

Data research based on ARIMA and Gray Relational Analysis

Yang Li¹, Haoliang Han²

¹Department of design studies, Tongji University, Shanghai, China

²Department of Information Science & Technology, Beijing Institute of Technology, Beijing, China

Abstract. The New York Times' word-guessing game Wordle has swept the internet and ignited a Wordle craze on Twitter. In this paper, we build four mathematical models for the Wordle game puzzle report and do further research. First, we used the ARIMA model for prediction, and the Gray Correlation model was used to find the relationship between word attributes and the percentage of hard-mode participants. Next, we resorted to the BP neural network model to forecast the distribution on the number of tries, and the result was (1, 9, 26, 29, 15, 10, 11). Then, we applied the K-means model to categorize all the words based on their difficulty level, and EERIE was found to be in the "easy" range. In the end, most players are devoted game enthusiasts, and roughly 10 percent of players consistently picked the hard mode. Additionally, most players preferred simpler games for fun and relaxation, and the percentage of hard-mode participants increased among the overall reported number.

Keywords: ARIMA, BP neural network, K-means clustering model, Grey relational analysis.

1. Introduction

Wordle, a word-guessing game released by The New York Times, has gone viral on social media. In the game, players have six or fewer attempts to guess a five-letter word and receive feedback after each guess. It offers two modes of play, including a regular mode and a harder mode that requires players to use previously guessed letters in subsequent attempts. Intrigued by the Wordle craze, we aims to analyze a dataset of Wordle puzzle reports on Twitter from January 7th, 2022, to December 31st, 2022 to gain some insights into Wordle gameplay.

2. ARIMA and Gray Relational Analysis

2.1. The ARIMA model

2.1.1 Construction of the ARIMA

The Autoregressive Integrated Moving Average (ARIMA) model is a popular linear model used in time series forecasting, which combines autoregressive (AR), integrated (I), and moving average (MA) methods. AR describes the relationship between current and historical values, I involve differencing to make a non-stationary series stationary, and MA models the accumulation of errors in the AR model. The model is denoted as ARIMA (p, d, q) where p , d , and q represent the AR, I, and MA terms, respectively. Building an ARIMA model involves the following four steps:

- a) Stabilization - Check if the time series is stationary using the Augmented Dickey-Fuller (ADF) test. If it is not, convert the non-stationary process into a stationary process using the differencing method.
- b) Selection of p and q - Use the autocorrelation function (ACF) and partial autocorrelation function (PACF) to determine the appropriate values for the autoregressive (p) and moving average (q) terms.
- c) White noise test - Check if the residual sequence is a white noise, which indicates that all useful information has been extracted and no further modeling is required.
- d) Prediction - Use the data of a certain number of consecutive days as the training set and the data of the next few days as the test set to predict future values. Evaluate the prediction results using the correlation coefficient R^2 .

2.1.2 Implementation of the ARIMA

a) After the ADF test, it appears that time series is stationary.
b) We apply ACF and PACF on the dataset when $d = 0$ and when $d = 1$. (See Figure 1 and 2)
Looking at Figure 1 and 2, the sequence is essentially stable when $d = 1$, as shown by the fact that the majority of the sequence's autocorrelation graph's (ACF) and partial autocorrelation graph's (PACF) numbers fall within the confidence range. Therefore, we pick the ARIMA (0, 1, 0) model for our forecast.

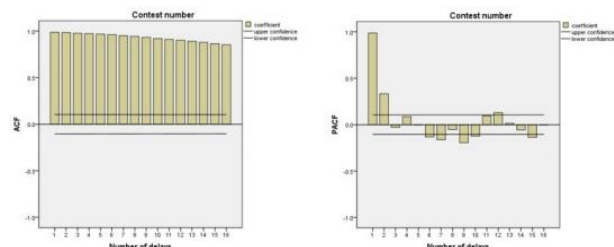


Fig. 1 ACF and PACF graph when $d = 0$

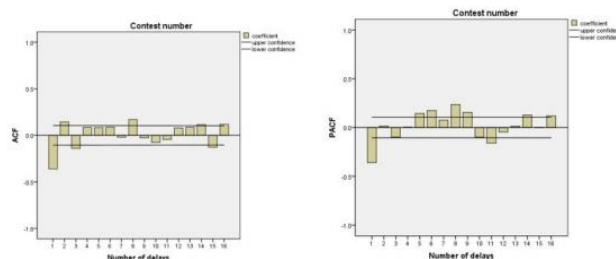


Fig. 2 ACF and PACF graph when $d = 1$

c) When we finish the white noise test, the outcome shows that the residual sequence is a white noise. (See Figure 3)
d) Following the above step 4, we use ARIMA to predict the number of reported results on March 1, 2023, and find that it passes the validity test. The range of estimation on the figure turns out to be (0.159771). (See Figure 4)

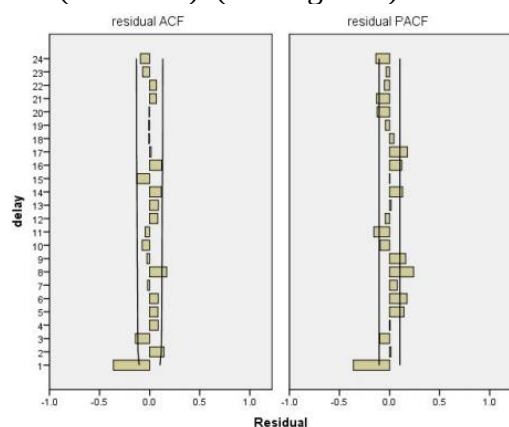


Fig. 3 The result of white noise test

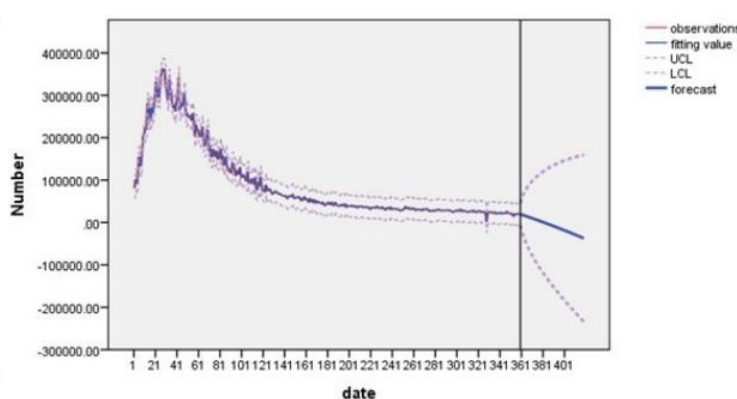


Fig. 4 Prediction of reported number on March 1, 2023

2.2. A Model based on Gray Relational Analysis

A multi-criteria decision-making method called Gray Relation Analysis aids in ranking alternatives when several criteria are present. Several fields, including engineering, economics, and finance, have extensively used it. The grey system theory, which deals with systems characterized by ambiguous and incomplete information, is the foundation of GRA. It offers a framework for handling systems in which the connections between the variables are not clearly defined.

The grey correlation analysis method is particularly practical since it can be used with any number of samples and any degree of sample regularity. It also has a modest calculation amount. The outcomes of the quantitative and qualitative analyses will be identical. As a result, obtaining both qualitative and quantitative effects is made simpler.

2.2.1 Construction of Gray Relational Analysis

According to the theory of the gray system, for this part of the problem, we can analyze the gray correlation degree of each subsystem, and use the method of dimensionless processing, calculating the difference, and calculating the correlation degree to overcome the dimensional difference of each factor. We also analyze each Quantitative subsystem analysis of the numerical relationship of the

system, to obtain the influence of each attribute on the result. After our observation of the data set, we pin down the following four attributes:

- Number of vowels in a word
- Number of repeated letters in a word
- Whether the word starts with vowels
- Average letter frequency

Then model could be split into three parts:

A. Define the number of players as the parent sequence, $x_0(k)$ as the influencing factor of each index, and the number of adjacent players as the sub-sequence.

B. Calculate the difference sequence $\Delta i(k)$ according to the following formula, and find the max-imum absolute difference $\Delta \max$ and the minimum absolute difference $\Delta \min$ in the calculation results.

$$\Delta i(k) = x'_i(k) - x'_0(k) \quad (1)$$

C. Find the correlation coefficient corresponding to each site, $q_i(k)$, $i = 1, 2, 3 \dots$ and correlation degree γ_i , $i = 1, 2, 3 \dots$

$$q_i(k) = \frac{\Delta \min + \rho \times \Delta \max}{\Delta i(k) + \rho \times \Delta \max} \quad (2)$$

$$r_i = \frac{1}{n} \sum_{k=1}^n q_i(k) \quad (3)$$

n is the sequence number of each word, ρ is the resolution coefficient, and the value is generally 0.5. The closer the value of γ_i is to 1, the better the correlation between this factor and the main direction of system development is.

2.2.2 Implementation of Gray Relational Analysis

Here are the steps of our analysis.

A. Dimensionless processing for the data.

B. Solve the gray correlation coefficient values between the parent series (comparison series) and the feature series.

C. Solve for the gray correlation values.

D. Rank the gray correlation values and draw conclusions.

The results are listed as follows:

The correlation coefficient represents the value of the degree of correlation between this subsequence, WordAvgFreq, VowelCount, RepeatCount, StartsWithVowel, HardModeCount pairs and the corresponding dimension of the parent sequence (the larger the number, the stronger the correlation). We are weight correlation coefficients to value between 0 and 1, and the larger the value is, the stronger the correlation with the “reference value” (parent series), which means the higher the evaluation. From the above table, we can see that HardModeCount has the highest evaluation (correlation: 0.953), followed by WordAvgFreq (correlation: 0.902). Besides, all the four factors mentioned above are related with the percentage of hard-mode participants. (See Figure 5 and 6)



Fig.5 The result of Gray Relational Analysis

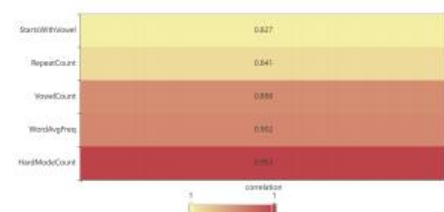


Fig.6 The result of Gray Relational Analysis

3. The BP neural network model

A typical example of a BP network is the BP neural network, which could be realized in Matlab and in which the error propagates the learning process in the opposite direction. To create a model that can produce results that are consistent with the expected outcomes, we carry out the error analysis repeatedly based on the training results and the expected results. Then we update the weights and thresholds incrementally. Here is the precise procedure:

3.1. Construction of the BP neural network

When we define the number of attempts in the hard mode as input, and the percentage of hard mode in all the reported cases as output, this problem is transformed into a complex function mapping problem where five layers make up the simulated hidden layer. This function serves as the activation function of BP neural network simulation, which onverts the processed values into the range of (0, 1), and discretizes continuous data into discrete data. The whole process could be divided into the following 2 steps:

A. The training set and test set use outlier removal and normalization processing to randomly serve as verification samples and training sets, and obtain a training set of 30 data out of order and a verification set including 5 data. Among them, the training set of 30 data is 1-6 attempts of 5 hidden layers, and the verification set is X attempts of 5 hidden layers.

B. The construction of the BP neural network maps the input of the neuron to the output by the function. In order to obtain the relationship between the input and output data, we use a function between the last hidden layer and the output layer. (See Figure 7 and 8) The expressions are as:

$$f(x) = \frac{1 - e^{-x}}{1 + e^{-x}} \quad (4)$$

$$yk' = \sum_{j=1}^r v_j \cdot f \left[\sum_{i=1}^m w_{ij} \cdot P_i \right] + \theta_j \quad k = 1, 2, 3, \dots, N \quad (5)$$

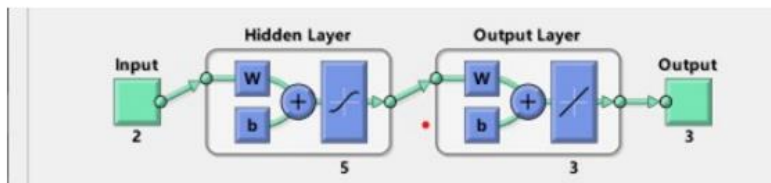


Fig.7 The structure of the the BP neural network maps(a)

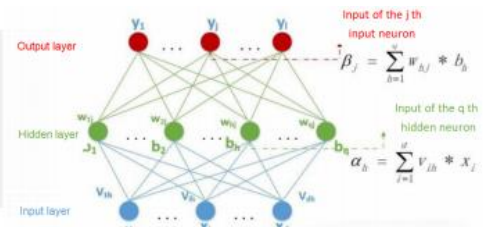


Fig.8 The structure of the the BP neural network maps(b)

3.2. Implementation of the BP neural network

Given the estimated and actual distribution of these seven sets of line graph in their trends and points, our BP neural network model has made good predictions in test sets. This shows that choosing five hidden layers for the neural network will make the results tend to be accurate. Of course, we also see points of discord in the test set, for example, the percentage of a fifth attempt and the predicted value have some deviation at Date=4, which also indicates that our neural network model still needs to be iterated and calibrated. Based on the same neural network as the previous training and test sets, we substitute the information related to EERIE, dated March 1, 2023, so that the number of people reported on that day can be predicted according to the ARIMA model from the first question, thus entering EERIE as a set of test sets into the BP neural network and obtaining the following distribution (1, 9, 26, 29, 15, 10, 11).

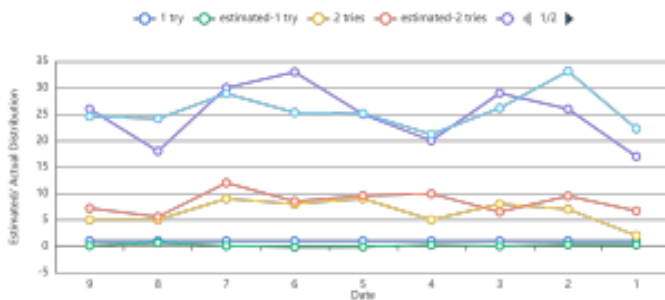


Fig.9 The result of the BP network

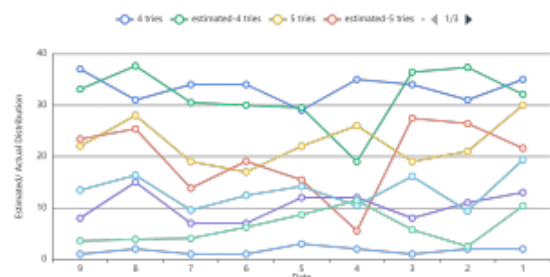


Fig.10 The result of the BP network

4. Model based on K-means

K-means is a popular unsupervised machine learning algorithm used for clustering data. The algorithm aims to group similar data points together into K clusters, where K is a pre-defined number chosen by the user. The algorithm iteratively assigns data points to the closest cluster center, and then updates the center of each cluster based on the mean of the data points assigned to it. The process continues until the cluster assignments no longer change, or a maximum number of iterations is reached.

4.1. Construction of K-means

We build K-means in two steps:

A. Decide how challenging the word is–Determine the difficulty level of various word characteristics to place the word EERIE in the appropriate category. Base on each word's score, we assess the complexity of the word. These are the guidelines:

- Make x_i the word with the simple mode scores of 1–7.
- Word scores 1 through 7 are, accordingly, y_i in tough mode.
- The outcome is A in regular mode and B in hard mode, but the difficulty level is $P_i = \frac{x_i}{A} + \frac{y_i}{B}$ for 1–7 points.

- Among them, the harder the word is, the lower the score.

B. Word classification–We use the K-means clustering model here. Proceed as follows:

- Randomly select k initial center points;
- For each sample point in the data set, calculate the distance between the sample point and k center points, and divide the sample point into the category corresponding to the nearest center point;
- After the category division, re-determine the center point of the category, and use the mean value of each feature of all samples in the category as the value of the feature corresponding to the new center point, that is, the centroid of all samples in the category;
- Repeat steps b and c above until a certain termination condition is reached (the number of iterations, the points in each cluster will not change). Next, calculate the proportion of word attributes of various elements.

C. Difficulty analysis–We incorporate the EERIE prediction findings from question 2 into the model (We determine difficulty value partially based on its attributes.) And create the aforementioned table for each polycenter simultaneously.

$$D_i = \sum (p_i - m_i)^2 \quad (6)$$

p_i and m_i represent the difficulty value of EERIE and cluster center on the i attribute, respectively. One can determine the word category when D_i is the smallest.

4.2. Implementation of K-means

The categorization of word EERIE be divided into the following four steps:

- A. Construct a word difficulty table.
- B. Output five clustering results:(See Figure 14 and 15)
 - a) Cluster 1: 95 samples, center point (3.96, 0.96842, 0.094737, 0.021053)
 - b) Cluster 2: 91 samples, center point (4.9267, 1.8571, 0.043956, 0.17582)
 - c) Cluster 3: 72 samples, center point (4.1627, 2.125, 0.11111, 0.16667)
 - d) Cluster 4: 75 samples, center point (5.6967, 1.9867, 0.053333, 0.2)
 - e) Cluster 5: 26 samples, center point (5.8231, 3, 0.038462, 0.57692)

C. Compare the word attributes of the center point in the clustering results with the word attributes in word attributes updated, find out the most similar words, and bring them into the word difficulty classification table generated in step1, which can be evaluated:

- a) Cluster1 corresponds to hard
- b) Cluster2 corresponds to medium-hard
- c) Cluster3 corresponds to medium
- d) Cluster4 corresponds to easy-medium
- e) Cluster5 corresponds to easy

based on our calculation, the four EERIE word attributes is (8.66, 4, 3, 1), so EERIE belongs to Cluster5, the easy class.

4.3. Accuracy analysis

Comparing the clustered word category with the identified word category, we count the difference between the two by how many gaps there are (e.g. clustered: “easy”; identified: “easy-medium”, then count as 1 gap), average the difference, and thus derive the degree of error.

According to our calculations, the average gap for each valid word given is 0.71, that means these two models we have built are generally interoperable and can roughly determine whether words are difficult or easy, but to continue to break down the difficulty levels and make valid judgments, we need to continue updating and improving our models in future studies.

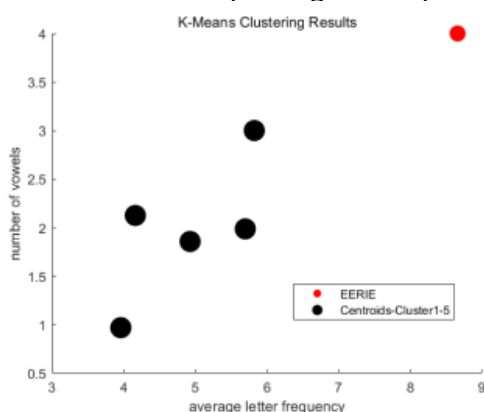


Fig.11 K-means clustering results(a)

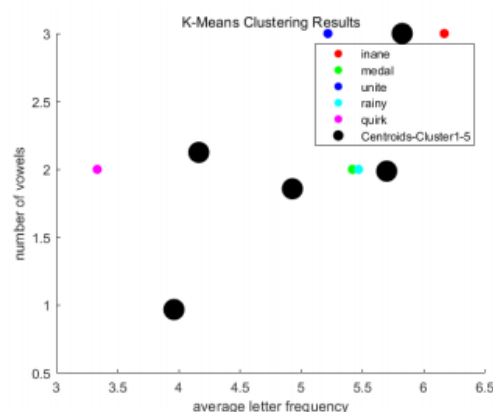


Fig.12 K-means clustering results(b)

5. Some other fascinating traits of the dataset

The quantity of Wordle reports can indicate the size of the word puzzle game’s fan base, and the popular trend of the game over time can be shown. Around January and February, the number of fans increased quickly, peaking on February 2, 2022, with a reported total of 361,908. Following that, the number of reports steadily fell, and the trend was quick at first until becoming full and stabilizing. The majority of those can be regarded as devoted game enthusiasts. The number of people who pick hard mode is consistent at roughly 10%, which is a substantial rise compared to the original 2% when the number of reports in hard mode is divided by the number of all reports. It happens partially

because the Easy setting of Wordle games is no longer challenging for many since they have gotten used to them. However, most individuals merely want to complete levels of fun and relaxation and do not want to spend time playing complex games.

6. Evaluation of Models

6.1. Strengths

A. The model is very simple and only needs endogenous variables without resorting to other exogenous variables.

B. It can effectively compare and evaluate the relationships between multiple attributes and the percentage of reported scores that were achieved in Hard Mode.

C. BP neural network has nonlinear solid mapping ability and fault tolerance.

D. K-means clustering boasts several benefits, such as its algorithm, which is simple to understand and implement, and its interpretability which is highly commendable.

6.2. Weakness

A. In essence, only linear relationships can be captured, but nonlinear relationships cannot be captured. To use the model, we could only assume that the dataset conforms to linear relationships at the stake of high inaccuracy.

B. Due to the limitation of time, we only take two attributes of words into account so we might neglect correlations between the percentage and other possible attributes.

C. The approximation and generalization ability of the network model is closely related to the typicality of the learning samples, which is a complex problem in selecting typical sample instances.

D. One major drawback of this method is the difficulty in determining the appropriate value of K. Our choice of K might be too subjective for this case, which could influence the accuracy of our results remarkably.

References

- [1] Hartigan, J. A., Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1), 100-108.
- [2] Kodinariya, T. M., Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal*, 1(6), 90-95.
- [3] Julong D. Introduction to grey system theory[J]. *The Journal of grey system*, 1989, 1(1): 1-24.
- [4] Brown, K. A. (2022). MODEL, GUESS, CHECK: Wordle as a primer on active learning for materials research. *npj Computational Materials*, 8(1), 97.
- [5] Ding, S., Su, C., Yu, J. (2011). An optimizing BP neural network algorithm based on genetic algorithm. *Artificial intelligence review*, 36, 153-162.
- [6] Pham, D. T., Dimov, S. S., Nguyen, C. D. (2005). Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 219(1), 103-119.
- [7] Ahmed, M., Seraj, R., Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8), 1295.
- [8] Bora, M., Jyoti, D., Gupta, D., Kumar, A. (2014). Effect of different distance measures on the performance of K-means algorithm: an experimental study in Matlab. *arXiv preprint arXiv:1405.7471*.
- [9] Hespanha, J. P. (2004). An efficient matlab algorithm for graph partitioning. University of California, 1-8. Team 2300828 Page 14
- [10] Zhao, Z., Xin, H., Ren, Y., Guo, X. (2010, March). Application and comparison of BP neural network algorithm in MATLAB. In *2010 International Conference on Measuring Technology and Mechatronics Automation (Vol. 1, pp. 590-593)*. IEEE.

- [11] Kim, P. (2017). Matlab deep learning with machine learning, neural networks and artificial intelligence. by Phil Kim.
- [12] Chekalina, V., Bondarenko, A., Biemann, C., Beloucif, M., Logacheva, V., Panchenko, A. (2021, April). Which is better for deep learning: python or MATLAB? Answering comparative questions in natural language. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (pp. 302-311).
- [13] <https://leancrew.com/all-this/2022/01/wordle-letters/>