ISSN 2616-5805 Vol. 4, Issue 2: 60-65, DOI: 10.25236/AJMS.2023.040209

# A study of the number of Wordle users and experience predictions

**Zhirui Min**

*Hohai University, Nanjing, China, 210098*

***Abstract:*** *Wordle, a popular word guessing game offered daily by The New York Times, has been widely loved and shared due to its straightforward rules and strong fun. This paper uses the* ***ARIMA time series prediction model*** *to predict future user number and then defines the word attribute by combining the word frequency and letter frequency through entropy weight method. To predict the percentage of tries in the future, we fit the percentage of tries with the word attribute from January 7, 2022 to December 31, 2022.This paper forecasts the number of Wordle users on March 1, 2023 and came up with a prediction of 16,458 users. Predicting the word "EERIE" on March 1, 2023 through* ***fitting function*** *and the corresponding percentage of tries is (0,13,35,33,14,2). This paper is instructive for setting the direction of future updates for Wordle as well as giving a forecast method for the future development of Wordle.*

***Keywords:*** *ARIMA, Curve fitting, Wordle, Prediction*

## 1. Introduction

Wordle is a popular puzzle currently offered daily by the New York Times. Players try to solve the puzzle by guessing a five-letter word in six tries or less, receiving feedback with every guess. For this version, each guess must be an actual word in English. Guesses that are not recognized as words by the contest are not allowed. Wordle continues to grow in popularity and versions of the game are now available in over 60 languages. In just a few months, Wordle has grown from a handful of players to millions. Therefore，knowing how fast users are growing and how well they experience is necessary for developers to optimize game modes and improve user experience.In order to better evaluate and predict the number and experience of Wordle users, this paper collects Wordle user experience data during 2022.1.7-2022.12.31[1]. Based on the data,this paper establishes the prediction model , which provides a basis for the formulation of the future development direction of Wordle[2].

## 2. Method

### *2.1 Future user number prediction Model*

#### *2.1.1 ADF stationary test*

Before using the time series to forecast the number of future users, we must first ensure that the number of users accurately reflects all pertinent information about the number of users. Therefore, to avoid the occurrence of false regression, it is vital to assess the validity of the research object before we establish the user number prediction model.

This study use the ADF unit root method in stationary test to assess the validity of the association between user data and time series. The relationship between stationarity and validity is as follows:

The validity of the relationship between user data and time series is not satisfied when the performance of time series is unstable. Otherwise, it is content.

#### *2.1.2 The overview of ARIMA time series prediction model*

ARIMA model uses the inertial change trend of things to establish a prediction model according to the correlation of data in different periods.According to the chronological order, we can obtain a time series by recording the number of Wordle users.Over time, the number of Wordle users has also grown to a certain inertia.In other words,there is a correlation between the before and after values of the series[3].The various information contained in the time series can be extracted by mathematical models, and then using the statistical software to show the relationship image,so that we can apply it to practice.

ISSN 2616-5805 Vol. 4, Issue 2: 60-65, DOI: 10.25236/AJMS.2023.040209

Auto Regressive Integrated Moving Average Model (ARIMA) is one of the methods for time series prediction which can be represented as ARIMA (p,d,q), where AR is "auto regression", MA is "moving average", and I can be understood as difference.

Basic setup steps[4]:①Establish a time series graph of the number of Wordle users, observe whether it is stable and there is a trend of change, if not, using difference method to transform the series into stable time series, d is the number of differences, the value of d is generally 1 or 2 .②The autocorrelation coefficient and partial autocorrelation coefficient are respectively obtained from the processed stationary time series, the optimal stratum p and order q can be known by analysing the pictures of autocorrelation coefficient and partial autocorrelation. ③ Establish the most suitable ARIMA model by using the d, q, p obtained in the first 2 steps, and perform model testing and predictions next. The number of daily Wordle users from 2022.1.7 to 2022.12.31 was used as the basic data for modeling, the ARIMA model is established through SPSS26.0 statistical software to predict the number of users on March 1, 2023.

#### *2.1.3 AIC Check*

In order to find the best model from these multiple models, we utilize the Lagged Information Criterion (AIC), which balances the likelihood of fit with the number of unknown parameters of the model. The calculation formula that results from taking this into account is as follows[5].

$$AIC = -\frac{2L}{n} + \frac{2k}{n} \tag{1}$$

Here, L is the log-likelihood function value, n is the number of observations, and k is the number of unknown parameters.

The value of AIC criterion follows the principle that smaller is better. It can be seen from the formula that the AIC criterion takes into account both the quantity of unknown parameters and the fitting accuracy. Therefore, it is important to consider both the quantity of unknown factors and the precision of the fitting comprehensively to reduce the AIC.

### *2.2 Word attribute*

#### *2.2.1 Define word attribute*

The user's initial guess of a five-letter word is not hinted in the Wordle game, which is very random. The attributes of the word determine whether the user will choose it as the initial input.

The composition of the word and how often it is used affect its attribute. Therefore, this paper uses the English frequency dictionary (the rank of a word in the frequency dictionary) and the frequency of letters in the previous Wordle games to quantify the attributes of words. This paper performs positive normalization on the two during the procedure to make sure that their influence on the attribute is appropriate and in the same direction. The formula is as follows:

For the negative index rank of a word in the frequency dictionary:

$$R' = \frac{R_i - \min(R_i)}{\max(R_i) - \min(R_i)} \tag{2}$$

$$x_{i1} = \frac{R_i}{R'} \tag{3}$$

Here, $R_i$ is the rank of the word in the frequency dictionary, and $X_{i1}$ is the frequency of the word after forward normalization.

For the positive indicator frequency of letters in past Wordle games:

$$L' = \frac{\max(L_i) - R_i}{\max(L_i) - \min(L_i)} \tag{4}$$

$$x_{i2} = \frac{L_i}{L'} \tag{5}$$

ISSN 2616-5805 Vol. 4, Issue 2: 60-65, DOI: 10.25236/AJMS.2023.040209

Here, $L_i$ is the rank of the ith word in the frequency dictionary, and $X_{i2}$ is the frequency of the letter after forward normalization.

At this point, the attributes of a word can be defined as:

$$Word\ attribute\ =\ word\ frequency \oplus letter\ frequency \tag{6}$$

That is, the attributes of words can be divided into two categories: word frequency and letter frequency.

#### *2.2.2 Calculate word attribute*

According to the word frequency and letter frequency defined above, firstly calculate the proportion $p$ between the word frequency and letter frequency of the word and their entropy $e_1, e_2$[6].

$$p_{ij} = \frac{x_{ij}}{\sum_{i}^{n} x_{ij}} (j = 1,2) \tag{7}$$

$$e_j = -\frac{1}{\ln n} \sum_{i=1}^{n} p_{ij} \ln p_{ij} \tag{8}$$

Here $n$ is the number of words, which should be 359 in this paper.

Next, calculate the coefficient of difference between word frequency and letter frequency as $g_i = 1 - e_j (j = 1,2)$. Then calculate the weight coefficient of word frequency versus letter frequency as $\omega_1$, $\omega_2$.

$$w_j = \frac{g_j}{\sum_{j=1}^{4} g_j} (j = 1,2) \tag{9}$$

Finally we can get the attribute value $s_i$ of the word.

$$s_i = \sum_{j=1}^{4} w_j p_{ij} \tag{10}$$

### *2.3 Guess times prediction Model Establishment*

In order to predict the percentage of tries (0,1,2,3,4,5,6,X) that a certain word will be guessed in the hard mode in the future, this paper obtains the functional link between the word attribute and the percentage of tries (0,1,2,3,4,5,6,X) and discovers the trend of the number of guesses changing with the word attribute by polynomial fitting of least squares.

Take the word attribute $(s_i)$ with 4 tries $(c_i)$ as an example, the process of building the fitting multinomial is as follows:

Suppose the fitting function[7] for the word attribute -4 tries is:

$$\varphi^*(s) = a_0\varphi_0(s) + a_1\varphi_1(s) + a_2\varphi_2(s) + \cdots + a_n\varphi_n(s) = \sum_{i=0}^{n} a_i\varphi_i(s) \tag{11}$$

Then the error between it and the true attribute of the word is:

$$\delta(a_0, a_1, \cdots a_n) = \sum_{i=1}^{m} \left[ \left( a_0\varphi_0(s_i) + a_1\varphi_1(s_i) + \cdots a_n\varphi_n(s_i) \right) - c_i \right]^2 \tag{12}$$

To reduce the error, take the partial derivative of a and let it be zero, that is, $\frac{\partial \delta}{\partial a_k} = 0$, then we can

ISSN 2616-5805 Vol. 4, Issue 2: 60-65, DOI: 10.25236/AJMS.2023.040209

get:

$$\sum_{i=1}^{m}\varphi_k(s_k)\left[\left(a_0\varphi_0(s_i)+a_1\varphi_1(s_i)+\cdots a_n\varphi_n(s_i)\right)-c_i\right]=0(k=0,1,\cdots n) \tag{13}$$

Introduce notation $(h,g)=\sum_{i=1}^{m}h(x_i)g(x_i)$ for any function $h(x)$, $g(x)$:

$$a_0(\varphi_k-\varphi_0)+a_0(\varphi_k-\varphi_1)+a_0(\varphi_k-\varphi_2)+\cdots+a_0(\varphi_k-\varphi_n)=(\varphi_k-c) \tag{14}$$

Write it in matrix form as:

$$\begin{bmatrix}(\varphi_0,\ \varphi_0) & (\varphi_0,\ \varphi_1) & \cdots & (\varphi_0,\ \varphi_n)\\ (\varphi_1,\ \varphi_0) & (\varphi_1,\ \varphi_1) & \cdots & (\varphi_1,\ \varphi_n)\\ \vdots & \vdots & \vdots & \vdots\\ (\varphi_n,\ \varphi_0) & (\varphi_n,\ \varphi_1) & \cdots & (\varphi_n,\ \varphi_n)\end{bmatrix}\begin{bmatrix}a_0\\ a_1\\ \vdots\\ a_n\end{bmatrix}=\begin{bmatrix}(\varphi_0,\ c)\\ (\varphi_1,\ c)\\ \vdots\\ (\varphi_n,\ c)\end{bmatrix} \tag{15}$$

Call it a normal system of equations. When $\varphi_0(x),\varphi_1(x),\cdots\varphi_n(x)$ is linearly independent, the system of equations has a unique solution. Take $\varphi_0(x)=1,\varphi_1(x)=x,\cdots\varphi_n(x)=x^n$ and the corresponding system of equations is:

$$\begin{bmatrix}m & \sum_{i=1}^{m}x_i & \cdots & \sum_{i=1}^{m}x_i^{\,n}\\ \sum_{i=1}^{m}x_i & \sum_{i=1}^{m}x_i^{\,2} & \cdots & \sum_{i=1}^{m}x_i^{\,n+1}\\ \vdots & \vdots & \cdots & \vdots\\ \sum_{i=1}^{m}x_i^{\,n} & \sum_{i=1}^{m}x_i^{\,n+1} & \cdots & \sum_{i=1}^{m}x_i^{\,2n}\end{bmatrix}\begin{bmatrix}a_0\\ a_1\\ \vdots\\ a_n\end{bmatrix}=\begin{bmatrix}\sum_{i=1}^{m}x_i\\ \sum_{i=1}^{m}x_ic_i\\ \vdots\\ \sum_{i=1}^{m}x_i^{\,n}c_i\end{bmatrix} \tag{16}$$

## 3. Results

### *3.1 The result of ADF*

The ADF test of the time series data by SPSS software and the ADF test[8] after the second-order difference processing of the data show that the Dickey-Fuller value is -4.242, and the corresponding p value is 0.001, which is less than the significance level α=0.01.The difference result is shown in Figure 1.

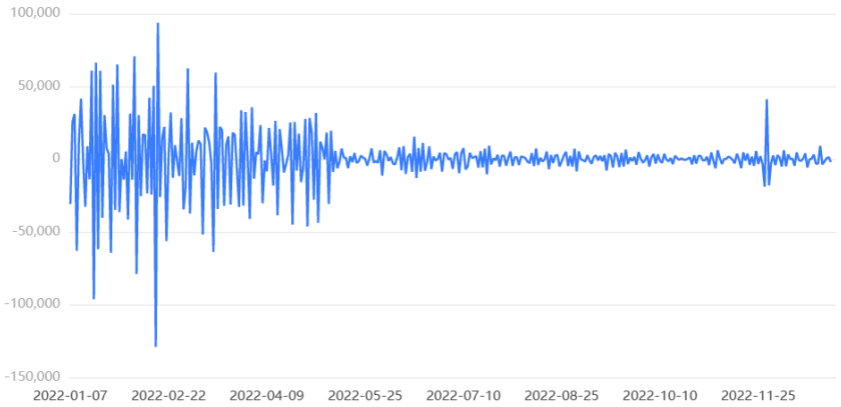


*Figure 1: Differential comparison*

ISSN 2616-5805 Vol. 4, Issue 2: 60-65, DOI: 10.25236/AJMS.2023.040209

### *3.2 ARIMA*

#### *3.2.1 Model identification*

According to the above analysis, non-stationary single sequence data can be processed into stationary time series by 2-order difference, so we choose ARIMA (P, D, Q) model to conduct modeling analysis on the data.

#### *3.2.2 Parameter estimation*

In the process of data preprocessing, we obtain the stationary data after the second-order difference of the data, and identifie the parameter $d = 2$ in ARIMA (P, D, Q).

Next, we plot the Autocorrelation graph and Partial autocorrelation graph of the series after difference[9]which are shown in Figure 2.

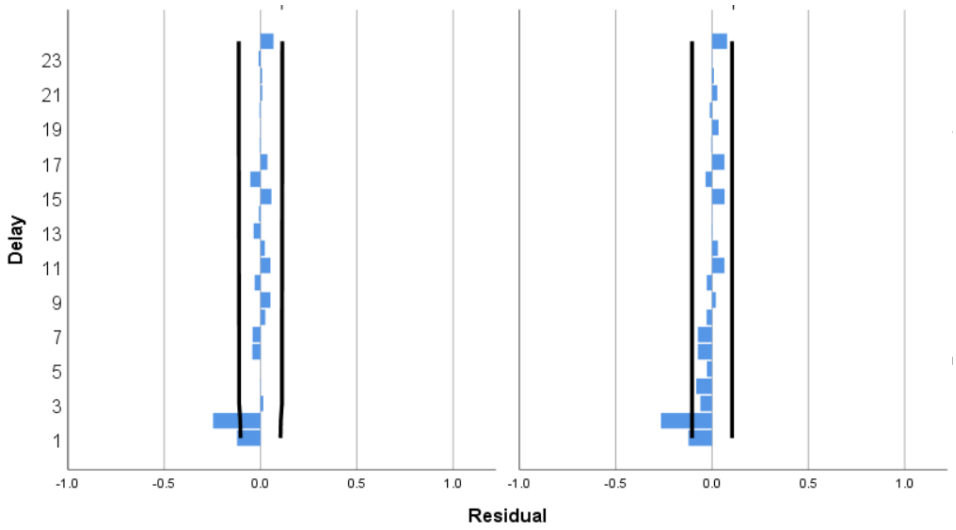


*Figure 2. Autocorrelation graph and Partial autocorrelation graph*

The autocorrelation coefficient of this series exhibits a trailing feature, which is notably non-zero after one lag order, as can be seen by looking at the autocorrelation graph. We can also see from the partial autocorrelation graph that the sample's partial autocorrelation coefficient exhibits censoring characteristics and deviates noticeably from zero at lag order one.

Consider that the values of parameters p in the model are 0 and 1, and the values of q are 0, therefore we can obtain ARIMA(0,2,0) and ARIMA(1,2,0).

#### *3.2.3 Model Check*

We use AIC as the criterion to calculate the AIC values of the above two models, which are AIC (0, 2, 0) =7195.208, AIC (1, 2, 0) =7176.608, respectively. The AIC value of ARIMA (1, 2, 0) is small. Therefore the ARIMA (1, 2, 0) model is used in this study to forecast the number of the users in the future through time series. The analysis chart of the forecast results is shown in figure 3.

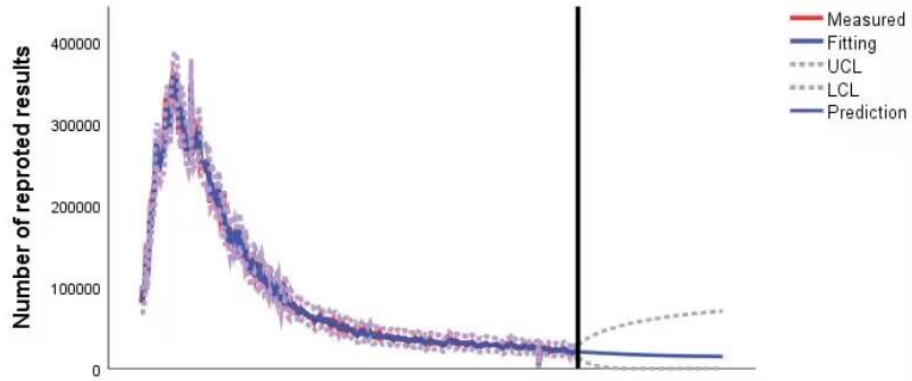


*Figure 3. Forecast analysis chart of future number of users*

ISSN 2616-5805 Vol. 4, Issue 2: 60-65, DOI: 10.25236/AJMS.2023.040209

Taking March 1, 2023 as an example, we use the ARIMA (1,2,0) model established in this paper to predict that the number of users on that day is **16458**.

## 4. Conclusions

The Future user number prediction model is based on ARIMA model with strong mathematical theory support. Moving average algorithm is used to reduce errors caused by data redundancy. We use letter frequency and word frequency as two attributes of a word and define the attribute definition from both the word itself and the characteristics in Wordle, so as to make the attribute definition as complete as possible and match the Wordle game characteristics. The model is heavily focused on real-time feedback from Wordle users to Woedle, flexible combination of word characteristics, suffering ratio and so on. It can also be applied to user usage analysis of other games.

## References

*[1] Boxin Le, Xiaofeng Liu, Na Wang, Weihong Hu, Taicong Feng, Bo Cai. Prediction and analysis of tuberculosis epidemic trend before and after the new crown pneumonia epidemic based on ARIMA [J]. Practical Preventive Medicine,2022,(11):1299-1302.*

*[2] Jing Ma, Mei Wang, Tan Xiaowei, Wenpei Cao, Juansheng Li, Xiaowei Ren, Yuhong Wang, Xiaoning Liu. Application of ARIMA seasonal model in predicting the incidence of hepatitis C in Lanzhou City [J]. China Health Statistics, 2022, (01):98-100+105.*

*[3] Xinmiao Sui. Time series analysis of the relationship between air pollutants and non-accidental death of residents in Hefei City during 2013-2018[D].Hefei: Anhui Medical University, 2021.*

*[4] Huannian Lin, Yizhuang Lu, Huankun Huang, Peng Wei, Sudong Ma. Epidemiological characteristics of mumps and prediction analysis of ARIMA model in Xixiangtang District, Nanning City [J]. Medical Animal Control, 2023, (01):46-51.*

*[5] Xiaofeng Chen. Research on AIC criteria and its application in econometrics [D]. Tianjin: Tianjin University of Finance and Economics. 2012*

*[6] Jinhuang Mao. Improvement and empirical research on the construction method of rural revitalization evaluation index system. Journal of Lanzhou University (Social Science Edition). 2021: 53-64*

*[7] Qiuying Gao, Lili Wang, Rongzhong Wang. Study on curve fitting and optimization algorithm of least square method [J]. Industrial Control Computers. 2021: 103-104*

*[8] Qiaoping Nie, Xiaozheng Zhang. Study on the joint test F statistic in the ADF unit root test [J]. Statistical Research. 2007: 75-82*

*[9] Xiangyue Gu. Establishment and comparative study of outpatient volume prediction model for hand, foot and mouth disease in hospitals[C]. Southwest University, 2022.*