

Prediction and Word Difficulty Classification Model Based on Wordle

Summary

As the wordle crossword game was spread internationally, We aims to analyze the time series characteristics of Wordle and describe the word attributes that appear in the game.

For task 1, our paper presents a prediction model based on the life cycle theory and time series. We propose a **piecewise linear trend extraction algorithm** that identifies the appropriate sliding window sequence by targeting the local minimum RMSE. Using this algorithm, they divide the time series into four slices - the rising , saturation, falling and stable period - with an average RMSE value of 2563. We then use the stationary period series to predict the data for March 1st, 2022, which is estimated to be 19369. However, the problem of incomplete use of time series data remains unsolved. To complement the prediction part, we use an **ARIMA (3,1,12) - GARCH (1,1) model** with a goodness of fit of 0.986. The also forecast the data for March 1st, 2023, which is estimated to be 20982. Finally, we correct the results of the life cycle model using the results from the ARIMA-GARCH model. The final result is 20982, with a 95% confidence interval of [17869, 23354].

To analyze the words in the Wordle game, we select six word attributes and calculate the correlation between the difficulty mode percentage and these word attributes using the **Spearman correlation coefficient**. We find that there is no significant correlation, except for a moderate correlation of -0.144 for word frequency.

For Task 2, our paper uses the **Apriori algorithm** to establish the correlation rule between classification variables and word difficulty. We establish the mixture index (Mi) based on the confidence level of these rules. Then, we use the **partial least squares regression model** to calculate the percentage of different answer times using the aforementioned word attributes. The final predictions are as follows: 1=0%, 2=2%, 3=15%, 4=33%, 5=30%, 6=16%, and X=4%. We identify social media, global epidemics.etc. as the uncertainty factors. The model can reach a minimum RMSE is 0.74, indicating good performance.

For task 3, we first used a **system clustering** to cluster 359 words into three levels of difficulty: simple, moderate, and difficult based on their seven percentage attributes. Then, we used **Fisher linear discriminant analysis** to evaluate the classification results with a 7:3 split between training and testing sets, and found that the clustering accuracy was about 95%, indicating strong classification ability of the model. we substituted the seven percentage data of EERIE words predicted in Task 2 into the discriminant equation, and obtained the result that EERIE belongs to the simple class.

For Task 4, we have observed that as the number of players decreases, the percentage of difficulty mode increases. Additionally, we have found that the distribution of the percentage conforms to a normal distribution.

Finally, sensitivity tests were conducted to verify the robustness of the regression equation coefficients and ARIMA-GARCH model coefficients to ensure the reasonableness of coefficient selection.

Key Words: life cycle theory; piecewise linear trend extraction algorithm; ARIMA-GARCH Model; correlation coefficient; Apriori algorithm; PLSR;

Contents

1	Introduction	2
1.1	Background and problem summary	2
1.2	Ourwork	3
1.3	Assumptions and symbols	3
2	Task1: Report quantity interpretation model	4
2.1	Preprocess Data	4
2.2	Piecewise Linear Regression for Life Cycle shape extraction	4
2.3	Time series supplementary model	6
2.4	Model inspection	9
2.5	Model predictions	9
2.6	Research on the attribute of words	10
3	Task 2:Report result forecast	11
3.1	Classification variable processing based on Apriori algorithm	11
3.2	Establishment of Partial Least Squares Regression Analysis Model	13
3.3	Uncertainties in Models and Forecasts	15
3.4	Evaluate the model prediction results	15
4	Task 3: Word difficulty classification model	15
4.1	Hierarchical clustering	16
4.2	Fisher linear discriminant analysis	17
4.3	Classification accuracy	18
5	Task 4:Analysis of data set characteristics	18
6	Model sensitivity analysis	20
6.1	Tests for the parameters of ARIMA	20
6.2	Test of Regression Equation Coefficients	20
7	Conclusion	21
8	MEMO	23
	Appendices	24

1 Introduction

1.1 Background and problem summary

Wordle, a word guessing game rose to global popularity in the December of 2021[2]. The goal of the game is to guess a five-letter English word within six attempts. Each attempt will provide a prompt to the player through the color change brick, telling the player whether a character is part of the solution and whether it is in the correct position.

In view of the popularity of the game, many people have studied the game, and the medical field has begun to analyze the value of skill development in participating in this daily intelligence activity; At the same time, people are constantly studying how to maximize the game's winning and solve daily challenges. The work by Kandabada [3] suggested manually selected four words, [SPORT, CHEWY, ADMIX, FLUNK], as the best starting words. Sidh[4] searches for the best starting word from the perspective of linguistics. Anderson and Meyer[5] use machine learning to find the best strategy for solving puzzles.

Wordle is divided into regular mode and hard mode. In the hard mode, players are required to use the correct letter in a word (the tile is yellow or green) in subsequent guesses, which greatly reduces the efficiency of the strategy. We need to use the data set *Problem_C_Data_Wordle.xlsx* provided by MCM, which includes the scores reported by users on Twitter from January 7 to December 31, 2022, Conduct multi-angle analysis on the popularity of the game and the difficulty of daily word titles.

After through in-depth analysis and research on the background of the problem, we can specify that our article should cover the following aspects:

- Establish a model to describe the change of report results, and use the model to create a forecast interval for the number of reported 2023 on 1 March.
- This paper explores the influence of the attributes of words on the percentage of players' scores in difficult mode, and discusses the establishment of a regression model to determine the average number of answers of different words, and explores the reasons for the influence of different attributes on players.
- Develop a model to predict the proportion of results players will get on a given solution word at a future date, and explore the prediction of player results for the 2023 word EERIE on March 1, the prediction accuracy of the model is analyzed and evaluated.
- This paper sets up an evaluation model of the word difficulty, classifies the words in the game according to the difficulty, carries on the feature engineering to establish the word classification attribute, and evaluates the word EERIE according to the difficulty, the accuracy of our classification model is also discussed.
- Mining interesting features in a dataset.
- Finally, summarize our results in a one- to two-page letter to the Puzzle Editor of the New York Times.

1.2 Ourwork

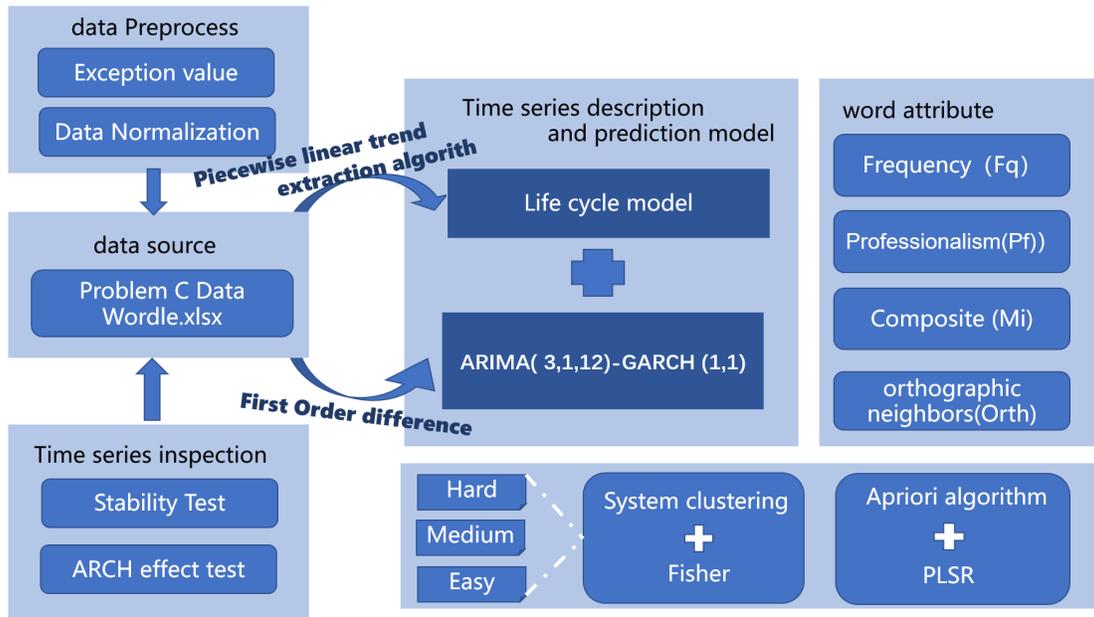


Figure 1: Ourwork

1.3 Assumptions and symbols

- **The player result data is stable to a certain extent:** the data will not fluctuate violently, so if the data on a certain day is too different from the surrounding data, it will be considered as error data.
- **The situation that in which player fails to pass the customs is considered as seven times:** if the player fails to find the answer six times in the game, the system will give the correct answer, and the player will definitely find the answer the seventh time.
- **The word difficulty in the Wordle game is considered to be evenly distributed:** therefore, we can completely rely on this data set to divide the word difficulty.

Symbols	significance
$RmseThreshold$	Time series sliding window error threshold
$X_i(i=1,2,3...359)$	time series
Y_{t-1}	The observations in the t time series
Fm	Word form
Pf	Word Professionalism
Fq	Frequency

2 Task1: Report quantity interpretation model

2.1 Preprocess Data

This paper introduces the life cycle theory to describe the number of changes in the results reported by players in Wordle from January 7, 2022 to December 31, 2022. We use the piecewise linear trend extraction method to extract the shape of the model and then identify that the time series model is divided into four stages: growth, maturity, decline and stability.

- **Exception value handling**

We use Twitter data to correct the abnormal words in the data, such as the number of letters is 4 or 6, and the spline interpolation method is used to process the abnormal value data.

- **Data Smoothing**

First, smoothing is performed to remove weekly seasonality. The overall trend of a series can be better identified if the impact of low-level fluctuations due to weekly seasonality is removed. Thus, a moving average with a window size of 7 days is applied to each player population series to extract the first approximation of the trend by excluding the impact of weekly seasonality. Here a value at a certain point of the series is approximated by calculating the mean of values within a 7 days window surrounding that data point. Figure 2 is the time series after smoothing.

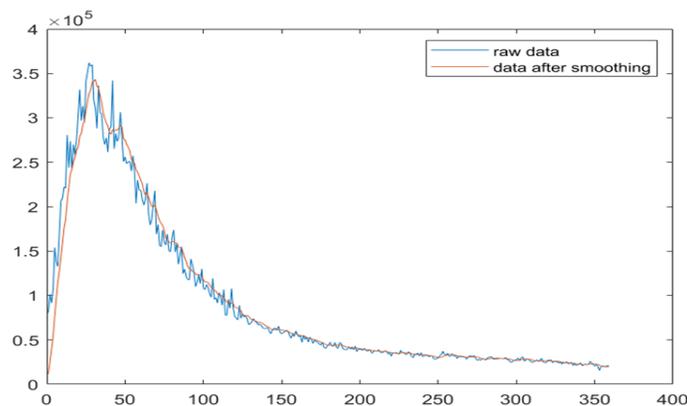


Figure 2: Smooth series to eliminate seasonal trend

2.2 Piecewise Linear Regression for Life Cycle shape extraction

Because we do not know how many segments the time series should be divided into, we use the segmented unknown time series extraction algorithm [1].

The algorithm we designed can extract the time series linearly according to its characteristics by iterative method, take one month (30 days) as the shortest time series, use the sliding window to continuously add a single life cycle segment, take RMSE as the local minimum as the target, and extract the segment. Figure 4 is the schematic diagram of the sliding window.

Find out that the reason why RMSE is the local minimum is that there is a certain contradiction between the error and the length of the time series to a certain extent. Therefore, it is necessary to

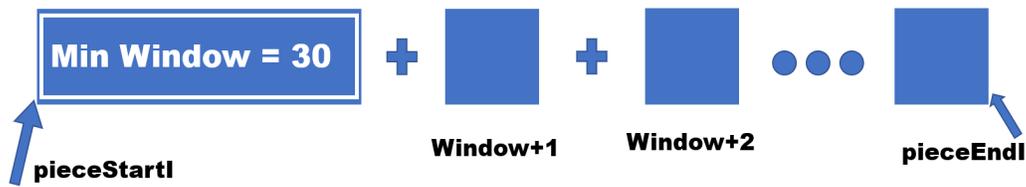


Figure 3: Schematic diagram of sliding window

balance it in the algorithm, and find out that the error of window length+1 (RMSE) is less than or equal to the error of the original window length, so as to find the length of the time series with the local RMSE minimum. The specific linear trend extraction code is detailed in the appendix.

At the same time, in the time series division, we also observed that part of the sequence will increase the error rate with the increase of the window length, which is not consistent with our previous criteria. In view of this situation, we decided to relax the criteria of the local optimal solution and judge it by the artificially given RMSE threshold. The function(1) are as follows:

$$\frac{error_{i+1} - error_i}{error_i} \leq RmseThreshold \tag{1}$$

According to formula(1), when the RmseThreshold value is set to 0, it is the first case mentioned in this paper. With the increasing of the threshold value, it can be assumed that the window of the local optimal solution will also increase. Therefore, to prevent the error from increasing with the increase of the window sequence, we select the RmseThreshold value to 0.01 to avoid excessive increase of the window length and increase the time complexity.

• **Results and Discussion**

In this section, we use the time series extraction algorithm above to divide the time series of Worldeer games into four stages: up, saturation, down and stationary, as shown in Figure 4.

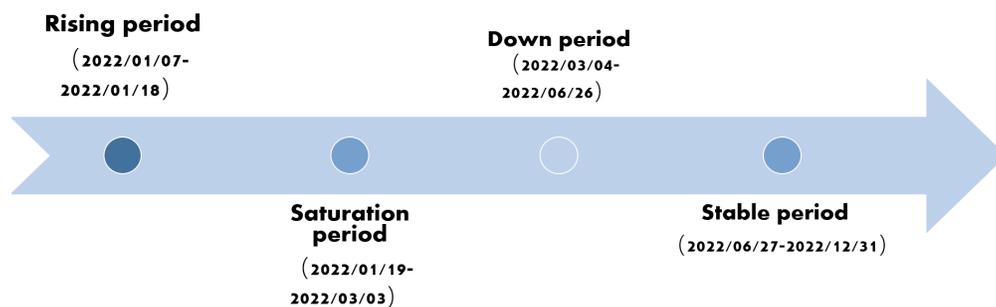


Figure 4: Life cycle time division

Our article serializes time to X_i , for example, 2022-1-7 is converted to $X_1=1$, and so on, converting 359 date values to $X_i(i=1,2,3...359)$, and set the number of reports to $f(X_i)$, $i = 1, 2, 3...359$. Figure 5 shows the curve fitting of four time series. In the fitting, we choose the method with the greatest goodness of fit for each curve. The average RMSE of the four curves is 2563.

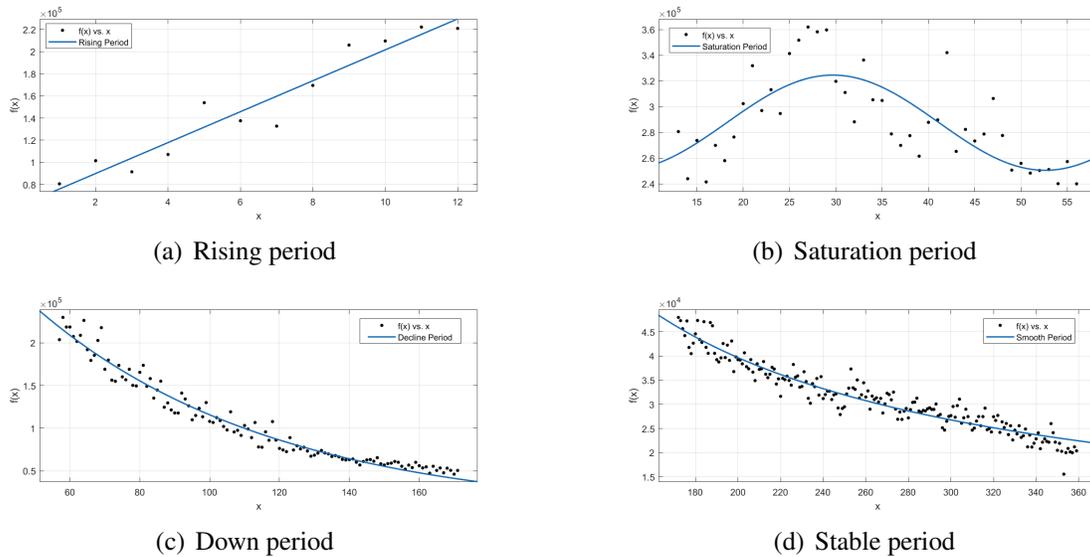


Figure 5: Life cycle curve fitting

Rising period($X_i = 1 \sim 12$):

$$f(x) = 1.396 * 10^4 x + 6.202 * 10^4 \quad R^2 = 0.9256$$

Saturation period($X_i = 13 \sim 56$):

$$f(x) = 2.875 * 10^5 - 2.226 * 10^4 \cos(0.137x) - 2.956 * 10^4 \sin(0.137x) \quad R^2 : 0.559$$

Down period($X_i = 57 \sim 171$):

$$f(x) = 5.069 * 10^5 \exp(-0.01477x) \quad R^2 = 0.9648$$

Stable period($X_i = 172 \sim 359$):

$$f(x) = 6.722 * 10^6 x^{-0.9688} \quad R^2 = 0.9171$$

2.3 Time series supplementary model

The data of life cycle model is missing in the prediction, so this paper makes use of time series model to supplement the prediction value. Therefore, we use the ARIMA time series model to describe the past, analyze the law and predict the future.

- **Stability Test**

In the stationary test, the single root test (ADF) is used. The ADF test is to determine whether a sequence has a unit root: if the sequence is smooth, there is no unit root. The table 1 is the result of ADF test, the p value is significantly greater than 0.05, indicating that the time series data is not stable. At the same time, according to the results of the t-test, the original hypothesis is rejected.

- **First Order difference**

Because of the instability of time series data, the first-order difference processing is used to transform the time series formed by adjacent periods, that is, subtracting the previous period from

Table 1: Stability Test

Variable	t	P	Threshold		
			1%	5%	10%
USD	-1.256	0.6490	-3.451	-2.876	-2.570

the latter period[6].

$$y_t = y_0 + \sum_{i=1}^t \varepsilon_i \quad (2)$$

Using Formula (7), we carry on the difference processing to the data,MacKinnon approximate p-value for $Z(t) = 0.0000$, which is far less than 0.05, indicating that the data has been stable.

- **Build ARIMA model**

By combining the autoregressive model (AR), moving average model (MA) and difference method, we get the differential autoregressive moving average model ARIMA (p, d, q) , where d is the order of the data to be differentiated.

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \dots + \alpha_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \dots + \beta_q \varepsilon_{t-q} \quad (3)$$

For a time series, there may be more than one significantly effective model. At this time, the range of p and q can be determined by autocorrelation diagram and partial correlation diagram. To eliminate subjective factors, we use AIC and BIC to determine model parameters.

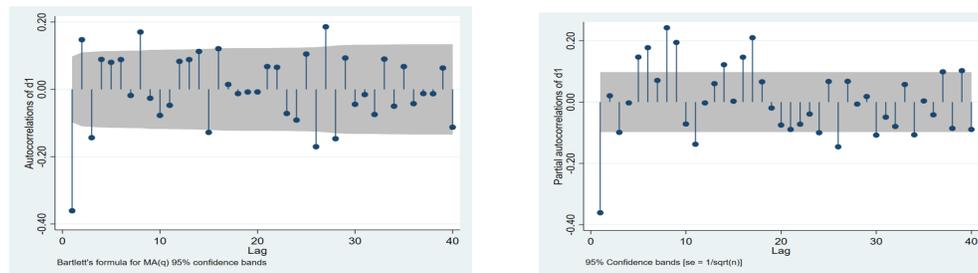


Figure 6: Autocorrelation and partial correlation

Table 2: ARIMA model evaluation index

	AIC	BIC
ARIMA(2,1,12)	8561.355	8625.298
ARIMA(3,1,12)	8529.636	8593.579
ARIMA(3,1,11)	8548.890	8612.833
ARIMA(4,1,12)	8550.963	8622.899

According to ACF, PACF diagram and table 2 we can see that the parameters of the Arima model are $p = 3, q = 12$ and $d = 1$ because we use the first order difference. The ARIMA(3,1,12) model was selected, the least squares estimation result of the model is as follows:

$$\begin{aligned}
 y_t = & 250.163 + 0.949y_{t-1} - 0.881y_{t-2} + 0.783y_{t-3} + \varepsilon_t - 1.594\varepsilon_{t-1} + 1.748\varepsilon_{t-2} \\
 & - 1.761\varepsilon_{t-3} + 1.117\varepsilon_{t-4} - 0.285\varepsilon_{t-5} + 0.161\varepsilon_{t-6} - 0.059\varepsilon_{t-7} - 0.002\varepsilon_{t-8} \\
 & - 0.305\varepsilon_{t-9} + 0.158\varepsilon_{t-10} - 0.189\varepsilon_{t-11} + 0.377\varepsilon_{t-12}
 \end{aligned}
 \tag{4}$$

Where y_{t-1} represents the number of reported results in the previous moment, ε_i represents Residuals at i moments.

The emergence of GARCH model solves the problem that the traditional autoregressive conditional heteroscedasticity model can not be used to predict the situation with many lagging ord[8].

- **A Lagrange multiplier test for the ARCH effect**

The aggregation effect, namely conditional heteroscedasticity, often occurs in the prediction of time series, that is, the conditional variance σ_t^2 of the residual term u_t depends on the size of u_{t-1}^2 . We use LM test method, if $LM > \chi_\alpha^2(q)$, then refuse H_0 , thinking exists ARCH effect.

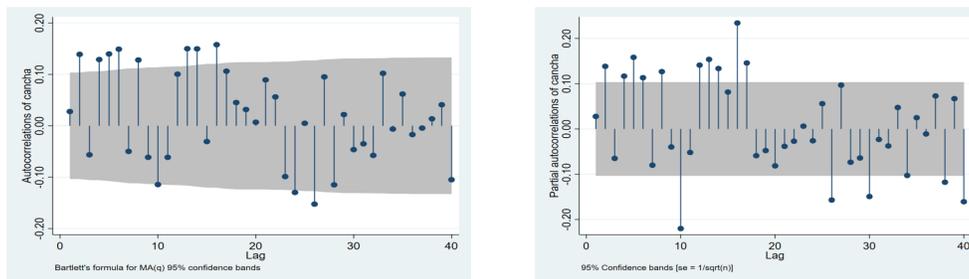


Figure 7: First Order post-difference time series

According to the test results, the p value of the residual sequence is significantly equal to 0 at about 20 orders, indicating that the residual sequence of the ARIMA (3, 1, 12) equation has ARCH effect, so the residual square term of the mean model is applicable to the GARCH model for fitting.

- **Establish ARIMA(3,1,12)-GARCH (1,1) model**

Table 3: GARCH model evaluation index

	AIC	BIC
GARCH(1,1)	7891.911	7962.429
GARCH(2,1)	7900.111	7968.135
GARCH(1,2)	7894.405	7963.936
GARCH(2,2)	7923.108	7995.134

The AIC and BIC values of GARCH(1,1) are small, so the GARCH(1,1) model is selected

The GARCH (1,1) model is established in this paper. The AIC, BIC and HQ values of the model are the smallest as a whole, and its effect is the best. The model expression is as follows:

Mean Value equation:

$$y_t = -50.570 + 1.778y_{t-1} - 1.263y_{t-2} + 0.465y_{t-3} + \varepsilon_t - 2.391\varepsilon_{t-1} + 2.313\varepsilon_{t-2} - 1.181\varepsilon_{t-3} + 0.263\varepsilon_{t-4} + 0.025\varepsilon_{t-5} - 0.030\varepsilon_{t-6} + 0.042\varepsilon_{t-7} - 0.079\varepsilon_{t-8} + 0.132\varepsilon_{t-9} - 0.023\varepsilon_{t-10} + 0.304\varepsilon_{t-11} - 0.152\varepsilon_{t-12} \quad (5)$$

Variance equation:

$$\sigma^2 = 240813 + 0.750\sigma_{t-1}^2 + 0.237\varepsilon_{t-1}^2 \quad (6)$$

2.4 Model inspection

Next, we will conduct ARCH-LM inspection. Observe whether the established ARIMA(3,1,12)-GARCH (1,1) model has ARCH effect.

Table 4: ARCH LM test

Heteroskedasticity Test:ARCH			
F-statistic	0.001065	Prob.F(1,400)	0.9740
Obs*R-squared	0.001071	Prob.Chi-Square(1)	0.9739

At this time, the concomitant probability is 0.974, and the original hypothesis is not rejected. It is believed that there is no ARCH effect in the residual sequence, indicating that the GARCH (1,1) model eliminates the conditional heteroscedasticity of the residual sequence.

To sum up, we establish a lifecycle-ARIMA(3,1,12)-GARCH(1,1) mixed model to fit the percentage data of 359 words, and the goodness of fit is 0.986, indicating that the model performance is good.

2.5 Model predictions

Next, we use a hybrid model of the life cycle model and the time series model to predict the 2023.03.01 data. Taking the mean of the two predicted values as the predicted value of the hybrid model, the final predicted number is 20,982, with a 95% confidence interval of [17869,23354].

In 03/01, we introduce the concept of sample confidence interval in statistics, give the distribution of the total parameters of the sample, and use Formula (7) to calculate the 95% estimated interval of the predicted value.

$$Standarderror(SE) = \frac{Standarderror(SD)}{\sqrt{n}} \quad (7)$$

And the 95% forecast interval is [14689,28354]. The following is our prediction of 2023 data using the time series prediction model, as shown in Figure 8.

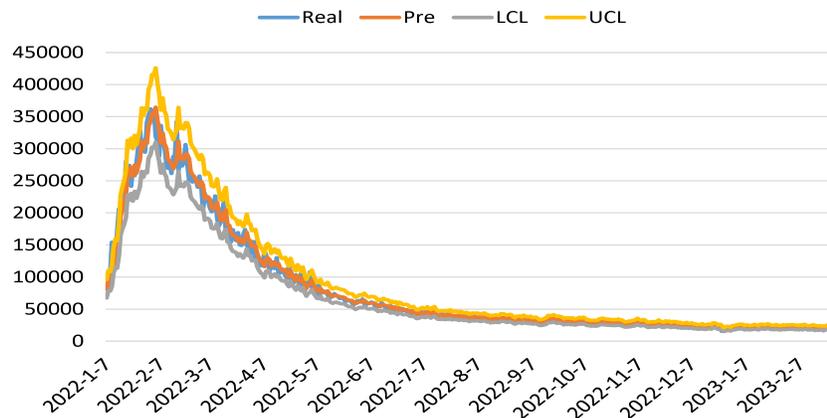


Figure 8: Time series prediction curve

2.6 Research on the attribute of words

To study attributes of the word affect the percentage of scores reported that were played in Hard Mode. This study used Pearson product-moment correlation coefficient to examine the degree of correlation of six word-related attributes.

- **Word form(Fm):**

The more forms a word has, the more likely it is to be used and the less difficult it is to guess, so the more word forms you set, the greater the property value.

- **Word Professionalism(Pf):**

Professional words are difficult to be used in daily life, so they are divided into professional words and non-professional words.

- **Frequency(Fq):**

The number of times each word is used per million in the Corpus of Contemporary American English (COCA) ("Corpus of Contemporary American English," n.d.) as its frequency in English.

- **Loanwords(Fo):**

Some words are derived from Latin and are therefore classified into native and foreign words, such as mummy and tiara.

- **number of repeated letters(Rp):**

The number of repeated letters in each word, e. g there are two 'a' in abbey, the number of repetitions is 2.

- **Orthographic neighbors(Orth):**

This is the number of orthographic neighbors that a string has. An orthographic neighbor is defined as a word of the same length that differs from the original string by only one letter. For example, given the word 'cat', the words 'bat', 'fat', 'mat', 'cab', etc. are considered orthographic neighbors.

Under the condition that the null hypothesis is established, a statistic is constructed by using the Spearman correlation coefficient so that it conforms to the standard normal distribution:

$$r_s \sqrt{n-1} \sim N(0, 1) \quad (8)$$

	Pre	Pf	Fq	Fo	Fm	Rp	Or
Pre	1	-0.073	★-0.144	-0.018	-0.032	0.063	-0.065
Pf	-0.073	1	0.053	0.075	0.019	-0.045	★0.106
Fq	★-0.144	0.053	1	0.099	★0.243	-0.046	0.091
Fo	-0.018	0.075	0.099	1	0.081	★-0.131	★0.124
Fm	-0.032	0.019	★0.243	0.081	1	★-0.125	★0.12
Rp	0.063	-0.045	-0.046	★-0.131	★-0.125	1	-0.049
Or	-0.065	★0.106	0.091	★0.124	★0.12	-0.049	1

★ Correlation is significant at the 0.01 level (two-tailed).
 ☆ Correlation is significant at the 0.05 level (two-tailed).

Figure 9: correlation matrix

We calculate the test value and find the corresponding P value to compare with 0.01, 0.05. According to Spilman's correlation coefficient test (Figure 12), the percentage of difficult modes in total players was not correlated with the word attributes we established, only the word frequency had a correlation of -0.144, we think a correlation between word frequency and percentage with 99% confidence.

3 Task 2: Report result forecast

3.1 Classification variable processing based on Apriori algorithm

Association rule mining allows us to discover the relationship between items from the dataset [11]. Our paper uses this algorithm to mine the relationship between word classification attributes and word difficulty. Then the problem is transformed into, when the word has which attributes, it is most likely to be difficult.

Step 1 Data processing: We transform several classification variables of word form (noun, verb, adjective, adverb, word specialization, foreign word) into frequent itemsets between the form search of the fact table and the difficulty of words. In terms of word difficulty processing, we cluster words by k-means according to seven attributes of the number of attempts, and divide them into two categories: difficult and simple. Table 5 is the data we have processed. Among them, 1 represents that the word belongs to this category. For example, the difficulty of main is divided into difficulty, and the word form is adjective, which is not a foreign word.

Table 5: Transaction table

Word	difficulty	Pf	Fo	n	v	adj	adv	pron
manly	1	0	0	0	0	1	0	0
molar	0	1	0	1	0	0	0	0
havoc	1	0	0	0	1	0	0	0
impel	0	0	0	0	1	0	0	0
condo	1	0	1	1	0	0	0	0

Step 2 Formulated description: Let $I = i_1, i_2 \dots i_n$ be a set of n binary attributes called items. Here $I = \text{Pf, Fo, n, v, adj, adv, pron}$. Let $D = t_1, t_2 \dots t_m$ be a set of transactions called the database. Here $D = \text{manly, molar} \dots \text{cramp}$ transaction in D has a unique transaction ID and contains a subset of the items in I . A rule is defined as an implication of the form:

$$X \Rightarrow Y, \text{ where } X, Y \subseteq I \quad (9)$$

Every rule is composed by two different sets of items, also known as itemsets, X and Y , where X is called antecedent or left-hand-side (LHS) and Y consequent or right-hand-side (RHS). In order to select interesting rules from the set of all possible rules, constraints on various measures of significance and interest are used. Here we use minimum thresholds on support and confidence method, we first need to define the definition of support and confidence. The support of X with respect to T is defined as the proportion of transactions in the dataset which contains the itemset X .

$$\text{supp}(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|} \quad (10)$$

Confidence is an indication of how often the rule has been found to be true. The confidence value of a rule, $X \Rightarrow Y$, with respect to a set of transactions T , is the proportion of the transactions that contains X which also contains Y . Confidence is defined as:

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} \quad (11)$$

Step 3 Association rule found: The association rules are extracted by the algorithm mentioned above, and the discrete variables are converted into rules and their support is taken as the score of each attribute value, and applied to the regression of the second question. Table 6 shows the size of the found association rules and confidence.

According to our association rules (table 6), we use the confidence degree to assign the difficulty to each situation. The difficulty of foreign words in the guessing game is significantly increased. At the same time, the difficulty of adjectives is significantly greater than that of nouns and verbs. At the same time, when the professional words are adjectives, the number of games is also significantly increased. This is consistent with our common sense.

From the word attributes of 5.1, we use each attribute as an independent variable to build a model. From the correlation coefficient diagram in 5.2, it is found that there is a relatively strong

Table 6: Association rules

rules	confidence	rules	confidence
{Pf} =>{difficulty}	0.273	{Pf,n} =>{difficulty}	0.200
{adv} =>{difficulty}	0.318	{n,v,adj} =>{difficulty}	0.200
{v} =>{difficulty}	0.416	{adj,adv} =>{difficulty}	0.286
{adj} =>{difficulty}	0.553	{Pf,adj} =>{difficulty}	0.333
{Fo} =>{difficulty}	0.778	{Fo,n} =>{difficulty}	0.778
{n} =>{difficulty}	0.418	{Fo,adj} =>{difficulty}	1.000
{n,v} =>{difficulty}	0.363	{n,adj} =>{difficulty}	0.450
{v,adv} =>{difficulty}	0.375	{Fo,n,adj} =>{difficulty}	1.000
{v,adj} =>{difficulty}	0.417	{v,adj,adv} =>{difficulty}	0.400

correlation between certain attributes, for example, the correlation coefficient between Rp and Fo is -0.16, and the correlation coefficient between Rp and Fm is -0.13, but the correlation is not particularly significant. Then use SPSS to test for multicollinearity, the variance ratio is close to 1, which indicates the existence of serious collinearity.

Then refer to the eigenvalues and conditional indicators in the collinearity diagnosis. When the eigenvalue is approximately equal to 0, the value of the conditional index is greater than 10, and the variance ratio is close to 1 (one of them is enough), it can indicate that there is relatively serious collinearity.

Table 7: Multicollinearity test

Dimension	Eigenvalues	Condition Index	Variance Ratio	Fq	Mi	Rp	Orth
1	2.96	1.00	0.02	0.01	0.02	0.00	0.01
2	0.86	1.86	0.00	0.99	0.00	0.00	0.00
3	0.68	2.09	0.01	0.00	0.03	0.00	0.98
4	0.38	2.78	0.02	0.00	0.22	0.04	0.00
5	0.12	5.00	0.95	0.00	0.73	0.88	0.00

The result of multicollinearity test(table7): ① The eigenvalue of dimension 5 is approximately equal to 0; ② The variance ratio of Fq dimension 2, the variance ratio of Orth dimension 3, the variance ratio of Mi and Rp dimension 4 are all close to 1. It shows that there is serious collinearity.

3.2 Establishment of Partial Least Squares Regression Analysis Model

Due to the small amount of data in this question and the serious collinearity among the independent variables, the partial minimum double regression model was used to establish the model[12].

Then we perform partial least squares regression. Using partial least squares regression analysis, we obtained a regression equation with 7 trial values as dependent variables and 4 word attributes as independent variables.

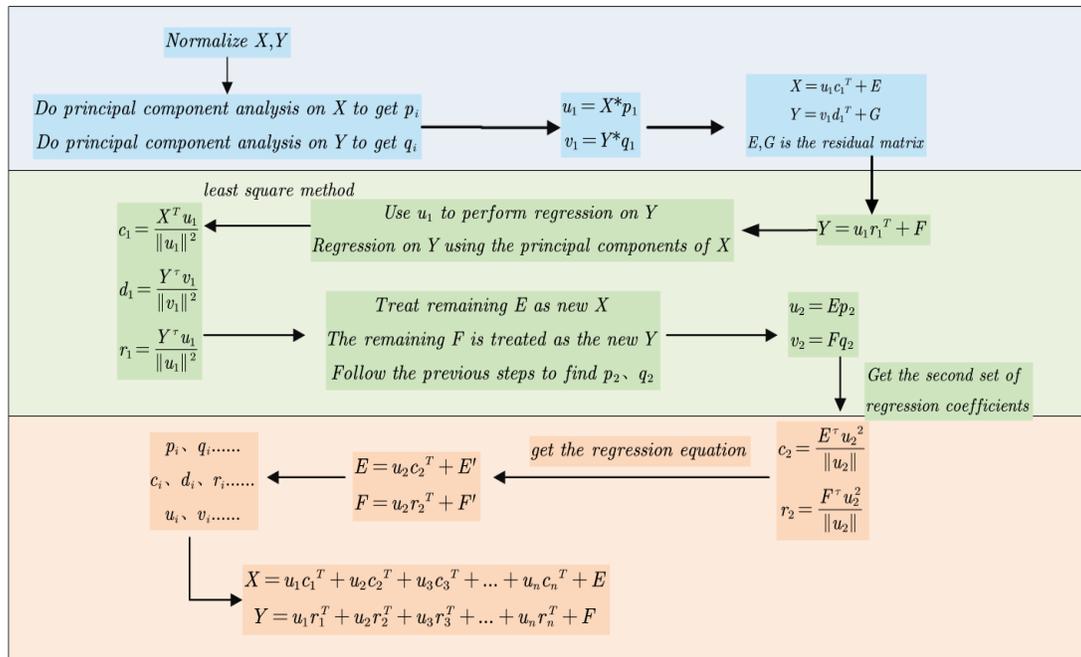


Figure 10: Flow chart of partial least squares regression analysis

$$N(x) = \begin{cases} 0.464 + 0.001 * Fq - 0.154 * Rp + 0.061 * Or - 0.747 * Mi, x = 1 \\ 5.923 + 0.008 * Fq - 1.372 * Rp + 0.27 * Or - 2.335 * Mi, x = 2 \\ 23.871 + 0.02 * Fq - 3.214 * Rp - 0.185 * Or + 2.768 * Mi, x = 3 \\ 34.441 - 0.001 * Fq - 0.549 * Rp - 0.796 * Or + 6.606 * Mi, x = 4 \\ 23.783 - 0.014 * Fq + 2.361 * Rp - 0.295 * Or - 0.55 * Mi, x = 5 \\ 10.189 - 0.01 * Fq + 2.107 * Rp + 0.409 * Or - 3.52 * Mi, x = 6 \\ 1.288 - 0.003 * Fq + 0.777 * Rp + 0.524 * Or - 1.776 * Mi, x = 7 \text{ or } X \end{cases} \quad (12)$$

The index of word frequency has a positive impact on 1, 2, and 3 tries, and a negative impact on the value of the remaining multiple attempts, which means that the higher the frequency, the simpler the word. In contrast, the number of word repetitions had a positive effect on multiple attempts, meaning that the more word repetitions, the harder the word. The impact of orthogonality and mixed indicators is relatively random, indicating that their data are random.

Then we analyze the word attributes of EERIE, substitute it into the regression equation, and get the prediction of the percentage of EERIE attempts.

Table 8: EERIE’s forecast results

	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)
EERIE	0	2	15	33	30	16	4

3.3 Uncertainties in Models and Forecasts

- **Social media:** Social media platforms are an important channel for Wordle games to spread online. If a certain social media platform has many users sharing their game results and experiences, then this may attract more players to try the game.
- **Current news:** Certain current news events or hot topics may directly or indirectly affect the popularity of Wordle games. For example, the frequent occurrence of a certain word in a certain news event may drive more people to try that word in the game.
- **Word of mouth and recommendation:** Wordle is a popular game that benefits greatly from positive recommendations and reviews from players. When individuals share their positive experiences playing the game on social media or forums, it can attract more people to try it out.
- **Global epidemic:** The global epidemic has also had a certain impact on the popularity of Wordle games. Due to people being restricted at home during the epidemic, the demand for games has increased a lot, which may also be one of the important reasons why Wordle games are popular during the epidemic.

3.4 Evaluate the model prediction results

Table 9: evaluating indicator

	MSE	RMSE	MAPE	MAE
1 try	0.55	0.74	48.45%	0.51
2 tries	13.75	3.71	66.70%	2.69
3 tries	48.27	6.95	32.44%	5.68
4 tries	23.82	4.88	12.62%	3.72
5 tries	27.83	5.28	20.71%	4.28
6 tries	32.79	5.73	51.97%	4.46
7 or more tries (X)	14.62	3.82	87.99%	2.01

It can be seen from the prediction results that() the Mean Absolute Error and Root Mean Square Error are relatively small, indicating that the prediction results of the model are good; and in the evaluation index of Mean Absolute Percentage Error, "7 or more tries (X)" is removed This item, the rest of the MAPE values are within the acceptable range. Therefore, we can consider the model's prediction to be good, and we have confidence in it.

4 Task 3: Word difficulty classification model

In task 3, this paper uses the system clustering method to cluster words into three categories: "difficult", "medium" and "easy", with the contour coefficient of 0.567. In order to test the clustering effect, this paper uses Fisher discriminant analysis to back judge based on training samples, and measures the reliability of clustering analysis by the obtained confusion matrix.

4.1 Hierarchical clustering

We use the seven percentage values of words as the clustering index, and use the system clustering model for clustering. We find out the cluster centers, so that we can use the Euclidean distance to divide the difficulty of words, and divide the difficulty into three categories through the elbow rule of the system clustering, that is: simple, moderate and difficult. Figure 11 is our clustering result graph.

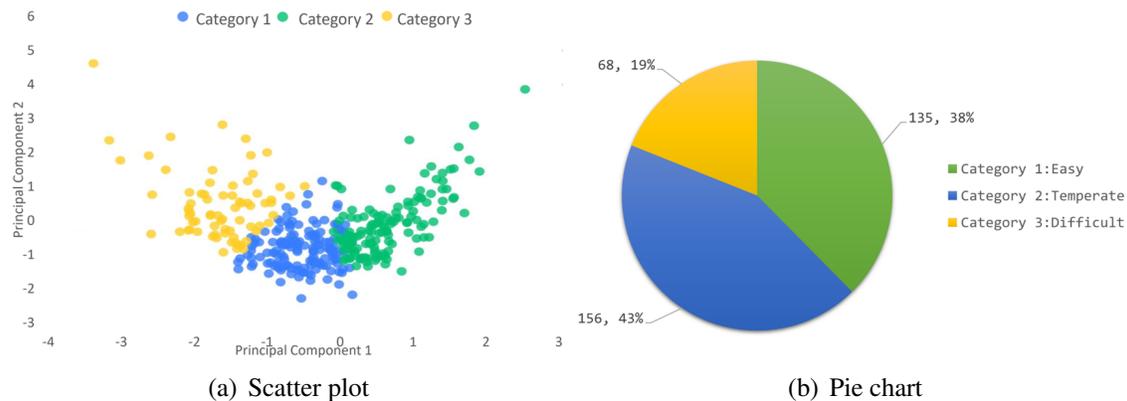


Figure 11: Clustering result graph

The first category (Figure 11): simple category, a total of 135 words, the average difficulty score is 3.74. The second category: Moderately difficult, with a total of 156 words and an average difficulty score of 4.13. The third category: Difficult category, a total of 68 words, with an average difficulty score of 4.20.

Table 10: Category attribute recognition results

	Frequency	Blending indicators	Repetitions	Orthogonality
Category 1: Easy	2.19	0.16	0.19	1.90
Category 2: Temperate	1.95	0.16	0.64	1.57
Category 3: Difficult	1.73	0.15	0.91	2.30

It can be seen from the cluster center data that the word frequency attribute value of category 1 is relatively large, while the letter repetition value is small. It can be seen that relatively simple words often appear frequently in life, and people can easily think of them; and their letters there are fewer repetitions, and people don't need to worry about how to arrange the repeated letters when playing the game.

For category 2, the indicators of the words in this category are relatively balanced, and there are no extreme values, which shows that the words with moderate difficulty selected by the game are also in line with common sense.

For category 3, its word frequency is the least, indicating that those words that do not appear in life, people often ignore them and it is difficult to think of them; at the same time, because people

often subconsciously want to try out more letters when playing games, will cause them to tend to guess different letters when guessing, and it is difficult to find the appearance of the same letter, so there will be more words containing the same letter in the difficult category; the orthogonality value of words in this category is higher, and it may be because people ignore words that have the same word length but differ by only one letter.

4.2 Fisher linear discriminant analysis

According to the above clustering results, we then perform Fisher[13] linear discriminant analysis based on the percentage data of 359 words to judge the clustering results.

The core idea of Fisher's discrimination is projection, trying to find an optimal projection vector or optimal discriminant function, so that the sample data is projected in this direction, and the discriminant is determined based on the principle that the within-group dispersion is as small as possible and the inter-group dispersion is as large as possible. function, and then determine the sample category according to the discriminant function. Assuming that there are l populations G_1, G_2, \dots, G_l and the observation samples are $x_{i1}, x_{i2}, \dots, x_{iq_i}$ ($i = 1, 2, \dots, l$), then the sum of squares of variance between groups and the sum of squares of variance within a group of sample data x_{ij} are respectively

$$SSA_x = \sum_{i=1}^l q_i (\bar{x}_i - \bar{x})^2, SSE_x = \sum_{i=1}^l \sum_{j=1}^{q_i} (\bar{x}_{ij} - \bar{x}_i)^2 \quad (13)$$

Suppose the projection vector is p , then the projected data of the observation sample is $y_{ij} = p'x_{ij}$, and the linear relationship is the corresponding discriminant function, then the sum of squared deviations between groups and the sum of squared deviations within a group of projected data are y_{ij} respectively

$$SSA_y = \sum_{i=1}^l q_i (\bar{y}_i - \bar{y})^2 = p' SSA_x p, SSE_y = \sum_{i=1}^l \sum_{j=1}^{q_i} (\bar{y}_{ij} - \bar{y}_i)^2 = p' SSE_x p \quad (14)$$

When the objective function $f(p) = (p' SSA_x p) / (p' SSE_x p)$ reaches the maximum value, the projection vector p obtained at this time is the best. In order to ensure the uniqueness of the solution, assuming that SSA_x / SSE_x is the unit matrix E , the partial derivative is derived

$$(SSE_x)^{-1} \cdot SSA_x p = \lambda p \quad (15)$$

The maximum eigenvalue of $f(p)$ obtained at this time is the maximum value of the objective function $(SSE_x)^{-1} \cdot SSA_x$, and the corresponding eigenvector is the optimal projection vector, so that the linear discriminant function is obtained as $y_{ij} = p'x_{ij}$ ($i = 1, 2, \dots, l; j = 1, 2, \dots, q_i$). According to the discriminant function, the projection matrix of the observation sample and the corresponding Based on the group mean projection matrix, the sample discrimination result can be obtained based on the principle of the minimum distance between the vectors of the distance discrimination method.

We get the discriminant functions as:

$$y_1 = -8739.167 + 152.863 * N_1 + 169.141 * N_2 + 175.642 * N_3 + 177.733 * N_4 + 170.664 * N_5 + 179.317 * N_6 + 172.743 * N_7 \quad (16)$$

$$y_2 = -8754.238 + 153.61 * N_1 + 169.251 * N_2 + 175.21 * N_3 + 178.164 * N_4 + 170.975 * N_5 + 179.603 * N_6 + 173.218 * N_7 \quad (17)$$

$$y_3 = -8786.136 + 153.849 * N_1 + 169.284 * N_2 + 175.447 * N_3 + 178.103 * N_4 + 171.31 * N_5 + 180.62 * N_6 + 173.859 * N_7 \quad (18)$$

4.3 Classification accuracy

Table 11: Evaluation metrics for hierarchical clustering

Contour factor	DBI	CH
0.567	0.919	309.688

According to the results of system clustering evaluation(table 13), the silhouette coefficient is 0.567, which is high in the range of [-1,1], DBI is less than 1, and CH is high, indicating that the clustering effect is better.

Table 12: Evaluation indicators for discriminant analysis

	Accuracy	Recall rate	Positive Accuracy	4 tries
Training set	0.956	0.956	0.958	0.956
Test set	0.944	0.944	0.947	0.945

In the linear discriminant model(table 12), the precision rate, recall rate, positive precision rate, and F1 of the training set and test set are all around 0.95, which means that the clustering reliability is high and the classification accuracy is good.

5 Task 4: Analysis of data set characteristics

The difficulty mode scale has shown a consistent upward trend over time. With the outbreak of COVID-19 in late 2020, the game attracted a surge of new players, although many lost interest due to the game's weak appeal. However, the remaining players became more daring and adventurous.

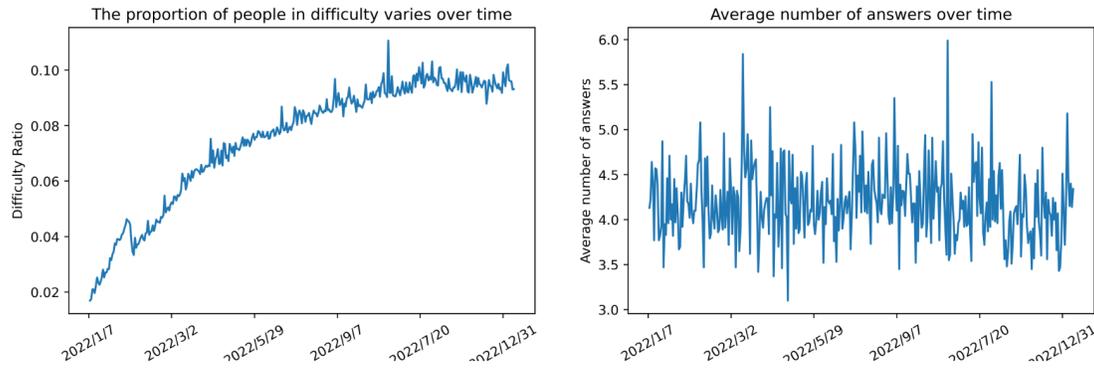


Figure 12: correlation matrix

With continuous promotion of the game, more people have now mastered the rules and are keen on pushing their limits by taking up tougher challenges.

The right graph is The average number of challenges with The change of time series, you can see The overall trend of 4times, only a few times and a number of times The situation, we can think of The difficulty of The problem every day is random, occasionally difficult and simple words appear.

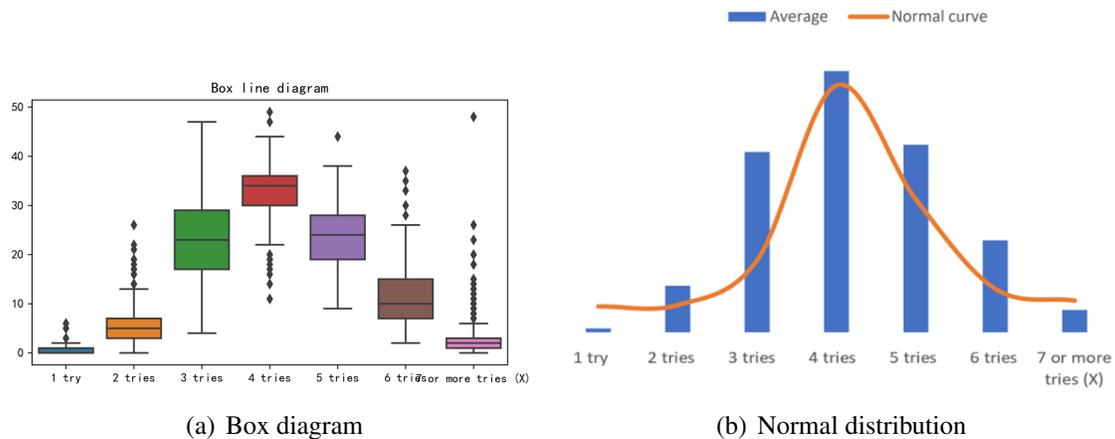


Figure 13: Analysis of game times

We draw a box chart and a bar chart for the percentage of each number of times. It can be seen that the number of answers conforms to the normal distribution. At the same time, four times are the mode of the number of answers. The number of outliers of number 7 is the largest, and the variance of number 3 is large. Through the analysis of the number of answers, we can see that our previous method of clustering by the number of answers is reasonable.

Table 13: Comparison of ARIMA models with different parameters

	ARIMA(3,1,12)	ARIMA(0,1,13)
R-Square	0.988	0.986
RMSE	9823.287	10319.81
MAPE	7.524	7.092
MAE	5492.645	5717.193

6 Model sensitivity analysis

6.1 Tests for the parameters of ARIMA

We compare the model ARIMA (0, 1, 13) automatically generated by SPSS software with our ARIMA (3, 1, 12) model determined by the IC criterion, autocorrelation diagram and sliding window data of the life cycle, we can see, when $p=0$, $q=13$, the late prediction accuracy of the model is not high, and the difference from the actual value is large, and the R-Square, RMSE and other indicators are also large, indicating that the prediction error is large. It can be seen that the model parameters we choose are more reasonable and the model sensitivity is better.

6.2 Test of Regression Equation Coefficients

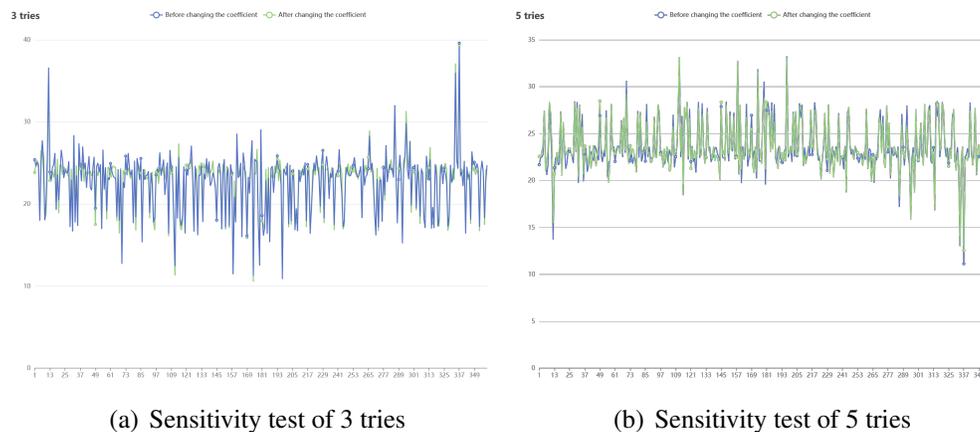


Figure 14: Sensitivity test

In order to determine the sensitivity of our regression model to different time series data, a sensitivity analysis was carried out. When establishing the regression equation prediction, the standard error of the regression coefficient measures the reliability of the estimated value of the regression coefficient. Generally, the smaller the standard error, the higher the sensitivity of the model. We manually change the coefficient of the independent variable of the word occurrence frequency of the regression equation by 10%, and then bring the data in. It can be seen that the percentage of 3 attempts and 5 attempts after the change is relatively small, but the frequency of word occurrence is the biggest factor affecting the regression equation. Therefore, our regression equation has no obvious change to the coefficient, and the model has good sensitivity.

The analysis shows that when the maximum factor changes within a certain range, the change range of the final percentage is 3.68%. Therefore, when other factors with small influence change, the percentage of attempts will not fluctuate significantly, which indicates that the proposed model is quite robust to the word attribute.

7 Conclusion

Strengths

- The model and strategy based on pairs trading are scientific and reasonable, which can maximize the income. The results obtained have strong confidence.
- The model uses the AP algorithm to assign the classification variables, which is more accurate and accurate than using only the classification variables to build the model. At the same time, the partial least squares regression model is mostly used for multi-output calculation, which is more in line with the actual situation of the topic.
- The model uses Fisher algorithm to test its clustering results in clustering, and the model is more reliable.

Weaknesses

- The word attribute used by the model is only the most commonly used attribute variable in current linguistic research, while there is not much research on uncommon variables, such as vowel consonant combination, irregular spelling and other attributes, which will be cited in our subsequent research.
- The model is directly smoothed in the processing of time series without analyzing factors such as special festivals, which is also reflected in the uncertain factors in Task 3.

Future outlook

- In our model, only quantifiable word attributes are considered, and the essence of language, such as vowel and consonant frequency, word pronunciation and spelling, etc., we have not considered them. We will conduct a comprehensive analysis of them in our future work. At that time, there will be new progress in the research work on word attributes, difficulty, etc.

References

- [1] Wannigamage D. Player Population Patterns in Digital Games: A Data Analytics and Machine Learning Approach. UNSW Sydney, 2021.
- [2] De Silva N. Selecting Optimum Seed Words for Wordle using Character Statistics//2022 Moratuwa Engineering Research Conference (MERCCon). IEEE, 2022: 1-6.
- [3] <https://thamara.blog/a-wordle-hack-aa83912cd979>
- [4] Sidhu D. Wordle—the best word to start the game, according to a language researcher[J].
- [5] Anderson B J, Meyer J G. Finding the optimal human strategy for wordle using maximum correct letter probabilities and reinforcement learning[J]. arXiv preprint arXiv:2202.00557, 2022.
- [6] Hamilton J D. Time series analysis[M]. Princeton university press, 2020.
- [7] Wang Yan. Research on stock price prediction based on ARFIMA-GARCH-LSTM hybrid model. Shanghai Normal University,2022.DOI:10.27312/d.cnki.gshsu.2022.000752.
- [8] Bauwens L, Laurent S, Rombouts J V K. Multivariate GARCH models: a survey[J]. Journal of applied econometrics, 2006, 21(1): 79-109.
- [9] Hassani H, Yeganegi M R. Selecting optimal lag order in Ljung–Box test[J]. Physica A: Statistical Mechanics and its Applications, 2020, 541: 123700.
- [10] <http://www.neuro.mcw.edu/mcword/>
- [11] Efrat A R, Gernowo R. Consumer purchase patterns based on market basket analysis using apriori algorithms[C]//Journal of Physics: Conference Series. IOP Publishing, 2020, 1524(1): 012109.
- [12] Maulud D, Abdulazeez A M. A review on linear regression comprehensive in machine learning[J]. Journal of Applied Science and Technology Trends, 2020, 1(4): 140-147.
- [13] Li C, Wang B. Fisher linear discriminant analysis[J]. CCIS Northeastern University, 2014: 6.

8 MEMO

To:Times puzzle editor

From:MCM Team 2314475

Subject:Prediction Model and Word Difficulty Classification Model Based on Wordle

Date:February 20, 2023

Dear Times puzzle editor

I am writing to you as MCM Team 2314475 to summarize our results on predicting Wordle.

In the first question, we developed a life-cycle model to account for the variation in the number of reports, using a segmented linear trend extraction algorithm to find a suitable sliding window sequence with a minimum local RMSE as the objective, and dividing the time series into four periods. Based on this, we build an ARIMA(3,1,12) model and introduce a GARCH(1,1) model for residual correction. We use a life-cycle-ARIMA(3,1,12)-GARCH(1,1) mixture model to forecast the data for 2023-3-1. Second, we reviewed the data and identified multiple word attributes that were not correlated with it, except for a significant correlation between the frequency of word occurrences and the percentage of scores reported in the difficulty model.

In the second question, we used the Apriori algorithm to produce associations between the three definite class word attributes and word difficulty to datum the variables; subsequently, a partial least squares regression analysis was used to establish a regression equation with the word attributes as independent variables and the seven percentages as dependent variables for the word 'EERIE', the word was predicted and its percentage results were obtained. Then, the uncertainties of the model are given, such as media coverage and game word-of-mouth. Finally, the regression model is evaluated and it is concluded that the model is more effective.

In the third question, we first used systematic clustering to predict the 359 words that were classified as "easy", "moderate", and "difficult". The words in the "difficult" category tended to be less frequent, with more letter repetitions and higher orthogonality, while the words in the "simple" category were less frequent. In order to verify the correctness of the clustering analysis, we introduced Fisher's linear discriminant analysis to test the clustering results back to the judgment. Based on this model, the word "EERIE" was categorized as "simple".

Finally, there are some other interesting features of this dataset that are worth mentioning. For example, we note that the overall number of players decreases over time but the percentage of those who pass the hard mode tends to increase, implying that veteran players prefer to challenge themselves and may be gradually discovering patterns and improving their solving skills. In addition, we found that the percentage of different attempts conformed to a normal distribution.

In summary, we believe our analysis provides valuable insights into the patterns and trends that exist in the dataset provided. We hope these findings will be of interest to your readers, and we look forward to hearing any feedback or comments you may have.

Sincerely

Your friends

Appendices

Algorithm : Piecewise Linear Trend Extraction

```

Input : Xdata, Ydata, minPieceSize, threshold
seriesLen = length(Ydata)
pieceStartI = 1
while (pieceStartI < seriesLen) do
    pieceEndI = pieceStartI + minPieceSize - 1
    if pieceEndI > seriesLen then
        pieceEndI = seriesLen
    end if
    bestFit = find the best linear fit for the data subset (pieceStartI : pieceEndI)
    error = calculate the error (RMSE) of the fit
    if error  $\leq$  threshold then
        save pieceStartI and pieceEndI
        pieceStartI = pieceEndI + 1
    else pieceEndI = pieceEndI + 1
        while (pieceEndI < seriesLen) do
            newFit = find the best linear fit for the data subset (pieceStartI :
pieceEndI)
            newError = calculate the error (RMSE) of the new fit
            if newError > threshold then
                break
            else
                bestFit = newFit
                error = newError
                pieceEndI = pieceEndI + 1
            end if
        end while
        save pieceStartI and pieceEndI
        pieceStartI = pieceEndI
    end if
end while
if remainingLen  $\leq$  minPieceSize then
    merge remainingLen to the last piece
end if

```
