Judges' Commentary: Predicting Wordle Results

Richard J. Marchand
Mathematics and Statistics
Slippery Rock University,
Slippery Rock, PA
richard.marchand@sru.edu

Introduction and Overview

This is the eighth year that a Problem C has been offered in the MCM®. The primary focus of a Problem C is to incorporate real-world data in the mathematical modeling process in one of two ways, either by applying a mathematical model to a data set or by using the data set to develop a mathematical model [Oliveras et al. 2018]. Like the 2022 Problem C [Olwell 2022], this year's problem required teams to develop predictive models based on time-series data, though the duration of the data was a little less than a year.

An important aspect of any predictive model is an analysis of the uncertainty associated with the model. That was again an important part of this year's problem, with even more specific guidance compared to previous years. Teams that addressed the uncertainties associated with their models well were among the best papers.

The Problem

The popular Wordle puzzle published in the *New York Times* requires players to identify a five-letter word within six attempts. After each attempt, players can identify which of the letters that they guessed are in the word, and which are in the correct location (see New York Times [2023] for a full description of the rules).

Players have the option to play in regular mode or hard mode. In hard mode, once a player determines a correct letter, they are required to use it in subsequent guesses, which makes the game more challenging.

The UMAP Journal 44 (4) (2023) 429–439.

Teams were given a data set derived from Twitter users that listed the daily results reported by players from January 7, 2022 through December 31, 2022. Specific quantities reported included the contest date; word of the day; number of people reporting scores that day; number of players on hard mode; and the percentage of players who guessed the word in one, two, three, four, five, or six tries, or did not solve the puzzle (X).

There were five parts to the problem:

- Develop a model to explain the variation in the daily number of results reported and use the model to create a prediction interval for the number of reported results on March 1, 2023. Determine any word attributes that affect the percentage of scores played in hard mode and how they affect the number of results reported.
- Develop a model that predicts the daily distribution of the percentage of attempts (1, 2, 3, 4, 6, X) on a future date and identify any uncertainties associated with the model and the prediction. Apply the model to the word "EERIE" on March 1, 2023 and determine the level of confidence associated with the prediction.
- Develop and summarize a model to classify solution words by difficulty. Identify the attributes of a given word that are associated with each classification. Use the model to classify the difficulty of the word "EERIE" and discuss the accuracy of the model.
- List and describe other interesting features of the data set.
- Communicate the results in a two-page letter to the Puzzle Editor of the *New York Times*.

Overview of the Judging Process

The judging process has been thoroughly explained in several commentary articles; see, for example, Olwell [2022] and Black [2013]. It begins with triage in which judges strive to identify the best papers to move forward to final judging. Judges generally assessed the executive summary and the letter to the editor to get an overview of the paper and a quick assessment of its quality before reading the rest of the paper. As those two elements form the basis of the judges' first impression, spending extra time on them is worth the investment. Both parts of the paper should provide a concise overview of the problem, the modeling process, and the results. Many papers achieve one or two of these goals, but the highest-rated papers effectively address all three, which is highly regarded by judges. The better teams also strike a good balance between clarity and rigor here, with greater preference given to clarity. The executive summary should be written from the perspective that the reader has not previously read the problem statement—a common element lacking in most papers. Summaries

and letters to the editor that are excessively technical were generally less effective. The letter to the editor must be written appropriately for a non-technical reader.

Judges then apply a rubric to assess how well the team addresses each of the required elements of the problem. Teams missing any required elements generally earn no more than Successful Participant. An important characteristic that judges look for, which is often deficient or incomplete, is a good justification for the modeling approach. This problem is particularly exacerbated by the proliferation of software tools that make it easy to apply predictive models with little consideration for why they were used. Teams that did so, or created multiple models indiscriminately, did not score well.

Data Preprocessing

As is often the case in a Problem C, there are several errors in the data set that must be addressed prior to model building. The errors in this year's data set included several misspelled words and at least one inaccurate user report that was off by an order of magnitude. The better papers acknowledged these errors. Although many teams used some form of imputation strategy to address them, the best teams simply corrected them—since the data could be verified online.

Predicting Results

Many teams employed a variety of sound strategies for providing a point estimate of the number of Wordle users on March 1, 2023, primarily relying on some form of regression, ARIMA, or machine learning model. However, an important requirement of any prediction interval is to provide an associated confidence level or some other robust measure of uncertainty to assess the quality of the estimate. As in previous years, this appeared to be the greatest challenge associated with this problem [Olwell 2022]. That challenge was particularly difficult for teams that exclusively used certain machine learning algorithms for which a prediction interval and confidence level were not inherently included in the algorithm.

Many teams additionally confused a prediction interval with a confidence interval or resorted to ad hoc methods for creating some sort of interval estimate for the number of users. The best papers provided well-defined prediction intervals and attached an effective measure of the uncertainty or level of confidence associated with it. Team 2310767 from Tianjin University, China provided a nice visualization of the prediction bands, as shown in **Figure 1**. Team 2318982 from East China Normal University

also did a nice job of justifying a point estimate for the reported number of users, along with a prediction interval and associated confidence level. The completeness of the analysis in those two papers was a contributing factor to their rank as Outstanding.

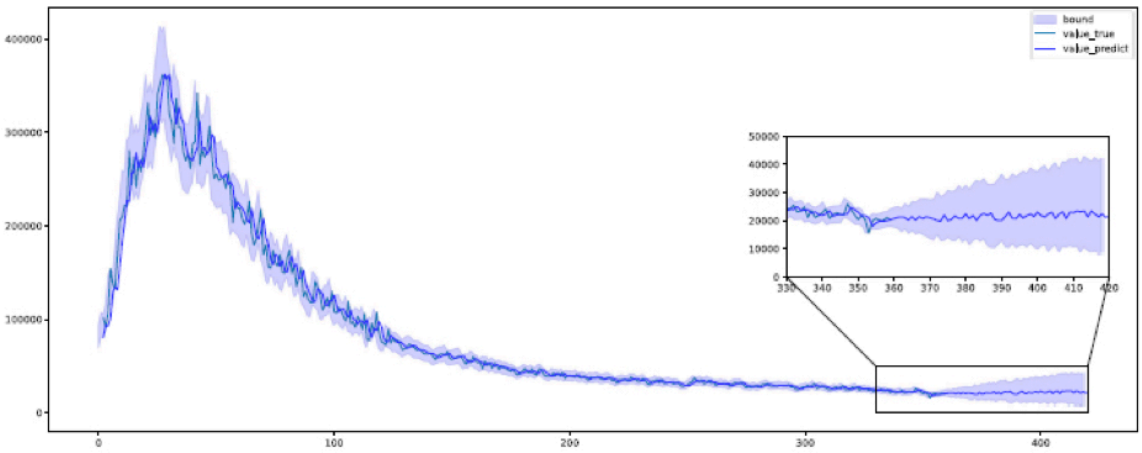

Figure 1. Prediction interval bands by Team 2310767 Tianjin University, China.

Another aspect of this part of the problem was to identify any attributes that may affect the percentage of scores played in hard mode. Teams were required to identify specific attributes of words that would affect the number of scores reported in hard mode. Most teams conducted some type of correlation analysis between various word attributes (e.g., commonness, number of repeated letters, etc.) and the number of hard-mode users, with the best papers clearly justifying their conclusions computationally and providing graphical visualizations including correlation matrices, heat maps, etc. Team 2307946 from Shanghai University of Finance and Economics, China showed that the most important attribute affecting the number of players in hard mode had more to do with the difficulty of the words in the few days prior to any given day. Team 2314151 from Northeastern University of China provided a neat heat map, as shown in **Figure 2** to illustrate their results.

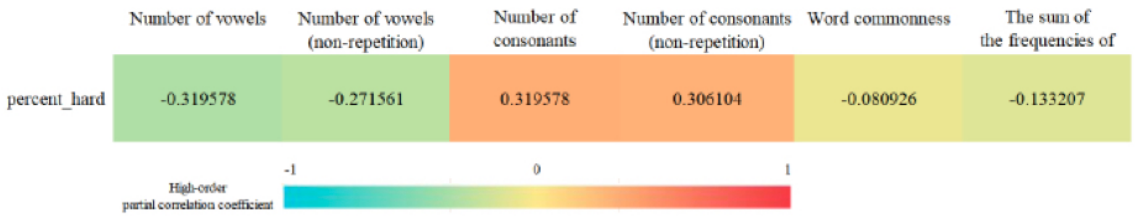

Figure 2. Heat map of word attributes and players in hard mode, from Team 2314151 from Northeastern University of China.

Another effective strategy was to employ some type of regression model with the attributes represented by variables and assessing the statistical significance of regression coefficients. This was done particularly well

by Team 2311035 from the Chinese University of Hong Kong, Shenzhen, China and Team 2309397 from Nanjing University of Posts and Telecommunications, China, the latter providing an excellent interpretation of the coefficients in their model, a rarity that resonates very well with judges.

Distribution of Reported Results

Perhaps the most challenging part of this problem was predicting the distribution of attempts at a future date. The ideal result, as is the case with any predictive model, is to provide a prediction interval with an associated confidence level or other robust measure of uncertainty. A variety of methods were used to create predictive models using regression and machine learning techniques, but most teams struggled to analyze effectively the uncertainties in the model and state an explicit confidence level associated with the predictions. Logical and thorough analysis of the uncertainties and confidence levels associated with the predictions separated the better papers from the others. The more successful teams also achieved better results for goodness of fit to justify their models to some degree. Team 2322645 from Columbia University, NY, USA used a novel Bayesian approach that led to an exceptional presentation of the prediction intervals for the distribution of the number of attempts and the associated confidence levels for the word "EERIE" on March 1, 2023 as show in **Table 1**. They referred to the confidence levels as variables as indicated in the table.

Table 1. Distributions of predictions by Team 2322645 from Columbia University, NY, USA.

Variable	95%	80%	50%	Median
Number of reports	[20,238, 27,876]	[21,479, 26,365]	[22,622, 25,169]	23,884
Number of hard mode	[2,194, 3,239]	[2,355, 3,048]	[2,509, 2,870]	2,683
Percentage of hard-mode	[9.97, 12.62]	[10.41, 12.15]	[10.79, 11.72]	11.25
Percentage in 1 guess	[0, 2.4]	[0, 0.82]	[0, 0.15]	0
Percentage in 2 guesses	[1.09, 14.01]	[2.08, 10.45]	[3.39, 7.70]	5.23
Percentage in 3 guesses	[12.16, 35,85]	[15.34, 31.03]	[18.49, 26.82]	22.46
Percentage in 4 guesses	[21.29, 48.16]	[25.19, 43.2]	[29.2, 38.51]	33.79
Percentage in 5 guesses	[12.48, 37.09]	[16.16, 32.19]	[19.4, 27.96]	23.54
Percentage in 6 guesses	[3.73, 21.33]	[5.6, 16.99]	[7.6, 13.53]	10.37
Percentage failed	[0.06, 7.77]	[0.24, 4.91]	[0.66 3.09]	1.6

What was particularly impressive and effective about their work was that they tested their model on a word, "FUNGI", for which the results were known. Such a strategy of validating a predictive model is generally considered a best practice when assessing mathematical models yet scant few teams did so. The team effectively represented the results of the test using a good visualization as shown in **Figure 3**.

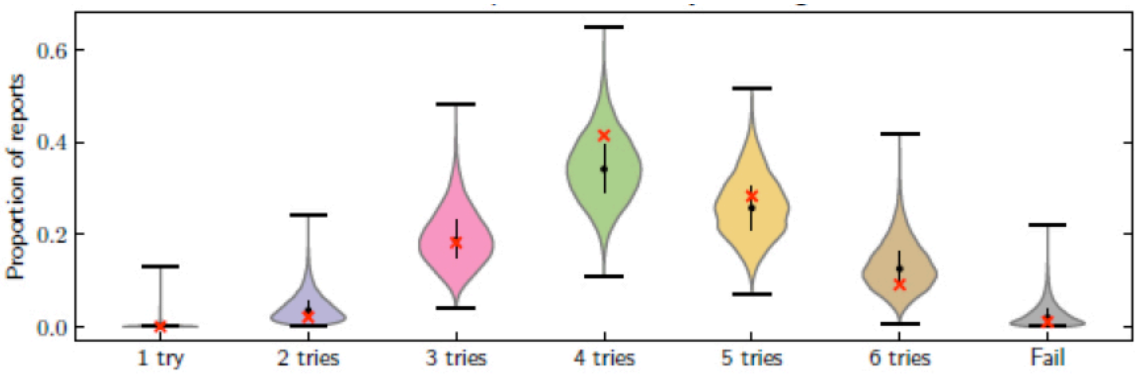

Figure 3. Test of the model for predicting the distribution of attempts on the word fungi by Team 2322645 from Columbia University, NY, USA. The **x** marks the estimated proportion of the number of tries and the black dot • indicates the median of the interval.

Comparison with Actual Data

Figure 4 shows the actual statistics for the Wordle of March 1, 2023, which turned out to be not "EERIE" but "MOOSE."

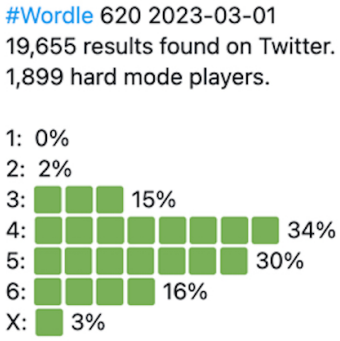

Figure 4. Statistics for on March 1, 2023 for the Wordle word "MOOSE."

A common assumption among many teams was that Wordle players play "fairly." That assumption was explicitly called into question by Team 2318982 from East China Normal University, which noticed that some words had an unusually-high percentage of players correctly identifying the target word on the first attempt. Their suspicion was supported in a recent playful paper positing rampant cheating among Wordle players [Dilger 2023] and summarized in The Physics arXiv Blog [2023] of *Discover Magazine* and by Knight [2023]. Knight notes:

> Of the roughly 2 million daily Wordle players, about 860 people should be guessing the right word on their first try. If they avoid any previous daily words (and most don't) that might rise to 1,320. And given that many players remain loyal to the same opening word, or cycle among several, that number should be even smaller. Instead, daily first-word-winners number between 4,000 and 10,000.

A respondent to the Knight article, Walter Legault, remarked:

"What is the point of cheating in a solo game?"

Difficulty-Ranking Model

In this section, teams had to make justifiable decisions about how to classify the difficulty of a word. The better teams identified specific word attributes (such as the types of letters in the word, the number of vowels, repeated letters, commonness of the word, parts of speech, etc.) to reveal correlations with the distribution of the number of attempts listed in the data. By far, the most common approach used here was to use some type of clustering algorithm coupled with correlation investigations. What separated the Outstanding teams from the others was a clear explanation of the clustering process and the interpretation of their results. The better teams also supported their conclusions with specific examples that could be verified through the data accompanied by effective visualizations. The visualization of Team 2311717 from Xidian University, China (**Figure 5**) was an excellent example of model validation.

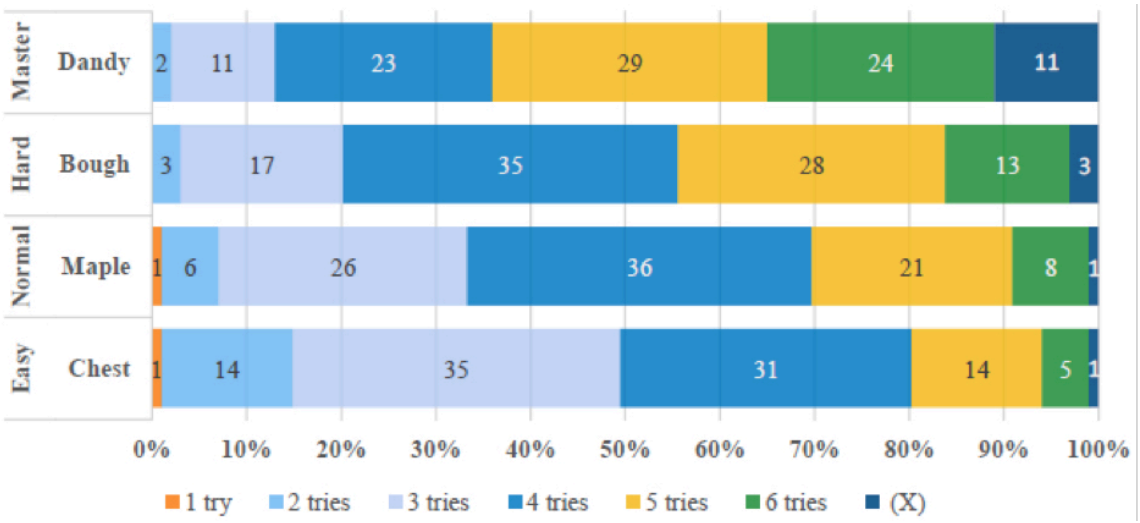

Figure 5. Visualization of difficulty-ranking model validation used by Team 2311717 from Xidian University, China, with number of tries increasing from left to right.

Another common approach was to use principal component analysis (PCA) to reduce the dimensionality of the variables used in the clustering. Team 2301192 from Zhejiang University of Finance and Economics, China did a good job of justifying their use of this strategy and the resulting visualization in **Figure 6** supported their results well.

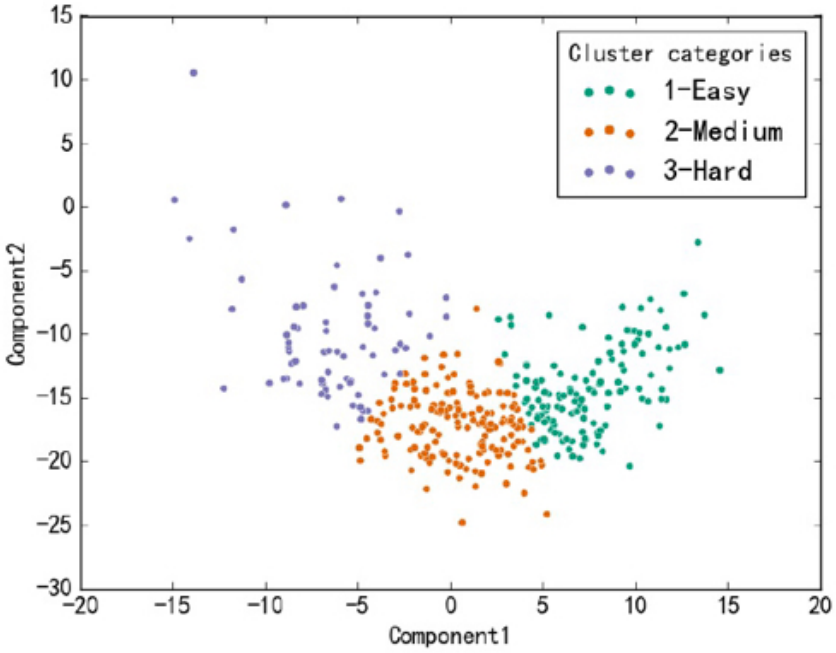

Figure 6. Clustering PCA components for the difficulty-ranking model used by Team 2301192 from Zhejiang University of Finance and Economics, China. The easy words cluster on the right, the medium ones in the middle, and the hard ones on the left.

Other Interesting Features

In general, most teams were able to identify several interesting features of the data motivated by insights acquired while completing other problem requirements. This part of the problem was an opportunity for teams to showcase their creativity and test conjectures. Typical observations originated from

- analysis of the most difficult word, "PARER";
- the most common starting letter of the solution;
- most common parts of speech used;
- correlation between the day of the week and the number of results reported,
- seasonal effects, and
- many others.

The better papers not only made astute observations but also provided clear explanations of the analysis and justifications for their results. Team 2307166 from Renmin University of China used a correlation matrix to effectively support several conclusions about relationships between word features and the distribution of attempts and other variables.

Sensitivity Analysis

The sensitivity of a model is a good indicator of model performance, so it is an important feature of any modeling effort. It is also a good opportunity to assess the robustness of a model to assumptions. How well a team completes this requirement is often a good discriminator for the best papers.

Any parameter that is estimated in a model is a good target for sensitivity analysis. Another possible target is the input variables themselves. Teams that used some kind of weighting system for input variables in their model without testing the weights generally scored poorly.

Team 2300348 from the University of Electronic Science and Technology of China made excellent use of a radar chart to assess the sensitivity of their predictive model to changes in word and letter frequencies, which facilitated the interpretation of their results. Team 2318036 from Renmin University of China assessed the sensitivity of their model to changes in the input variables. The analysis allowed them to test the stability of their model. The attention to detail and communication of results supported by appropriate data visualizations certainly contributed to their Outstanding rankings.

Strengths and Weaknesses

Another important requirement of the modeling process is to critically assess the performance of models while identifying limitations and how they could be mitigated. Judges clearly recognize the time limitations imposed on teams, so it is important for each paper to identify what could have been done differently if given more time. While some teams merely list the strengths and weaknesses of their models, the best papers justify their conclusions and provide substantive insight into how their models could be improved. The better teams also demonstrate an understanding of the relative importance of the limitations of their models.

Clarity of Writing

A well-written report that sufficiently explains a team's work is an important discriminator throughout all levels of the judging process. Even the best models, if poorly communicated or poorly organized, will find it difficult to earn the highest ranking. Judges routinely note a lack of sufficient detail among papers that makes it challenging for the judges to understand the rationale of the authors. Appropriate citations that support

justifications, conjectures, or conclusions must be included. Other common errors observed by judges are a lack of appropriate labels and titles for graphics, and data visualizations that do not clearly convey what they represent.

Judges are always looking for good interpretations of a team's model. A common example is interpreting the coefficients of a regression model within the context of the problem. The coefficients can be used to say a lot about the model that teams often fail to mention. Another classic example is principal component analysis, a common data reduction tool. It is rare to see teams attempt to interpret the resulting components in context. It is important to attach meaning to the components to adequately assess the quality of the modeling results.

Outstanding Papers

All Outstanding papers have the following qualities:

- They address all problem requirements.
- They effectively account for uncertainties.
- They achieve plausible results that are tested and communicated well.
- They leverage appropriate data visualizations to adequately support interpretations and conclusions.
- They clearly identify and analyze strengths and weaknesses of their models.

Conclusion

Problem C continues to be a very popular problem due to its accessibility to participants. It also provides a framework in which an incredibly broad set of legitimate modeling strategies, tools, and techniques may be applied.

References

Black, Kelly. 2013. Judges' commentary: The Ultimate Brownie Pan papers. *The UMAP Journal* 34 (2–3): 141–149.

Dilger, James P. 2023. Wordle: A microcosm of life. Luck, skill, cheating, loyalty, and influence! https://arxiv.org/abs/2309.02110.

Knight, Chris. 2023. Mathematician says thousands of Wordle players are cheating every day. https://ottawacitizen.com/news/wordle-cheating-claims/wcm/5cb025d8-979a-407a-b2af-96d80903ad2f.

New York Times. 2023. Wordle—The New York Times. https://www.nytimes.com/games/wordle/index.html.

Oliveras, Katie, Stacey Hancock, and David H. Olwell. 2018. Judges' commentary: The Southwest states' energy compact. *The UMAP Journal* 39 (3): 343–350.

Olwell, David H. 2022. Judges' commentary: The bitcoin and gold portfolio problem. *The UMAP Journal* 43 (4): 455–460.

The Physics arXiv Blog. 2023. Data analysis [sic] reveal surprisingly high number of Wordle cheaters. https://www.discovermagazine.com/technology/data-analysis-reveals-surprisingly-high-number-of-wordle-cheaters.

About the Author

Richard J. Marchand is a Professor and Chairman of the Dept. of Mathematics and Statistics and Assistant to the Dean for the College of Engineering and Science at Slippery Rock University. He earned a B.S. in mathematics from Clarion University and a Ph.D. in applied mathemati cs from the University of Virginia. He previously completed a civilian post-doc at the US Military Academy and a Distinguished Visiting Professorship at the US Air Force Academy after serving on the faculty of the State University of New York at Fredonia. Dr. Marchand has been an MCM judge for more than 15 years.