**Frontiers in Computing and Intelligent Systems**
ISSN: 2832-6024 | Vol. 3, No. 2, 2023

# Analysis and Prediction of Wordle Dataset Based on ARIMA and Pearson's Correlation Coefficient

**Zixin Wan**

School of Computer and Information Engineering, Hubei University, Wuhan 430062, China
3064253289@qq.com

**Abstract:** In this paper, a time series analysis was performed to ensure the parameters p, d, and q. Finally, the ARIMA (1,1,0) model was chosen to create a prediction interval for the number of outcomes to be reported in the future. The results of the precision test showed good confidence in the prediction with a goodness-of-fit of 98.2%. the prediction interval for March 1, 2023 was [10103,10523]. In addition, this paper used Pearson's correlation coefficient and chi-square test to verify the correlation between each attribute of the word and the percentage of scores in the hard model. In summary, this paper found that the attributes of words definitely affect the percentage of scores, and these scores are mainly the initials and lexicality of words.

**Keywords:** ARIMA; Pearson's correlation coefficient; Chi-square test.

## 1. Introduction

In recent years, Wordle, a word-guessing game with a matrix of yellow and green squares as an interactive interface, has become popular on social networks as people become more aware of innovation and prefer smart games [1]. Text analysis refers to the representation of text and the selection of its feature terms; text analysis is a fundamental problem in text mining and information retrieval, which quantifies the feature words extracted from text to represent text information.By examining Wordle dataset, this paper is able to predict the number of reports for Wordle operators and provide assistance with their word selection[2].

In this paper ARIMA is used, a more refined and accurate algorithm for analyzing and forecasting time series data is the Box-Jenkins method, whose common models include: autoregressive model (AR model), sliding average model (MA model), (autoregressive-sliding average hybrid model) ARMA model, and (differentially integrated moving average autoregressive model) ARIMA model. We consider building ARIMA model to make prediction on the Wordle Dataset, taking into account certain errors to obtain the predicted results.

## 2. Acquisition of Data and Assumptions

The data used in this paper are taken from problem C of the 2023 American College Mathematics Modeling Contest. To facilitate the analysis of the problem, the data used in this paper make the following assumptions: (1) The data provided are mostly correct and have negligible impact on the accuracy of the analysis of the results [3]. (2) There are no external factors to motivate users to play the game, and growth and decline in data is common. (3) There are no users sliding scores or making scores or making intentional mistakes.

## 3. Data preprocessing

The Wordle datasets gives the word statistics from January 7, 2022 to December 31, 2022, a total of 358 days. Since we are only allowed to use the data file ‘Problem C Data Wordle.xlsx’ provided by COMAP official, we need to pre-process the data before solving the problem.

To facilitate our later work and to ensure the reliability and plausibility of the result, data pre-processing was conducted on the datasets as follows.

### 3.1. Missing Value Processing

We first checked the whole table for missing values using Python and found that there are no missing values in 'Problem C Data Wordle.xlsx'.

### 3.2. Outlier Detection

In the process of outlier detection, we handle the daily words and data respectively to eliminate the abnormal ones.

#### 3.2.1. Word processing

According to the rules of the game, the word length can only be 5, so we need to remove the data of words whose length is not 5. First of all, we need to detect whether there are spaces and punctuation marks in the words, and this step removes the spaces in the word FAVOR; then the words whose length is not 5 are removed, and this step removes the words CLEN and TASH.

According to the rules of the game, words can only be composed of 26 English letters, so it is necessary to detect foreign words that are not composed of English letters, and this step removes the word naïve.

#### 3.2.2. Numerical processing

(1) Removal of outliers

We plotted the scatter plot based on the column Number of reported results and found that the word STUDY in number 529 was an outlier, which was not consistent with the statistical trend of the number in the last month, so we removed it [4].

(2) Calculating percentages

Considering that the percentages of players solving the puzzle are likely to be rounded and that the attempt limit should be no more than 7, therefore the calculating sum interval should be between 97 and 103, either all discarded or all reserved. For this reason, we find that the percentages of the word nymph sum to 126, which is obviously beyond the error interval [5].

(3) Reversing the date

The dates given in the question are from December 31, 2022 to January 7, 2022. To facilitate the later tasks such as predictions, we arrange the data in positive date order in this pre-processing section.

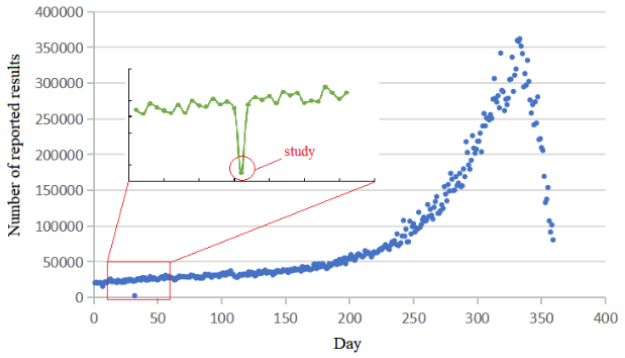



**Figure 1.** The scatter of Number of reported results and day that help pick out the outlier

# 4. The Interval Prediction Model Based on ARIMA

## 4.1. Time Series Analysis

The data to be analyzed in this task changes with time constantly, and our task is to predict the trend of data development over time, so we use a time series model to complete the job.

By means of time series analysis, we can achieve the purpose of describing the past, analyzing the regularity, and predicting the future. Also, we can use this as an access point to predict the number of reports in the future period [6].

## 4.2. Stationarity test

Before using a time series model, we have to test the data for stability to determine whether the data set can be used in a time series model; stability requires that the fitted curve obtained from the sample time series can continue inertially in the future period with the existing pattern.

(1) Time-series test diagram

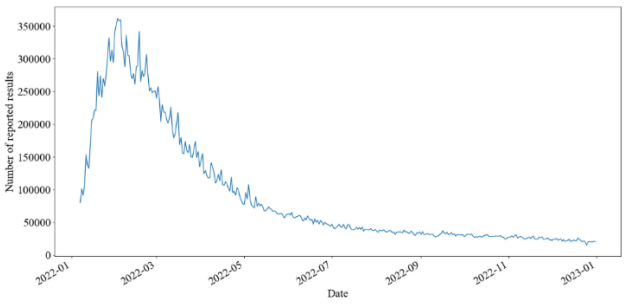



**Figure 2.** Time-series test diagram

From the above time series plot, it can be found that the trend from January to February 2022 shows an upward trend, reaching the highest point on February 2, and from February to December is a gradually stabilizing downward trend, and from July 2022, the curve can be fitted with the Sigmod function after translation.

Therefore, based on the trend of the time series graph, the data set can be predicted using a time series model from a subjective perspective.

(2) Unit-root test

The unit-root test is to test whether there is a unit root in the series and whether the presence of a unit root is a non-stationary time series. The most frequently used one is the ADF test.

Hypothesis of ADF test: We first assume the existence of a unit root and then perform a significance test. If the test statistic P is less than 10% (5% or 1%) confidence level, then we believe that the original hypothesis is rejected with 90% (95% or 99%) confidence, that is, there is no unit root and the series is smooth and stable [7].

**Table 1.** ADF test

| Difference order | P |
|---|---|
| 0 | 0.003*** |
| 1 | 0.001*** |

Note: *** stands for 1% significance level

The results of this serial test show that based on the variable Number of reported results, the significance P-value is 0.003 at the difference of order zero, which is less than 0.01 and presents significance at the level, indicating that the original series is a stationary sequence; at the difference of order one, the significance P-value is 0.001, which is also less than 0.01, indicating that the time series after the first order is also a stationary sequence with better stability[8].

In contrast, we choose the first-order difference, that is, the parameter "d" is 1, and initially determine the ARIMA model as ARIMA (p,1, q).

## 4.3. White noise test

After the stationarity test, we also have to do the white noise test, which is to test whether the series belongs to the pure random series. This is because if the series is completely random, the past behavior will not have any influence on future development, and the subsequent study will be meaningless [9]. According to the results of the stationarity test, the original time series of the Number of resulted reports is stable, so we perform a white noise test on the original series by lagging backward 40, and plot the range of p-value for the white noise test below.

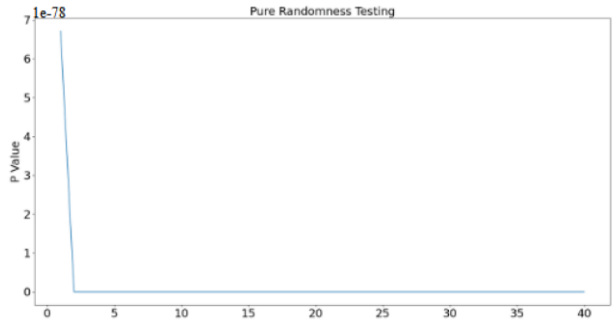


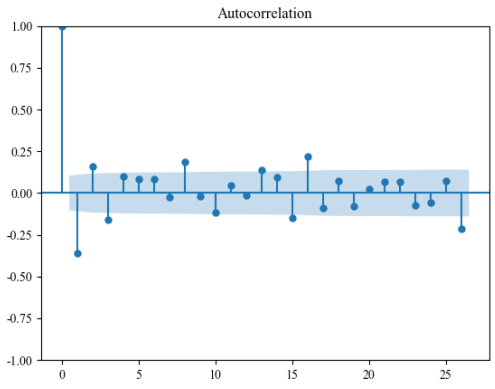

**Figure 3.** Pure Randomness Testing

As can be seen from the figure, the p-values of the randomness tests for the number of resulted reports are all significantly less than 0.05, indicating that the sequence is non-random, which provides a basis for subsequent prediction.

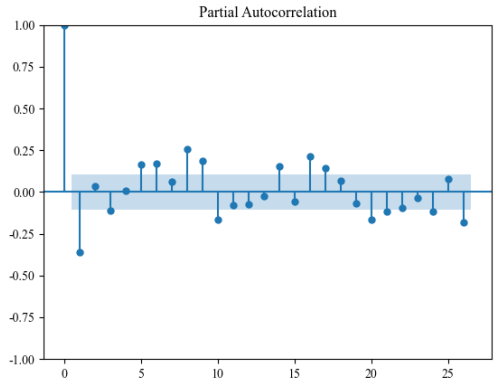
## 4.4. ARIMA Model based on Time Series Prediction

**(1) Selection of p and q**

By means of the unit root test, we determined that the series is a smooth series after the first order difference. Here, we make the autocorrelation and partial autocorrelation plots of this stationary series using python as follows.




(a)The autocorrelation




(b) the partial autocorrelation.

**Figure 4.** ACF and PACF

From Figure 4(a), we can see that the series is truncated at order zero with a preliminary judgment of q=0. From Figure 4(b), the series is truncated at order 1 with a preliminary judgment of p=1.Combining the above, we initially determined the ARIMA model as ARIMA (1,1,0).

(2) Prediction

Based on the model ARIMA (1,1,0), we forecast the number of resulted reports on March 1, 2023, and the fitted forecast graph is as follows.

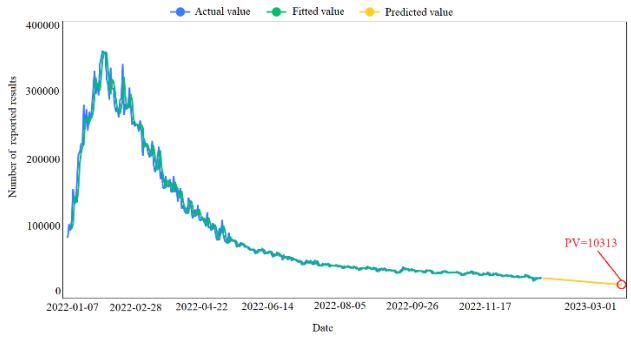



**Figure 5.** The prediction of the number of resulted report base on ARIMA

The above figure represents the original data graph, model fitted values, and model predicted values for the number of resulted report series. From the figure, it can be visually seen that the fitted curve overlaps well with the original data curve, and the changing trend is basically the same, which indicates that the model we used fitted very well. We concluded that the number of reported results on March 1, 2023 predicted by the ARIMA (1,1,0) model was [10103,10523] with a precision of 98.2%.

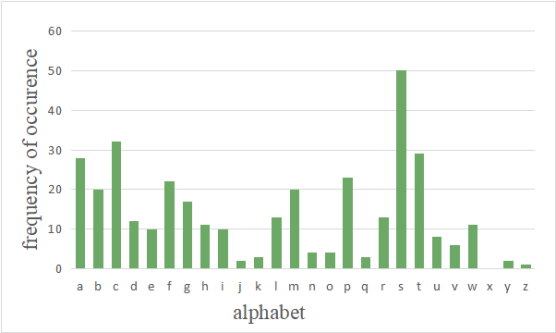
# 5. Attribute Extraction

## 5.1. Parts of Speech

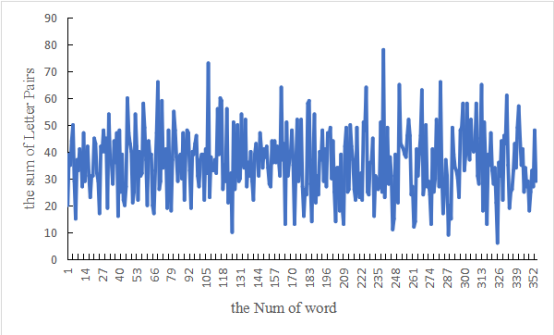
Part of speech is one of the top features of words. Given that this attribute may affect the correctness of users' guesses, here we apply Python's natural language processing NLTK library to get the part of speech of each word. Our result statistics show that the number of nouns is predominant, followed by adjectives, so we assume that the percentage of scores in hard mode with the same property can be correlated [10].

## 5.2. Word Initials

According to the number of words counted by the Oxford English Dictionary, the number of words beginning with different letters possess some characteristics, for example, the number of words beginning with the letters s and p is the largest while letters beginning with x and y has the smallest proportion. In light of this, we presume that the initial letters can have a certain influence on the correct rate of vocabulary guessing in Wordle. The following is the frequency graph of the initial letters of words after pre-processing the data using Python.




(a)The frequency graph of initial letters




(b)The sum of letter pairs

**Figure 6.** The frequency graph of initial letters and the sum of letter pairs

The result that the largest number of words beginning with the letter s and the smallest number of words beginning with the letters x, y, and z were obtained, which were approximately the same as the regular patterns in the Oxford English Dictionary. Therefore, the prediction of future words based on the statistical quantities of the whole has strong reliability to some extent.

## 5.3. Adjacent Letters Within a Word

In general, the composition of each word is a sequence of adjacent letters generated by phonetic symbols, take the word track for instance, pronounced /træk/, can be decomposed into

/tr/ and /k/ two common phonetic symbols, the letter pair of /tr/ is tr, /k/ has ck, ke and other letter pairs. From this, we numbered the word letters 1~5 from left to right, then match them adjacently to form a letter pair and count the total number of occurrences of 6 letter pairs in each word in all words. This is used to indicate the commonness of the word and judge its correlation with the percentage of difficulty pattern scores.

From Figure 6 (b), we found that the sum of letter pairs for each word was unstable, and speculate that this may have some relationship with the percentage of difficulty patterns.

### 5.4. Number of Identical Letters in Each Word

If two or more letters are identical in a word, the total number of guesses required will decrease to some extent. That is, the number of repetitions of letters in a word affects the level of difficulty of the question to a greater or lesser extent. Thus, we extracted the number of repetitions of the letters of all the words in the processed datasets. The result was that there were 254 words with different letters, 98 words with two identical letters, 2 words with three repetitive letters, and no words with four or more identical letters.

### 5.5. The Ratio of Vowel Letters to Consonant Letters

Consonants and vowels depend on each other to form the diverse syllables of words. After reading related articles, we found that people prefer to use words with more vowels, so we calculated the ratio of vowel letters to consonant letters and guessed that the number of vowel letters would be related to the percentage of scores in the difficult mode.

## 6. Correlation Analysis

### 6.1. Method Selection

For this part, we have to determine the relationship between the founded attributes (X) and the score played in hard mode (Y).

The dependent variable Y is quantitative in this question, while the independent variable X is composed of different types when representing different attributes. However, for the two attributes of part of speech and word initials, the relationship between them belongs to quantitative and categorial, so we cannot use the correlation coefficient directly. We expressed the quantitative representation of fractional percentages as multiple interval fixed-class representations and used the chi-square test to determine the relationship between them.

### 6.2. Correlation coefficient

Before choosing the correlation coefficient, we need to determine whether the data satisfy normality, and use Pearson correlation coefficient if it satisfies, or Spearman correlation coefficient if it does not.

The difference between Spearman correlation coefficients (SCC) and Pearson correlation coefficients (PCC) is that SCC takes the correlation coefficient as a ranking variable, so we first need to calculate the correlation coefficient $r$ by the following formula:

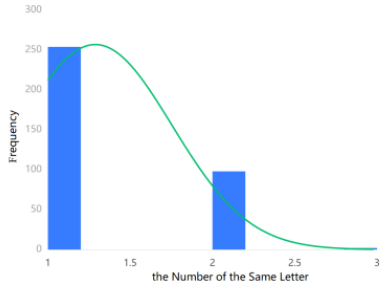
$$r = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \quad (1)$$

Here $X_i$ (i=1,2,...,354) represents the independent variable, meaning the attributes of the word, and $Y_i$ (i=1,2,...,354) denotes the dependent variable, meaning the scores in hard mode, and $X$ and $Y$ denote the mean of the two, respectively.

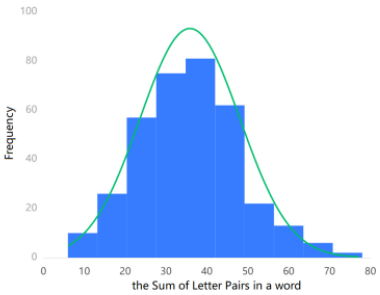
The data satisfying the normal distribution can be judged by calculating the correlation coefficient, and the closer the correlation coefficient is to 1, the stronger the correlation is; for the data not satisfying the correlation coefficient, we have to grade them after calculating the correlation coefficient, and then calculate it by the following formula:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (2)$$

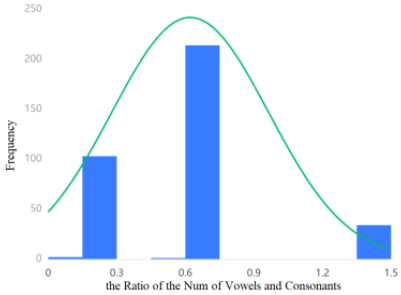
Here d denotes the descending position and n denotes the number of samples. The correlation between them is then determined by calculating out of ρ.




(a)




(b)




(c)

**Figure7.** (a) - (c) plot the frequency of 3 attributes, the number of the same letter, the sum of letter pairs in a word and the ratio of the number of vowels to consonants, respectively.

The number of same letters, the sum of letter pairs and the ratio of the num of vowels to consonants all basically satisfies the normal distribution, so for this attribute, we use the Pearson correlation coefficient, and calculate from the above formula are 0.086, 0.089 and 0.046 respectively, indicating that the correlation between the two isn't apparent.

## 6.3. Chi-square test

Before conducting the chi-square test we first classified the percentages of difficult mode scores (Y) into intervals, then counted the number of different categories of attributes in each interval, and finally conducted the chi-square test. We divided the data set of percentages of scores in hard mode evenly into four intervals as follows.

**Table 2.** Four intervals

| | | | |
|---|---|---|---|
| 0.012,0.063 | 0.063,0.083 | 0.063,0.083 | 0.093,0.133 |

For the two attributes of word initials and part of speech, the reliability of the hypothesis is first calculated by drawing separate frequency lists and assuming that they are correlated, using the following formula:

$$x^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} \tag{3}$$

in which $x$ (i=1,2, ...) denotes the group frequency, which is the true value in the table, $n$ denotes the total number of samples, and $p_i$ denotes the probability of occurrence of each sample. The greater the value of the hypothesis, the greater the reliability of the hypothesis and the correlation. The correlation reliability can be determined according to the following table.

**Table 3.** The precision of chi-sequare test

| $P(\chi^2 \geq k)$ | 0.05 | 0.025 | 0.010 | 0.005 | 0.001 |
|---|---|---|---|---|---|
| k | 3.841 | 5.024 | 6.635 | 7.879 | 10.828 |

(1) Initials words

**Table 4.** The number of different initials in different interval

| interval | a | b | c | ... | sum |
|---|---|---|---|---|---|
| [0.012,0.063) | 5 | 3 | 10 | ... | 87 |
| [0.063,0.083) | 8 | 7 | 7 | ... | 83 |
| [0.083,0.093) | 7 | 5 | 8 | ... | 94 |
| [0.093,0.133] | 8 | 5 | 7 | ... | 89 |
| sum | 28 | 20 | 32 | ... | 354 |

Using the above formula and the table, we calculate the value of x as 4.889 and indicate that correlation probability lies between 0.95 and 0.9725, so it can be determined that the word initials and the score in hard mode are highly correlated.

(2) Part of speech

Similarly, the value of x calculated here is 4.123, and the result is that there's a strong correlation in the attribute of part of speech.

**Table 5.** The number of parts of speech in different interval

| c | NN | JJ | MD | ... | sum |
|---|---|---|---|---|---|
| [0.012,0.063) | 44 | 24 | 2 | ... | 87 |
| [0.063,0.083) | 63 | 15 | 0 | ... | 83 |
| [0.083,0.093) | 45 | 23 | 0 | ... | 94 |
| [0.093,0.133] | 53 | 23 | 0 | ... | 89 |
| sum | 205 | 85 | 2 | ... | 354 |

# 7. Conclusion

In this paper, we perform time series forecasting and correlation analysis. To obtain more accurate results, this paper selects an effective model and continuously optimizes the parameters. Finally, we obtained the ARIMA (1,1,0) model with a fit of 98.2%, which shows that the model has a good predictive power. In addition, five attributes were derived and correlation analysis was performed between attributes and percentages in the hard model. It was found that the attributes of the words definitely affect the percentages of the scores, which are mainly the initials and the lexicality of the words.

# References


[1] Zeng B, Tong M, Ma X. A new-structure grey Verhulst model: development and performance comparison[J]. Applied Mathematical Modelling, 2020, 81: 522-537.

[2] Wang Z, Dang Y, Liu S. Unbiased grey Verhulst model and its application[J]. Systems Engineering-Theory & Practice, 2009, 29(10): 138-144.

[3] Ariyo A A, Adewumi A O, Ayo C K. Stock price prediction using the ARIMA model[C]//2014 UKSim-AMSS 16th international conference on computer modelling and simulation. IEEE, 2014: 106-112.

[4] Wang Z X, Li Q. Modelling the nonlinear relationship between CO2 emissions and economic growth using a PSO algorithm-based grey Verhulst model[J]. Journal of Cleaner Production, 2019, 207: 214-224.

[5] Zeng B, Ma X, Zhou M. A new-structure grey Verhulst model for China's tight gas production forecasting[J]. Applied Soft Computing, 2020, 96: 106600.

[6] Tang L, Lu Y. Study of the grey Verhulst model based on the weighted least square method[J]. Physica A: Statistical Mechanics and its Applications, 2020, 545: 123615.

[7] Hillmer S C, Tiao G C. An ARIMA-model-based approach to seasonal adjustment[J]. Journal of the American Statistical Association, 1982, 77(377): 63-70.

[8] Sowell F. Modeling long-run behavior with the fractional ARIMA model[J]. Journal of monetary economics, 1992, 29(2): 277-302.

[9] Fattah J, Ezzine L, Aman Z, et al. Forecasting of demand using ARIMA model[J]. International Journal of Engineering Business Management, 2018, 10: 1847979018808673.

[10] Nochai R, Nochai T. ARIMA model for forecasting oil palm price[C]//Proceedings of the 2nd IMT-GT Regional Conference on Mathematics, Statistics and applications. 2006: 13-15.