Judges' Commentary: The Wealth of Data Problem

Stacey Hancock
Mathematical Sciences
Montana State University
Bozeman, MT
stacey.hancock@montana.edu

Gregory Mislick
Operations Research
Naval Postgraduate School
Monterey, CA
gkmislic@nps.edu

David H. Olwell
Professor and Dean
Hal and Inge Marcus School of Engineering
Saint Martin's University
Lacey, WA
dolwell@stmartin.edu

Introduction and Overview

This is the fifth year that the MCM has offered a Problem C, which focuses on mathematical modeling based on real-world data. As mentioned in the 2018 Problem C commentary [Oliveras et al. 2018], the problems continue to be of two types: those developed from a large data set, and those applied to a data set. This year's problem was the former.

We continue to expect students to analyze and discuss the uncertainties of their results in the context of the uncertainties inherent in both the data and the modeling. We will continue to stress this emphasis in future problems. This analysis of uncertainty is what distinguishes Problem C from Problems A and B.

The growth in Problem C continues. While in 2018 we had 4,747 teams and 12,707 students, this year we had 7,446 teams and 22,130 students from 12 countries. That is 75% growth in just two years.

The UMAP Journal 41 (4) (2020) 447–459.

Six papers were selected as Outstanding. There were 64 Finalist papers, 484 Meritorious papers, and 1882 Honorable Mentions. The remaining papers (about 66%) were judged to be Successful Participants. Fewer than 100 papers were Disqualified (usually for plagiarism issues) or rated Unsuccessful (usually for not submitting anything meaningful to be judged).

The Problem

Three large datasets web-scraped from the online marketplace of Amazon were provided to the teams. The files included

- "star ratings,"
- text-based reviews with titles,
- helpfulness ratings of those reviews by other customers,
- date of the review, and
- other data about the purchaser and reviewer,

for three kinds of products: hair dryers, microwaves, and pacifiers. The addition of text fields in the data was new this year.

Teams were asked to look at the data from the perspective of consultants assisting with new product development by identifying key patterns, relationships, measures, and parameters in past customer-supplied ratings and reviews associated with other competing products to

1. inform their online sales strategy and
2. identify potentially important design features that would enhance product desirability.

Sunshine Company has used data to inform sales strategies in the past, but they have not previously used this particular combination and type of data. Of particular interest to Sunshine Company are time-based patterns in these data, and whether they interact in ways that will help the company craft successful products.

Teams were explicitly required to analyze the three data sets to find meaningful patterns, relationships, measures, and parameters between the variables and the three data sets, to help the client succeed in their new product offerings. There were five questions from the marketing director to answer. Teams were then required to write a letter to the Marketing Director summarizing their analysis and results.

The five questions were:

- Identify data measures based on ratings and reviews that are most informative for Sunshine Company to track, once their three products are placed on sale in the online marketplace.

- Identify and discuss time-based measures and patterns within each data set that might suggest that a product's reputation is increasing or decreasing in the online marketplace.
- Determine combinations of text-based measure(s) and ratings-based measures that best indicate a potentially successful or failing product.
- Do specific star ratings incite more reviews? For example, are customers more likely to write some type of review after seeing a series of low star ratings?
- Are specific quality descriptors of text-based reviews such as "enthusiastic," "disappointed," and others, strongly associated with rating levels?

Overview and Triage

The judging process relies heavily on a triage phase, in which each paper is read by two experienced judges. Those judges give an initial assessment, and those assessments are combined to determine which papers move further in the judging process. It is very important for student teams to understand what the judges weigh in this triage process.

- Judges first assess whether the paper addresses the required elements of the modeling process, and whether a team has answered the questions posed in the problem statement. Most judges begin by reading the abstract and the memorandum to get an overall sense of the paper before reviewing the main body. It is critical that both of these components be well-written crisp summaries of the results. The burden is on the student team to explain what they are doing and to summarize results for a nontechnical audience with recommendations and justifications in their letter to the marketing director. If a judge cannot follow what they say in the letter that they have done, then the marketing director would not be able to understand it, either.
- Judges then use a rubric to assess how well a team addresses each of the required elements. Teams that omit a required element usually score no higher than Successful Participant.

In the next section, we discuss the most common shortfalls observed by judges.

Common Issues

What Was the Goal of the Model?

Participants were asked to perform an in-depth data analysis of Amazon customer reviews using data concerning three new products from a large database provided. The goal of this analysis was to devise actionable recommendations (with justifications) to the Sunshine Company Marketing Director. Sunshine Company is planning to introduce and sell these three new products—a microwave oven, a baby pacifier, and a hair dryer—and is seeking advice on which product features would give them an edge over their competitors.

Each team was considered a consulting firm and asked to analyze data from customer ratings and reviews on similar competing products to help Sunshine Company

- establish effective advertising and product improvement strategies to maximize its product sales and profit, and
- identify those design features that most enhanced their product's desirability.

Essential to this analysis was a thorough understanding of time-based patterns in the data that could be used to maximize sales. Participants were provided with the complete data set for the study and no other external data sources were allowed. At the conclusion of the study, each team was required to write a letter to the Sunshine Company Marketing Director describing their analysis and discussing the recommendations that would most improve the chance of success with these three new products. Recommendations needed to be helpful, supported by the data analysis, and actionable.

How to Address Text-Based Analyses?

A random variable is a real number assigned to the outcome of an experiment. This year, we included a significant amount of textual data for students to analyze. How they decided to create random variables from the textual data was a new challenge for most students.

There are a number of software packages available to analyze textual data, with a significant number available for free in R. These packages can do sentiment analysis, latent Dirichlet allocation, and a number of other advanced techniques from machine learning. The judges did not expect that a team should use an advanced software approach such as these, although many did. The judges did expect that a teams use some modeling method to assign a numerical score or scores to the text data and incorporate those scores further into their modeling.

Successful teams filtered the quality of the text reviews, to remove "score robots" and potential competing companies intentionally providing poor reviews to sway public opinion. These "fake reviews" were discovered when a large number of unverified reviewers gave overwhelmingly poor evaluations.

Teams found associations between the text data and other features of the review. For example, they found that negative reviews tended to be longer and advised the director to monitor them. Good teams analyzed the star ratings and looked at ratings over time to determine various patterns in the data. The analysis of keywords for both one-star and five-star ratings was essential and enabled the best teams to determine desirable features as well as defects that should be fixed. Some teams considered words in "ALL CAPS" to be the most important and filtered out duplicate words. Many teams found success when they compared the average star value and their text review score. Others determined that the title of the review was more informative about how the customer felt about the product than the review itself.

Some teams tried various machine learning algorithms to find the best algorithm for extracting the most useful comment words. Other teams divided text emotions into three main categories: neutral, positive, or negative. Still others accounted for text readability, counted the number of words in a comment, and performed time-series analysis on reputation trends. And some weighted reviews from the customers in Amazon's Vine Voices program more heavily.

The judges were not looking for any particular method to handle textual data, but they did expect a method to be used and explained well, and for it to prove useful for further modeling in the rest of the paper.

Data Visualization

As in previous years, data visualization played an important role in uncovering relationships in the data. For many of the problem requirements, simple statistical plots could immediately illuminate the required relationships. For example, to examine the relationship between star ratings and helpfulness ratings, one could turn to side-by-side boxplots or pirate plots[1].

Boxplots are a simple and helpful tool for comparing the distribution of a quantitative variable across groups. Successful teams were adept at using such plots to explore relationships in the data. For example, team 2004156 from Xiamen University (Xiamen, Fujian, China) effectively used

[1] A pirate plot has jittered data, a vertical density shape showing distribution, and a rectangle representing an inference interval. For rationale, examples, and code in R, see Phillips [2016; 2017a; 2017b] and Weston and Yee [2017]. Phillips coined the term, based on his whimsical claim of finding an introduction to R written in pirate-speak, in which a major example dataset is from a survey of 1,000 pirates.

boxplots to compare the distribution of star ratings across high-frequency words in the reviews. After calculating a "review score," team 2012671 from Tsinghua University (Beijing, China) used boxplots to compare the distribution of this score across the five-star ratings. Though more informative than boxplots, pirate plots were rarely seen.

In this year's contest, the added component of text as data gave teams the opportunity to explore creative data visualizations. Yet many of the visualizations that humans tend to think are the "prettiest" are not often the ones that are most informative. For example, the most popular data visualization seen in this year's entries for product reviews was the word cloud. **Figure 1** shows a word cloud of the 100 most common words used in product reviews of hair dryers (with common words such as "a" and "the," as well as the product terms "hair" and "dryer," removed).

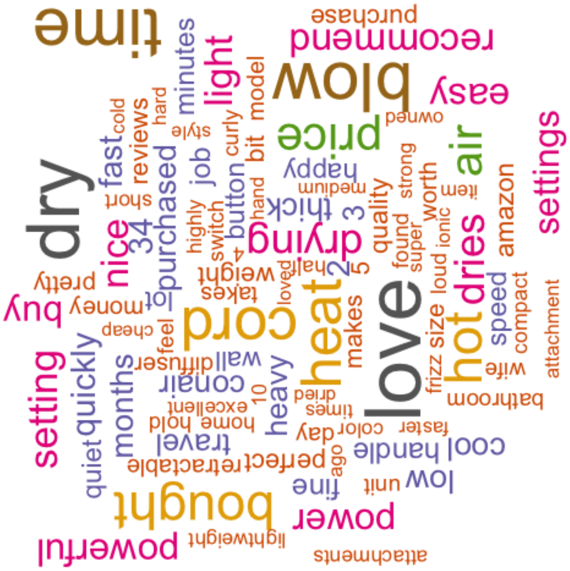

Figure 1. Word cloud of the 100 most-frequent words used in hair dryer reviews.

While the graphic is colorful and pretty to look at, it is difficult to gain from it accurate information about the distribution of these words. In a word cloud, the font size of the word is proportional to its frequency—how often it occurred. However, the length of the word also catches the human eye. Thus, words that are longer but with smaller font may appear larger than shorter words with larger font. Additionally, the human eye is generally terrible at comparing sizes of two words, let alone comparing a cloud of many words.

Another popular visualization seen in entries was the tree map (see **Figure 2**).

A tree map plots the most frequent words in the top left, and the least frequent in the bottom right. The area of the rectangle is proportional to the frequency of the word. While comparing areas of rectangles is slightly easier than comparing font sizes of words, one would still be hard pressed

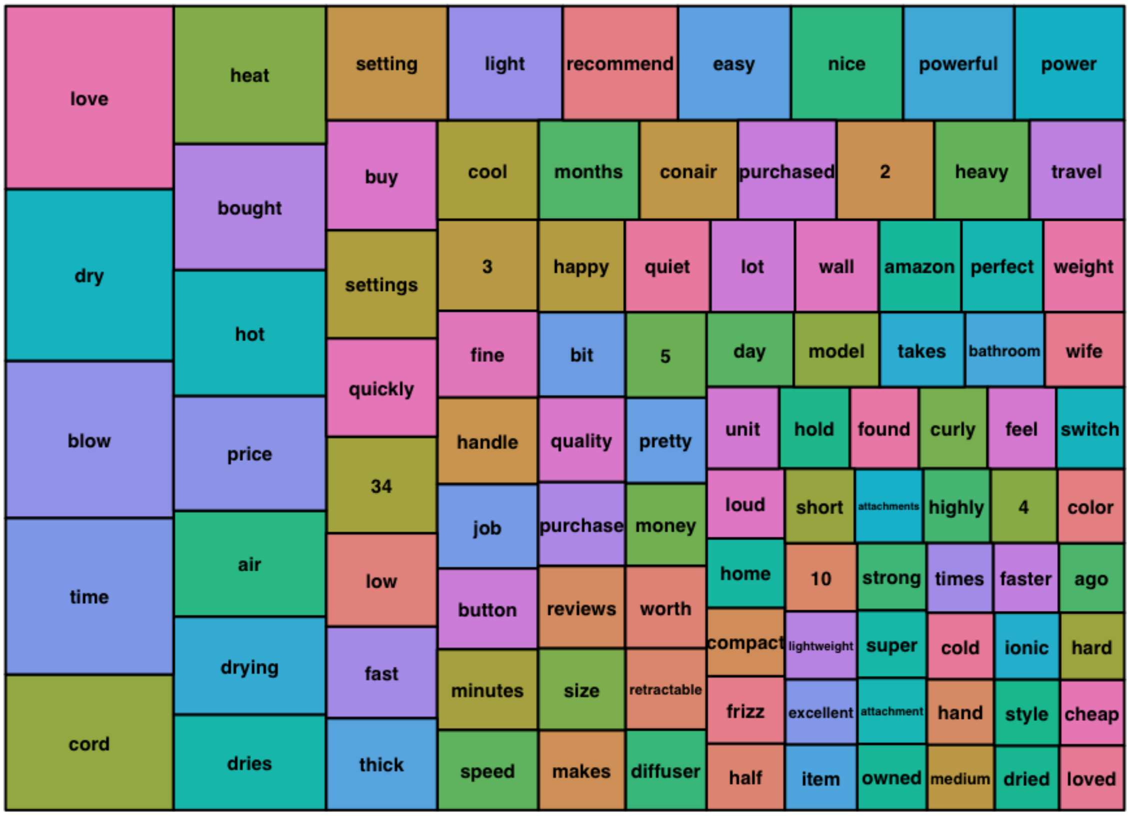

Figure 2. Tree map of the top 100 most-frequent words used in hair dryer reviews.

to determine if, for example, "love" or "dry" had a higher frequency in hair dryer reviews.

William S. Cleveland, one of the leaders in data visualization, defined "elementary perceptual tasks" [1984]. An elementary perceptual task is how people extract quantitative information from graphs. Cleveland lays out a hierarchy of elementary perceptual tasks, from easiest to hardest. The easiest is to compare the position of two points on a common scale, such as comparing the heights of two bars on a bar chart. Comparison tasks that are harder include comparing areas, volumes, angles, and shading. Unfortunately, many of these harder elementary perceptual tasks are the basis of some of the most popular data visualizations (case in point: pie charts). If the goal of a data visualization is to convey quantitative information, then it is best to adhere to the easier tasks in Cleveland's hierarchy.

Returning to our example, **Figure 3** shows a simple bar graph of the 20 most-frequent words used in hair dryer reviews. Since our eyes are accustomed to comparing positions on a common scale, the ease with which this plot conveys information is immediately apparent. More importantly, we can use this plot to compare different types of reviews.

Figure 4 displays a bar graph of the top twenty most frequently occurring words used in "poor" (star rating less than 3) hair dryer reviews, and the same for "good" (star rating greater than 3 reviews). Strong papers

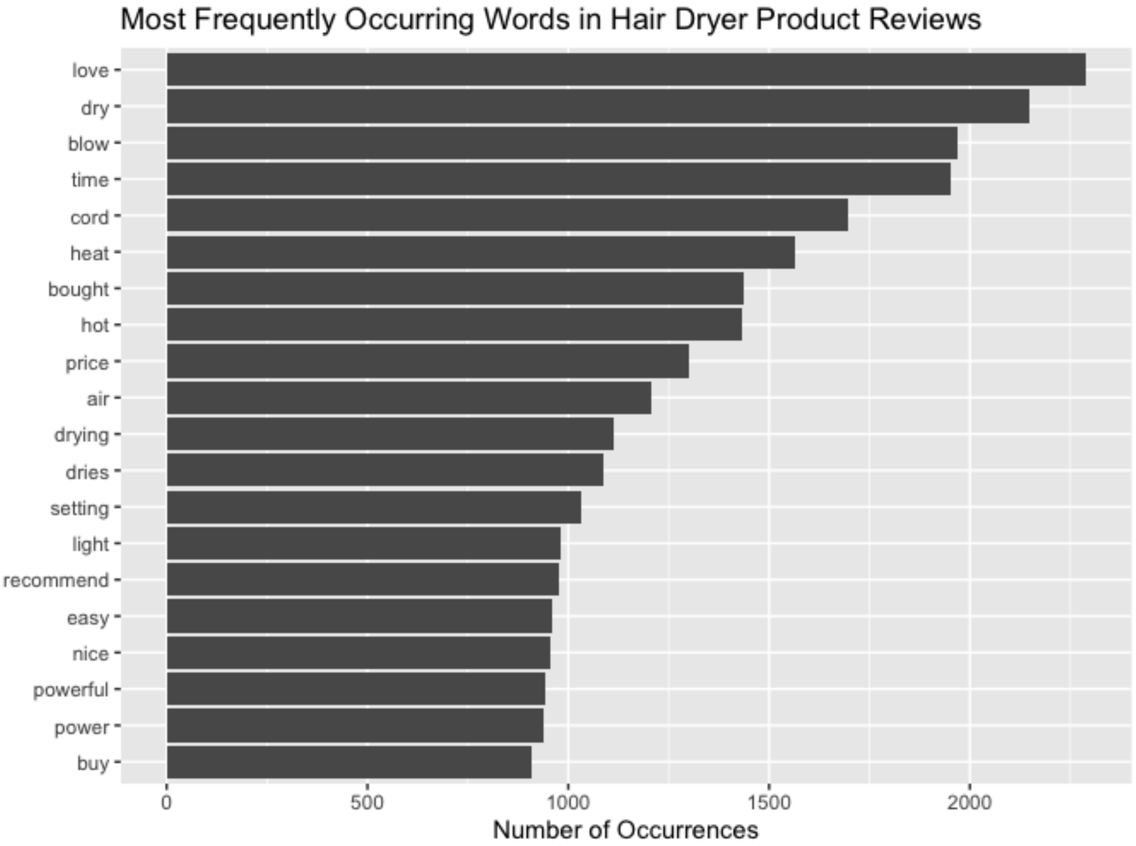

Figure 3. Bar plot of the 20 most-frequent words used in hair dryer reviews.

used data visualization methods that adhered to Cleveland's hierarchy, such as these bar graphs.

Handling Uncertainty

A central part of any statistical analysis is quantifying uncertainty—both in the data and in the modeling process. Any prediction is useless without an accompanying measure of uncertainty. As in previous years, many papers left this crucial component out of their analysis completely. After identifying time-based data measures that are most informative for Sunshine Company to track, successful teams produced time series models that forecast those measures into the future. The best papers included prediction bands with those forecasts, to give Sunshine Company an idea of the range of plausible outcomes that they might observe.

Handling uncertainty is not limited to prediction of quantitative variables. When predicting a binary outcome, such as whether a review is helpful, measuring uncertainty is again important. Team 2012671 from Tsinghua University (Beijing, China), for example, used the area under the curve (AUC) of a ROC (receiver operating characteristic) curve to quantify the uncertainty in such predictions. The resulting AUCs of between

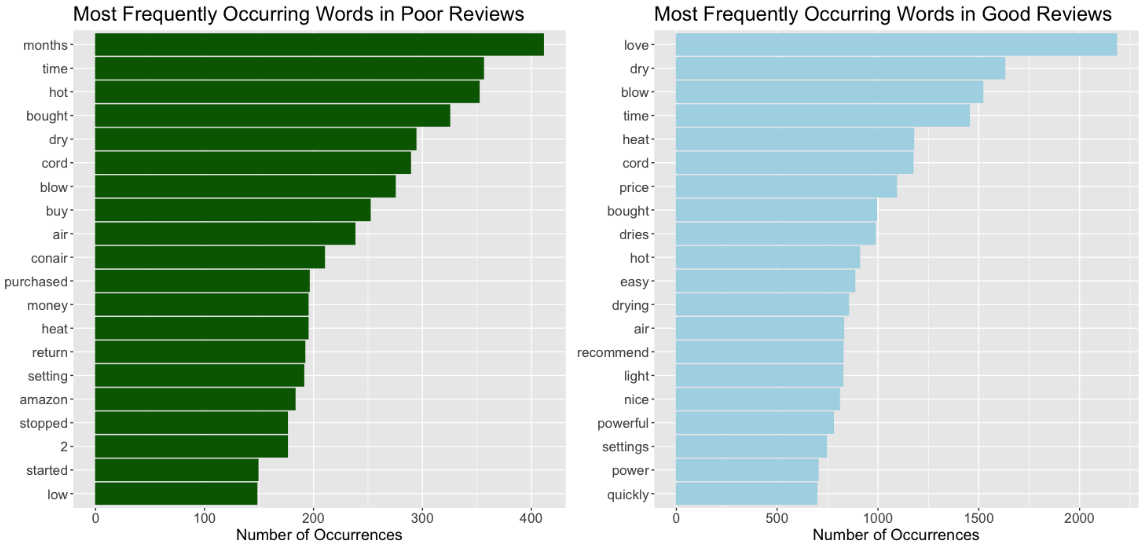

Figure 4. Bar plots of the top 20 most frequent words used in poor hair dryer reviews (left) and good hair dryer reviews (right).

64% and 69% helped communicate that, while somewhat predictive, the star rating, emotional score, and length of a review together are not strong predictors of whether the review is voted helpful by other customers.

Missing Required Elements

The main areas that needed to be included in the paper were:

- Restatement of the Problem. This area should possess three things:
  - A description of the problem
  - An overview of the team's approach
  - An overview of the team's results
- Assumptions and Justifications. There should be at least five of them to be at all effective. A good paper will interweave the assumptions throughout their paper and show how the assumptions were influencing decisions. Weak or minimal assumptions were a big detractor.
- Model construction and application. These include the analysis itself and was the bulk of the paper. A long description of what was attempted and accomplished needed to be well-described. A time-series analysis of the data was also required. Sometimes there was a complex analysis provided but then unfortunately they did not give any results for each method.
- Model Testing
- Sensitivity Analysis and Modeling Uncertainty
  - Was the model sensitive or not?

  - Did it add any value to the paper?
  - What did the results show?
  - How precise did the model predict future behavior?

- Analysis of Strengths and Weaknesses
- Overall Summary and Conclusions
- Letter to the Marketing Director of Sunshine Company for their three products. The letter needed the following attributes:
  - To be clear and offer helpful and viable conclusions for their products, and
  - To provide recommendations that were actionable to the Marketing Director

Points were deducted when areas were not included. The areas that were omitted the most were the Sensitivity Analysis section as well as the Strengths and Weaknesses. Many of the poor papers had very weak assumptions or just very few. Some of the papers provided conclusions that were long, but they described more about their efforts and how they accomplished their analysis than what the conclusions were.

Papers that omitted required elements were seldom rated above Successful Participant.

Writing Crisp Recommendations Grounded in Model Results

Organization and flow of the paper was essential, as was clear writing. Good graphs highlighting important data visualization trends and flow charts helped support the team findings.

As previously mentioned, the letter to the Marketing Director of Sunshine Company concerning the three products needed to reflect the following attributes:

- To be clear and offer helpful and viable conclusions for their products, and
- To provide recommendations that were actionable to the Marketing Director

Recommendations needed to discuss ways to improve product features and reveal which ones were the most important. The product features highlighted in the letter should have been extracted from repeated reviewer comments, which were both positive and negative. Positive comments reinforced the features that were most desirable and would most help increase sales overall, while negative comments revealed weak product features that needed attention and improving. These recommendations needed to be supported by modeling that forecasted sales out for five years.

Characteristics of Outstanding Papers

In addition to satisfying all of the required elements, applying appropriate modeling techniques, and communicating clearly, a few other qualities separated Outstanding papers from the rest.

Though many papers used appropriate and effective modeling methods, only the strongest papers described and explained these methods, along with their strengths and weaknesses, in a way that a reader unfamiliar with the methods could follow. Complete descriptions of what was attempted and why it was appropriate, how a method was applied, and the results of the analysis were present in every Outstanding paper.

Outstanding papers wrote clear and helpful letters to the Marketing Director of Sunshine Company, with actionable recommendations supported by the data for both product design and sales strategies. The Outstanding team from Southwestern University of Finance and Economics (Chengdu, Sichuan, China), the winner of the INFORMS Award, provided such detailed recommendations for both design and sales. For example, based on text analysis of negative reviews, this team recommended that the hair dryer design focus on efficiency (drying hair in a short amount of time), light weight, and low noise. Sales strategies included times of year when sales tended to peak, suggesting that the company focus advertising during those times.

Winning papers augmented clear communication of their analysis strategy and results with flow diagrams of the process and illuminating data visualizations. Such graphics allow the reader to get an overview of the analysis and results by looking at the figures alone, without reading the paper in detail.

As mentioned previously, quantification of uncertainty in the data and model predictions was expected but not often witnessed by the judges. Outstanding papers included this essential element in their analyses. The Outstanding team from South China University of Technology (Guangzhou, China), the winner of both the AMS Award and the COMAP Scholarship Award, for example, included 95% confidence bands when predicting a product's reputation over time, and displayed these confidence bands clearly in plots of the predictions.

Summary

Problem C continues to evolve as a data modeling and analysis challenge. Teams that do well on Problem C have a competitive advantage in the marketplace, where these skills are in great demand from employers.

As always, papers that answer all parts of the question, explain the models used, include appropriate graphics, and present a good summary scored well this year. Papers that explicitly addressed uncertainty did very well.

The judges were very pleased by the range of submissions this year. We look forward to next year's contest.

References

Cleveland, W.S. 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association* 79 (387): 531–554.

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

Phillips, Nathaniel D. 2016. The pirate plot: `pirateplot()`. In *YaRrr! The Pirate's Guide to R*, 138–142. https://drive.google.com/file/d/0B4udF24Yxab0S1hnZlBBTmgzM3M/view?usp=sharing and http://rpository.com/down/yarr/YaRrr_ch09+10.pdf.

________. 2017a. *YaRrr! The Pirate's Guide to R*, Chapter 11.6: `pirateplot()`. https://bookdown.org/ndphillips/YaRrr/pirateplot.html.

________. 2017b. `pirateplot`. https://cran.r-project.org/web/packages/yarrr/vignettes/pirateplot.html.

Weston, Sara, and Debbie Yee. 2017. Pirate plots in R: Plotting raw data, description, & inference. https://dmyee.files.wordpress.com/2016/03/pirateplots_workshop.pdf.

About the Authors

Stacey Hancock is an Assistant Professor of Statistics at Montana State University. She earned her B.A. in Mathematics and Music at Concordia College in Moorhead, MN, M.S. in Statistics at Montana State University, and Ph.D. in Statistics from Colorado State University. Her primary research interests lie in statistics and data-science education, and she serves as the American Statistical Association judge for the MCM competition.

Greg Mislick is a Senior Lecturer at the Naval Postgraduate School (NPS) in Monterey, CA. He earned his B.S. in Mathematics from the US Naval Academy and his primary master's degree in Operations Research from NPS. He served for 26 years as a helicopter pilot in the US Marine Corps. His research interests lie in statistics, optimization and cost estimation. He authored the textbook *Cost Estimation: Methods and Tools* (Wiley, 2015). He has been a judge in the MCM for more than 10 years.

David H. Olwell is Professor and Dean at the Hal and Inge Marcus School of Engineering of Saint Martin's University. He earned a B.S. at the US Military Academy, where he studied mathematics, and an M.S. in Mathematics, an M.S. in Statistics, and a Ph.D. in Statistics from the University of Minnesota. He has previously been on the faculty of the US Military Academy and of the Naval Postgraduate School. Dr. Olwell has been a problem author for both the MCM and HiMCM, and has been an MCM judge for almost two decades. He was the Head Judge for Problem C.