EnPress

About Wordle's difficulty classification model

Hui Huang

Hangzhou Normal University, Hangzhou 311100, China.

Abstract: Wordle, an easy-to-use, fun-filled game. The game is divided into two modes, normal and hard, where players have to try to guess a five-letter word that actually exists in six or fewer attempts, and after each submission, the player is given feedback through the color of the tiles. We examined the game from different perspectives based on the given data for almost a year.

Keywords: Wordle; ARIMA Model; SIS Model

Introduction

For the first question, we first predicted the data in the table based on a time series model, found a good fit, and went on to accept that the model creates a 95 % confidence prediction interval for the number of results reported in the future on March 1, 2023. We reasonably conjecture that the mechanism inherent in the popularity of this game is analogous to the mechanism of transmission of an infectious disease, and add to this prediction based on this. We then discuss the effect of word attributes on the percentage of reported scores played in hard mode in terms of three dimensions: number of vowels, number of repeated letters, and wordiness, giving a plot of the influence factors, which we believe contribute to some degree.

For the second question, in order to predict future dates regarding the number of successful answers required, we first conducted a multi-factor ANOVA on the overall sample, and the results showed that, to some extent, differences in these three factors lead to differences in the distribution of the number of successful answers required. Further, in predicting the word EERIE, we grouped it in one of the categories according to the presented word classification and used the least squares method to predict the distribution of the number of possible successful answers and give the possible results. In fact, we also pointed out the strengths and weaknesses of this model.

1. Model I: ARIMA model for prediction based on the number of reported result

1.1 Data pre-processing

Observing the original data and doing a pure randomness test on it yields $p>0.05$, which shows that its original data is a non-stationary time series and is not suitable for further time series model. Since ARIMA requires the time series to meet the requirements of smoothness and non-white noise, the difference and smoothing methods (rolling average and rolling standard deviation) are used to achieve the smoothness operation of the series. In general, the smoothness of the series can be achieved by performing the first-order difference method on the time series. Therefore, the pure randomness test for the original data is performed with one difference using the LBQ method, and the results are smooth. See the chart below.

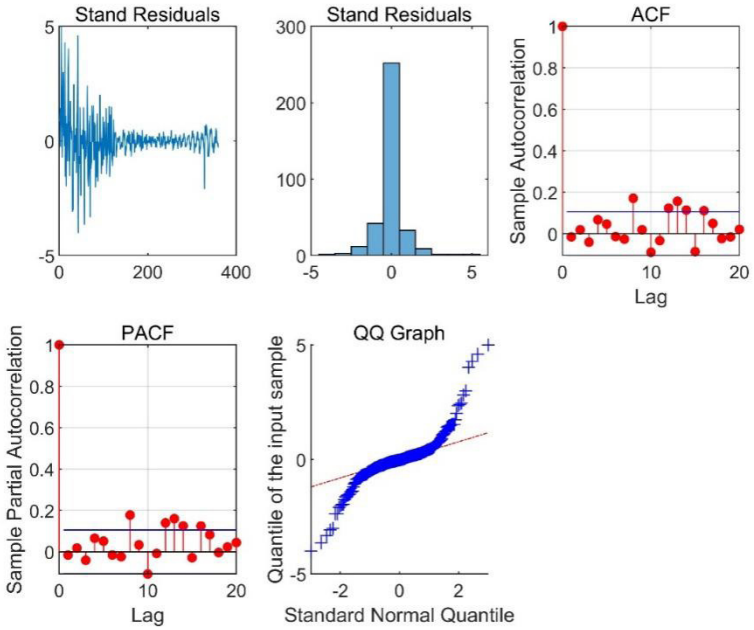

Figure1 : Residual Test

1.2 Construction of ARIMA model for prediction based on the number of reported results

To predict the number of reported results, we use a time series model. The steps are as follows.

The calculation of the general reversible ARIMA model based on the formula shows that it can be written as

$$e_1(\ell) = e_{r+\ell} + \psi_1 e_{r+t-1} + \phi_2 e_{t+-2} + \cdots + \psi_{t-1} e_{r+1}, \ell \geq 1$$

thus

$$E\big(e_t(\ell)\big) = 0, \ell \geq 1$$

and

$$\mathrm{Var}\big(e_t(\ell)\big) = \sigma_t^2 \sum_{j=0}^{\ell-1} \psi_j^2, \ell \geq 1$$

The non-stationary series is similar to the stationary series, but there are some significant differences. The ARIMA (p,1,q) model can be written as a non-stationary ARMA (p+1,q) model,which is written here as

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{r-2} + \varphi_1 Y_{r-3} + \cdots + \varphi_p Y_{t-p} + \varphi_{p+1} Y_{r-p-1} + e_t - \theta_1 e_{t-1} - \theta_2 e_{r-2} - \cdots - \theta_q e_{r-q}$$

Step 1 The time series is ordered, mainly using the BIC information criterion aided by ACF and PACF image tests, and the first-order difference of the preprocessed smooth time series is selected to fix the order. To determine the best lag, several models are fitted with different lag choices. The BIC is calculated for each fitted model with rows corresponding to AR degree p and columns corresponding to MA degree q. We give the relevant formulas for the corresponding BIC and ARIMA models.

$$BIC = k\ln(n) - 2\ln(L)$$

Where is the number of model parameters, n is the number of samples, and l is the likelihood function.

$$\left(1 - \sum_{i=1}^{p} \varphi_i L^i\right)(1-L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right)\varepsilon_t$$

Where l is the lag order, Pis the order of AR model, qis the order of MA model, and D is the difference number.

Step 2 After determining the order, the QQ plots were used to test whether the residuals satisfied the normal distribution; the D-W

test was used to test the autocorrelation of the residuals, and the autocorrelation and partial autocorrelation of the residuals were tested based on the ACF and PACF, with the purpose of testing whether the ideal results do not exceed the position of the blue line in the plots, in other words, whether the test results lie within the 95% confidence interval.

Step 3 The out-of-sample forecasting approach is used, i.e., the Forecast function is used to predict the ARIMA model response or conditional variance, and the confidence level is set to 95% in advance, and the output is predicted backward to the specified order, and based on this, the values are further predicted backward for 100 time points to obtain the prediction intervals and graphs for the reported number of results.

Step 4 A single-step prediction method was used, based on the first 30 data series with a step size of 3. The values of each subsequent point were predicted one by one, and the corresponding graphs were plotted by applying MATLAB to observe the fitting effect of the data.

1.3 Results

For the first question, we consider that at the beginning of Wordle's release, the game's simplicity and fun features led to an initial rise in popularity, which we consider to be consistent with the transmission mechanism of an epidemic, assuming that a person playing Wordle on one occasion would recommend it to the neighboring people, thus inducing the transmission mechanism of an epidemic. After a peak, the prevalence reaches its highest level, and then shows a faster decline, which ushers in a loss of novelty for some people, followed by a decline to a certain level and a leveling off to a certain stability.

Carrying on from the previous question, we finally determined the model of the time series as AMIAR (2, 1, 4) based on the obtained autocorrelation and partial autocorrelation plots. Based on this, we obtained the fit shown below, found a good fit,

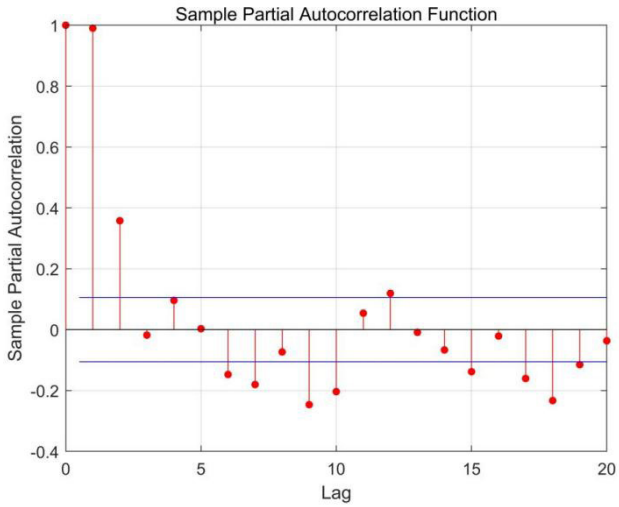

Figure 2: PACF

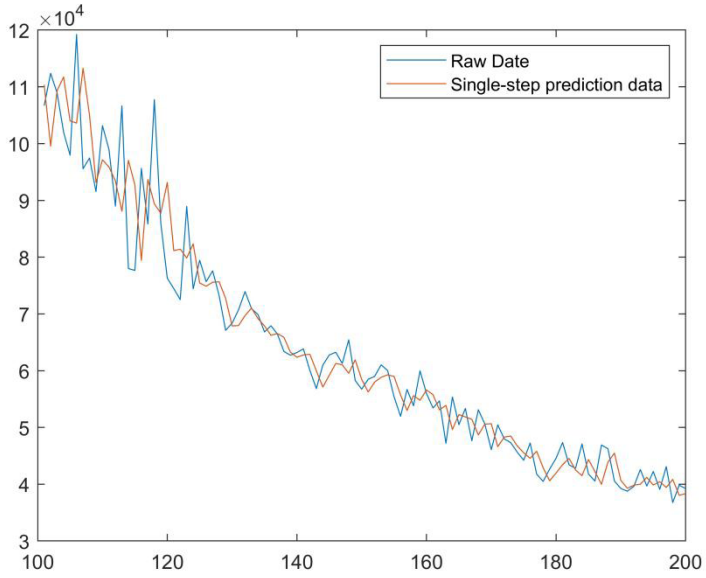

Figure 3: Single-step prediction model results graph

and continued with this method for March 1, 2023. A 95% confidence interval is given. The images are also placed below. The number of forecast report results for 2023/3/1 is 10,648.3.

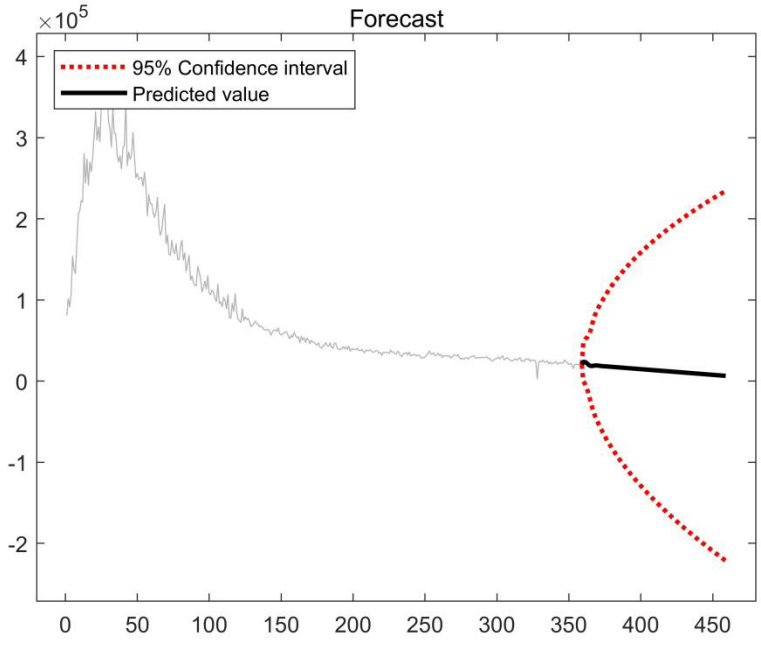

Figure 4: Time series prediction results graph

2. Model II : Word-based attribute factor analysis method

2.1 Construction of analysis model for Word-based attribute

To investigate the percentage effect of different word attributes on the number of reported results in the difficulty mode, we used a factor analysis model The steps are as follows.

Step 1 The possible influencing factors were identified, taking into account the number of vowels, the number of repeated letters and

lexicality, which come from the need for the distribution of vowels and consonants to conform to English grammar, the loss of information to a certain extent caused by repeated letters and the subjective factor of human instinctive guessing brought about by lexicality.

Step 2 Factor analysis was used to analyze the influence degree of the three factors.The number of vowels of all words in the table was counted as 0, 1, and 2 ..... The number of repeated letters was counted as 0, 1, and 2 .....

For lexicality, we specified adjectives as 1, nouns The data were preprocessed and the data were analyzed by using the sigmoid. After preprocessing the data, a factor analysis of these parameters was performed using SPSS to consider their effect on the percentage of reported scores played in hard mode.

2.2 Results

As shown in the table below, the number of repetitions of letters has the greatest impact on the target data, and the more repetitions, the greater the relative difficulty; followed by the number of vowels, the difference between the number of vowels in 2 and 3 is not specially significant, but when the number of vowels is 1, the difficulty of the word is low; and again, the lexical nature.

Table 2: Component Matrix

Name	Ingredients
Word Nature	-0.424
Number of repeated letters	0.634
Number of vowels	0.597

Accordingly, we make a reasonable guess: it is natural that the number of repetitions has the greatest influence on the result, because when a word has too many repetitions, this will lead to confusion for the respondent, because the guesser does not know the number of repetitions of the letter, but will choose to try to get an answer other than the one contained in the first guess, which leads to a significant increase in difficulty; on the other hand, the number of vowels should be consistent with the grammar of English, which would mean that the difficulty of the word is low.

On the other hand, the number of vowels should be in accordance with English grammar, which would mean that the number of vowels is high and the possibility of letter arrangement is reduced, thus leading to a decrease in guessing capacity, which may be one of the reasons for the effect of the number of vowels; for lexicality, the effect of lexicality is not so great, but a large part of the guessers will use nouns as the opening of a game.

References

[1] Rivera Frida; Ahn Kwang Woo; MunozPrice L Silvia, Predicting asymptomatic SARS-CoV-2 infection rates of inpatients: a time series analysis.Infection Control & Hospital Epidemiology, 2021.

[2] Moon JunHo,Kim MinGyu,Hwang HyeWon,Cho Sung Joo,Donatelli Richard E,Lee ShinJae. Evaluation of an individualized facial growth prediction model based on the multivariate partial least squares method.[J]. The Angle orthodontist,2022,92(6).