Cai Linlu, Jiang Zeshan, Fu Li Innovation and Construction of Network Communication Mode of Ideological and Political Theory in Colleges and Universities under the Background of "Internet plus" -- Creation and Application of Ideological and Political Contest applet "C Globe" [J]. Education Modernization, 2019,6 (34): 210-213. DOI: 10.16541/j.cnki.2095-8420.2019.34.083.