Wordle game prediction study based on ARIMA and grey correlation analysis model

Qiansong Zhang *

School of Electronic and Information Engineerin, Liaoning Technical University, Huludao, China, 125105

* Corresponding author: zqs003514@163.com

Abstract. The puzzle game Wordle is a huge hit and has attracted a lot of users on Twitter. In the era of big data, this study is dedicated to mining meaningful information from the data thus building a predictive model of the game, which can be used to assess its impact on human choices and behavioral outcomes. This experiment hopes that by predicting trends in multiple aspects of player behaviour, it will lead to more interesting and innovative improvements in puzzle games. By building an ARIMA model, it is possible to predict the number of people who will give feedback on the results each day. A grey correlation analysis model was developed to analyse the effect of three different factors on the percentage of people choosing the difficult mode in the game.

Keywords: Wordle, Prediction study, ARIMA, grey correlation analysis.

1. Introduction

Wordle is a crossword puzzle from the New York Times that is widely popular in Europe and the United States. The rules of the game are clear, players need to guess a word composed of five letters, if the use of less than or equal to six chances to spell out the correct word, the challenge is successful, otherwise the challenge fails. And, after each guess, you can know the result of your guess, the game uses yellow, green, gray three colors, respectively, representing the results of guessing three states.

For Wordle, a puzzle game, the level of difficulty can be reflected to some extent by predicting the number of daily feedback results. Hu Xianqin[1] used Hefei real estate price data from 2013-2019 to forecast using ARIMA model. Based on the change trend of construction human resources in Xinjiang from 2005 to 2019, Wang Hui [2] established an ARIMA model to forecast the number of employees in the construction industry from 2020 to 2025. Wu Xiangbin[3] established an ARIMA model for analysing the passenger flow of the city metro, and predicted the passenger flow of the metro in the city, with the error stabilised within 10%. Yang Yang [4] accurately predicts the population of cities in Sichuan Province by using the ARIMA model. Yiyi Zhang [5] established an ARIMA model based on the data of the number of inbound tourists in Shandong Province from 1991 to 2019 to predict the number of tourists, and the error between its predicted value and the real value was within 5%. Based on the population status of Inner Mongolia, Wang Huaizhao [6] established an ARIMA model to predict the population development trend of Inner Mongolia from 2021 to 2025, which is in line with the actual trend. Therefore, this study used the development of an ARIMA model to predict the number of daily feedback results in order to analyse the difficulty of the daily questions.

In order to gain further insight into this game and to analyse the effect of the attributes of the daily result words on the percentage of people who chose the difficult mode, a relevant correlation analysis model can be developed. Wang Li [7] obtained the primary and secondary factors affecting the water stability of steel slag asphalt concrete by grey correlation analysis. Meng Yishuang [8] conducted a grey correlation analysis on the development of exhibition industry in Hunan Province and found that the exhibition industry has the closest relationship with tourism. Jiang Qicheng [9] used grey correlation analysis to obtain that the largest contribution in the fishery structure of Guangdong Province is the fishery primary industry. Honglei Liu [10] used grey correlation analysis to derive the effect of limestone mineral composition on rock strength. Ni Wenqing [11] conducted a grey correlation analysis on the development capacity of provincial green transformation, and obtained that the tertiary industry has the greatest influence on it. This study used grey correlation analysis to

analyse the effect of word attributes on the percentage of the number of people choosing the difficult mode.

In summary, this study used ARIMA model and grey correlation analysis model to study this game respectively, and the results of the analysis can well predict the difficulty of the daily game and derive the effect of word attributes on the percentage of choosing the difficulty mode of the game.

2. Constraction of the model

2.1. The Establishment of ARIMA

Step1: the experiment requires an ADF test to check the smoothness of the time series and to determine the value of the difference order d. Data from mathematical modelling competitions.

Table 1. ADF Checking Table

ADF Checking Table						
Variable	Different order	t	p	Critical values		
				1%	5%	10%
Number of reported results	1	-4.239	0.001	-3.45	-2.87	-2.571

Combined with Table 1, it can be seen that when the difference is of order 1, the significance p-value is less than 0.05. Therefore, this experiment significantly rejects the original hypothesis and establishes that the series is a smooth time series. The experiment also verified that the time series is also smooth when the difference is of order 0, but its prediction accuracy is not as good as when the difference is of order 1. Therefore, the experiment determined that d = 1.

This led to the experimental decision to build an ARIMA (p,1,q) model. The finite difference time series are numbered such that the random variable at moment t is $x_t(t=1,2,3\cdots)$.

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \tag{1}$$

Equation (1) is the expression of the ARMA model. In order to make the prediction more reliable, the ARIMA model with a difference term of d is introduced in this experiment. Its expression is:

$$(1-\sum_{i=1}^{p}\phi_i L^i)(1-L)^d x_t = (1+\sum_{i=1}^{q}\theta_i L^i)\varepsilon_t \tag{2}$$

Where L is the causal operator. It is a positive integer. Where $\phi_1,\phi_2\cdots\phi_p$ is the parameter of the autoregressive model, $\theta_1,\theta_2\cdots\theta_q$ is the parameter of the moving average model, $\varepsilon_t,\varepsilon_{t-1}\cdots\varepsilon_{t-q}$ is the noise sequence. And the noise sequence obeys a normal distribution with mean 0 and variance σ^2.

The expression of the difference operator is given by:

$$\Delta^d x_t = (1-L)^d x_t \tag{3}$$

Let y_t be equal to

$$y_t = \Delta^d (1-L)^d x_t \tag{4}$$

This results in a deformed equation for the ARIMA model.

$$y_t = \phi_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \tag{5}$$

It should be noted that the ARIMA model has several limitations as follows.

$$\phi_p \neq 0,\ \theta_q \neq 0 \tag{6}$$

$$E(\varepsilon_t) = 0,\ Var(\varepsilon_t) = \sigma_\varepsilon^2,\ E(\varepsilon_t \varepsilon_s) = 0,\ s \neq t \tag{7}$$

$$E(\varepsilon_t \varepsilon_s) = 0, \forall s < t \tag{8}$$

Equation (6) ensures that the highest order of the model p, q.

The restriction of Equation (7) is actually random.

Step2: To better characterize the magnitude of the time series correlation, this experiment use ρ_l to denote *l* order autocorrelation coefficient between y_k and y_{k-1} with a lag of *l* periods.

$$\rho_l = \frac{Cov(y_k, y_{k-l})}{Var(y_k)} = \frac{\omega_l}{\omega_0} \tag{9}$$

In the above equation, ω_l denotes the autocovariance and ω_0 denotes the covariance, and they have the same unit of measure, so the autocorrelation coefficient has the unit of 1. This experiment understand that the autocorrelation coefficient measures the impact of a set of data before and after the time series. And the partial autocorrelation coefficient portrays the effect of a period in the past of a time series on the present.

$$\rho_u = Corr(y_k, y_{t-l} \mid y_{k-1}, y_{k-2}, \cdots, y_{k-l+1}) \tag{10}$$

Using Matlab, this study draw the ACF and PACF. both horizontal axes indicate the number of delays

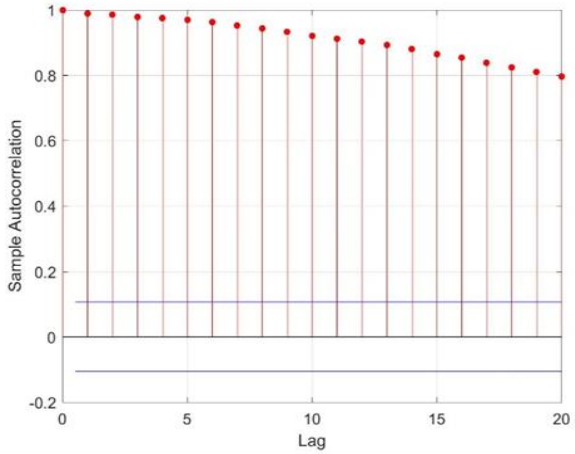

Figure 1. ACF

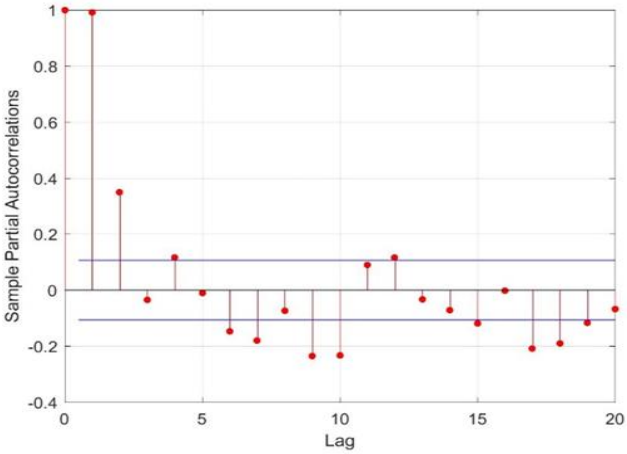

Figure 2. PACF

figures 1 and 2 show the coefficients with upper and lower confidence limits. The autocorrelation mapping is long-tailed at order q and partially autocorrelation mapping is truncated at order p. Combining the images and judgement criteria, with the help of Matlab, the optimal case is found for p = 3, q = 3.

Step3: With the help of SPSS, this study obtained the parameter table of ARIMA (3, 1, 3) model.

Table 2. ARIMA Model (3, 1, 3) Checking Table

Term	Symbol	Value
	Df Residuals	347
Sample size	N	355
Q statistics	Q6 (P value)	0.616 (0.433)
	Q12 (P value)	10.131 (0.119)
	Q18 (P value)	31.973 (0.001)
	Q24 (P value)	56.378 (0.000)
	Q30 (P value)	66.727 (0.000)
Goodness of fit	R^2	0.985

This study concludes from the analysis of the results of the Q statistic in conjunction with Table 2 that based on the test table, the hypothesis that the model residuals are a white noise sequence cannot be rejected based on the fact that the variable: Number of reported outcomes:Q6 is not significant at the level.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum (y_{text}^i - \hat{y}_{pred}^i)^2}{\sum (\bar{y}_{text}^i - y_{text}^i)^2} \tag{11}$$

Where SSE is the algorithmic model prediction error and SST is the baseline model error. Theoretically, we believe that the closer 2r is to 1, the better the model is. In this study, we calculated that the model's goodness of fit 2r is 0.985, the model performance is good, the model basically meets the requirements, therefore, this experiment takes p=3, q=3 can make the model meet the requirements.

This study performe tests of the model. This study use the standard deviation of the model, t-test and so on to analyze the model.

Table 3. Model parameter table

	Coefficients	Standard deviation	t	P>|t|	0.025	0.975
Constants	-170.198	458.36	-0.371	0.71	-1068.567	728.171
ar.L1.D.Number of reported results	0.481	0.132	3.631	0	0.221	0.74
ar.L2.D.Number of reported results	0.234	0.168	1.397	0.162	-0.094	0.563
ar.L3.D.Number of reported results	-0.508	0.083	-6.12	0	-0.671	-0.345
ma.L1.D.Number of reported results	-0.984	0.142	-6.951	0	-1.261	-0.707
ma.L2.D.Number of reported results	0.065	0.24	0.269	0.788	-0.405	0.535
ma.L3.D.Number of reported results	0.544	0.137	3.96	0	0.275	0.813

According to Table 3, the prediction model equations derived from this study are as follows.

$$y(t) = A - B \tag{12}$$

$$A = -170.198 + 0.481 * y(t-1) + 0.234 * y(t-2) - 0.508 * y(t-3) \tag{13}$$

$$B = 0.984 * \varepsilon(t-1) + 0.065 * \varepsilon(t-2) + 0.544 * \varepsilon(t-3) \tag{14}$$

2.2. Attribute Evaluation Model

Gray correlation analysis is a method that integrates quantitative and qualitative well. Therefore, the quantitative analysis results are in good agreement with the qualitative analysis results. Moreover, the results of quantitative analysis can fit well with the actual situation.

The experiment needs to extract the important attributes of words as the influencing factors of whether players play difficult games or not. The following 3 indicators were chosen as important attributes of words. Repetition:Whether there are repeated letters in the word. It is worth noting that letters are repeated two or more times in less than 1% of the words in the data. Therefore, we do not distinguish the number of times a word is repeated. Multiple words: Whether a word has more than one word. Commonality:How common a word is.

Step1: This study use the word attributes that have been processed previously as the signature sequence, and the percentage of the total number of people in the difficulty mode as the parent sequence for analysis. The signature sequence is expressed as:

$$\left[S_1^{'}, S_2^{'} \ldots S_n^{'}\right] = \begin{bmatrix} s_1^{'}(1) & s_2^{'}(1) & \ldots & s_n^{'}(1) \\ s_1^{'}(2) & s_2^{'}(2) & \ldots & s_n^{'}(2) \\ \ldots & \ldots & \ldots & \ldots \\ s_1^{'}(m) & s_2^{'}(m) & \ldots & s_n^{'}(m) \end{bmatrix} \tag{15}$$

The expression of the parent sequence is

$$S_0^{'} = \left(s_0^{'}(1), s_0^{'}(1), \ldots, s_0^{'}(m)\right)^{\mathrm{T}} \tag{16}$$

Step2: Because of the existence of different magnitudes of different series and magnitudes of magnitudes, here this study use the homogenization process to realize the dimensionless data. Thus, this research avoid the occurrence of unreasonable phenomena and reflect the real situation to the greatest extent.

$$f(s(k)) = \frac{s(k)}{\overline{s}} = u(k) \tag{17}$$

The above equation can be homogenized for different sequences. In other words, the elements of the sequence are divided by the average of the sequence elements. After the division, it can be normalized to the vicinity of 1, which is a good basis for model constructing.

In order to obtain the correlation coefficient, this experiment needs to be calculated using the following equation:

$$\Delta_{ik} = \left|s_0(k) - s_i(k)\right| \tag{18}$$

$$\Delta\min = \min_i \min_k \left|s_0(k) - s_i(k)\right| \tag{19}$$

$$\Delta\max = \max_i \max_k \left|s_0(k) - s_i(k)\right| \tag{20}$$

$$\Upsilon(s_0(k), s_i(k)) = \frac{\Delta\min + \rho\Delta\max}{\Delta_{ik} + \rho\Delta\max} \tag{21}$$

Where ρ is the resolution factor, which takes values within (0, 1). The smaller the resolution factor, the greater the difference between the correlation coefficients and the better the discrimination ability. Also, the resolution factor is usually taken as 0.5.

After the above processing, this experiment can obtain the correlation coefficient plot of each signature sequence with the parent sequence as follows.

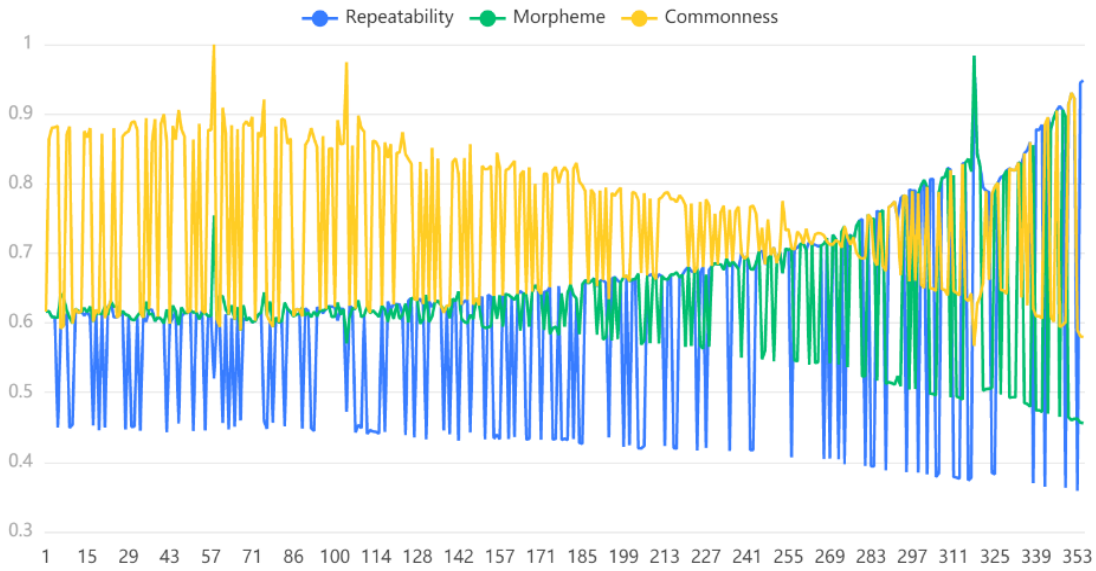

Figure 3. Correlation coefficient graph

The correlation coefficient represents the value of the degree of correlation between the sub-series Repeatability, Morpheme, and Commonness pairs and the corresponding dimension of the parent series. And, the higher value of the correlation coefficient represents the stronger correlation. As

shown in figure 3, the child sequences are correlated with the parent sequences. The yellow colour indicates that the value of "commonality" is more prominent and the correlation is stronger, which indicates that the experimentally established model is very reliable.

This study need to calculate the weighted average of each of its indicators with the corresponding elements of the reference series, using the equation respectively.

$$r_{0i} = \frac{1}{m}\sum_{k=1}^{m} W_k \varsigma_i(k) \tag{22}$$

3. Results

3.1. Forecast of the number of gamers

This experiment need to predict the interval of the number of reported results on March 1, 2023 with a confidence level of 95%. This study use Matlab to draw the prediction graph as follows, with intervals ranging from 1.0577e+04 to 1.1119e+04.

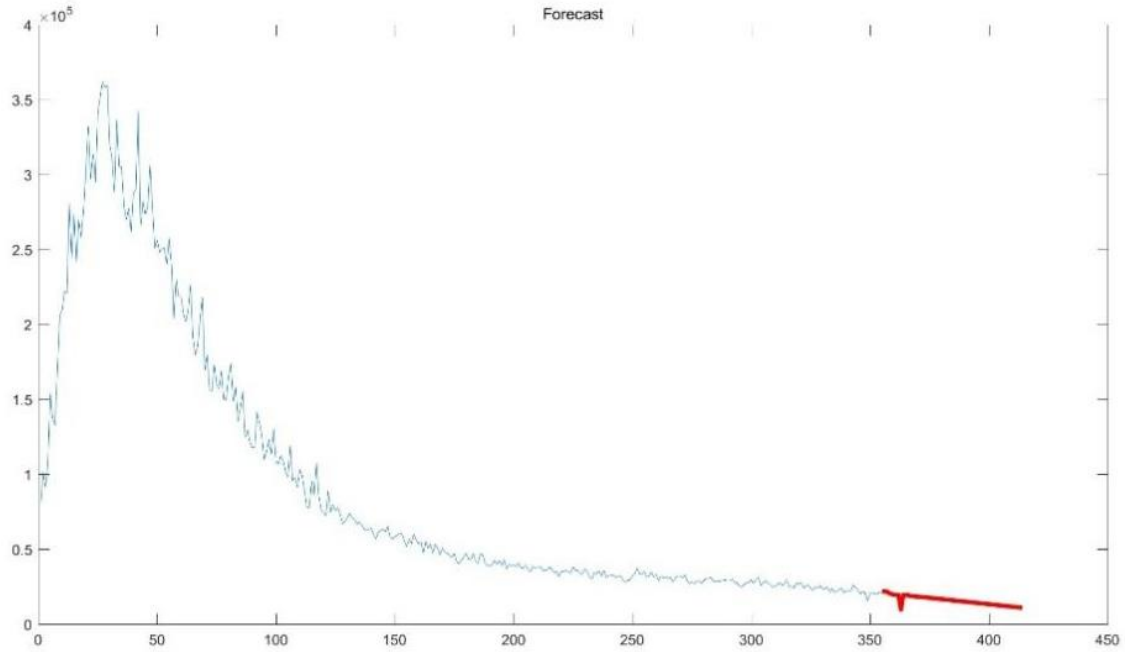

Figure 4. Outcome prediction charts

Combining the model formula with the predicted results, figure 4 shows that the number of people reporting results decreases over time. During the first month or so of World's release, the number of people reporting scores per day continued to rise and "explode", while after about five months, the number of people reporting scores per day began to level off and slowly decline.

3.2. Analysis of influencing factors

The association formula reflects the association between each manipulator object and the parent series, and the results are more precise. In this study, the calculated r_{0i} as the degree of association using the formula given in the model building section. This study obtain the following results from the calculation.

Table 4. Relevance

Evaluation items	correlation degree	Ranking
Commonness	0.745	1
Morpheme	0.633	2
Repeatability	0.613	3

From Table 4, it can be found that the repetitiveness and commonness of words are positively correlated with the proportion of the total number of people choosing the difficult mode. That is, if words are repetitive and common, the proportion of people choosing the difficult pattern increases. The number of word meanings is negatively correlated with this percentage. If a word has more than one lexical meaning, the proportion of people choosing the difficult pattern decreases.

4. Conclusion

In this experiment, an in-depth study of the puzzle game WORLD was conducted and an ARIMA model was developed to predict the number of feedback results per day based on the number of previous feedback results. The ARIMA model is very simple and requires only endogenous variables without the help of other exogenous variables. The results showed a gradual decrease in the number of people participating in the game after the outbreak of this game. In this study, a grey correlation analysis yielded that the number of people choosing the difficulty model was closely related to whether the word contained repeated letters, whether the word had multiple morphemes, and how common the word was. The data used in grey correlation analysis does not need a typical distribution pattern, and the amount of calculation is relatively small, and the results obtained will be more consistent with the results of qualitative analysis. In conclusion, if game officials want the game to attract more players and retain old players, they can try to introduce new mechanisms to the game to enrich the gameplay, and they can adjust the lexical properties of words and whether the words contain repeated letters. The difficulty of the daily game can be controlled by adjusting the word properties, whether the words contain repeated letters or not, and how common the words are, in order to stabilise the number of players.

References

[1] Hu Xianqin. Research on real estate price forecasting based on ARIMA model--Taking Hefei City as an example [J]. China Management Information Technology, 2022, 25 (05): 163 - 166.

[2] Wang Hui. Forecasting human resources demand in Xinjiang construction industry based on ARIMA model [J]. Real Estate World, 2022, (15): 26 - 28.

[3] Wu Xiangbin, Liu Zhifeng, Ding Chenglong et al. Analysis and prediction of metro passenger flow data based on ARIMA model [J]. Defence Manufacturing Technology, 2021, (04): 15 - 17.

[4] Yang Yang, Tian Dingsheng, Zhang Bao'an et al. Research on urban economy and population forecasting based on ARIMA model [J]. Comprehensive Transport, 2023, 45 (11): 79 - 85+97.

[5] Zhang Y. Research on forecasting the number of inbound tourists in Shandong Province based on ARIMA model [J]. Western Tourism, 2023, (07): 10 - 12.

[6] WANG Huaizhao, QIAO Tingting. Forecasting population trend of Inner Mongolia Autonomous Region based on ARIMA model [J]. Inner Mongolia Science and Economy, 2022, (16): 3 - 5+14.

[7] WANG Li, LIU Xianpeng, WEI Huan. Research on water stability of steel slag asphalt concrete based on grey correlation analysis [J]. Building Materials World, 2024, 45 (01): 57 - 60+105.

[8] MENG Yishuang, HUANG Xinyu. Study on the Relationship between Exhibition Industry Development and Service Industry Economic Growth in Hunan Province Based on Grey Correlation Analysis [J]. Foreign Trade and Economic Cooperation, 2024, (01): 19 - 22+111.

[9] JIANG Qicheng, XU Rui, LIU Xiaokun et al. Analysis of fishery industry structure in Guangdong Province based on grey correlation analysis [J]. Aquaculture, 2024, 45 (01): 58 - 62.

[10] LIU Honglei, LI Lingjie, YANG Xingwang et al. grey correlation analysis of the effect of limestone mineral composition on rock strength [J]. Road and Motor Transport, 2023, (06): 84 - 86.

[11] Ni WQ, Tao ZF. Grey correlation analysis of provincial green transition development measurement and industrial structure [J]. Journal of Southwest Forestry University (Social Sciences), 2023, 7 (06): 38 - 45.