Judges' Commentary: Understanding Used Sailboat Prices

Rodney X. Sturdivant
Associate Professor and Director of Statistical Consulting
Dept. of Statistical Science
Baylor University
Waco, TX
rodney_sturdivant@baylor.edu

Mark Arvidson
Professor and Director of Math Fellows Program
Department of Math, Physics & Statistics
Azusa Pacific University
Azusa, CA
marvidson@apu.edu

Introduction and Overview

In 2023, a second MCM/ICM contest ran after the initial contest was complete. The Sailboat Problem was the MCM problem Y for this "pop-up" contest. The problem was similar to the main contest Problem C in that it was a data problem. A data set was provided; but, unlike many of the C problems, students were able to include additional data as needed and—in fact—one part of the problem required that they find their own data. Further, as one can surmise, it is important to properly cite any data that is brought to bear on the problem. The integrity of any submitted paper must treat imported data and its subsequent analysis with the utmost care.

Data problems require use of statistical techniques and quantification of the uncertainty of those models. Teams that did well on Problem Y thus had to include an analysis of the uncertainty in their results. Many previous Problem C problems were focused on prediction. The Sailboat Problem did have a predictive element, but the key questions required explanatory analysis. The ability to develop models that could be interpreted to answer these questions proved the key differentiator of whether a paper achieved a higher designation in the contest.

The UMAP Journal 44 (4) (2023) 449–461.

The Problem

Teams were given data with information on used sailboat sales from December 2020 around the world. Two types of boats, catamarans and monohulled, were included, as well as the selling price and some information about each boat (make, year of manufacture, etc.). While the initial data set was instructive, it was not exhaustive. Moreover, teams were given tasks designed to provide brokers who sell sailboats a better understanding of the market:

- Develop a model that explains the listing price of sailboats, and include a discussion of the precision of the model estimate of price.
- Use the model to explain the effect of geographic regions and determine if the effect is consistent for the different variants of sailboats.
- Model the regional effect of Hong Kong on sailboat prices, and determine if the effect is the same for catamarans and monohulled boats.
- Identify a few additional inferences from the data that may interest a broker.
- Prepare a report for the Hong Kong broker that includes graphics to help understand conclusions.

The problem had a few features not typical of a Problem C (data) problem in the past:

- While the problem required teams to use the provided data, it did not preclude the use of additional data. In fact, for the model of the Hong Kong regional effect, teams were tasked to find some comparable boat listing data from Hong Kong.
- The report included a requirement to include an informative graphic.
- Finally, the key task for the statistical modeling was to "explain" the listing prices rather than "predict" them.

Triage and Final Judging

Triage judges gauged suitability of papers to become finalists. Failure to address all requirements of the problem and clarity in describing models and results were among primary screening criteria. The summary and memorandum were important as well. Teams that could not articulate the basics of their models and give meaningful and useful results generally were less likely to move forward in the contest. Since papers advancing generally addressed all requirements, the final judging rubric placed the most weight on how well teams interpreted and explained their models, with emphasis on actually addressing the specific questions posed. The problem statement specifically asked in various tasks for measures of precision in prediction, and statistical and practical significance of results. Adequately addressing these requirements were critical to achieving an Outstanding rating.

In the next sections, we discuss some key items that challenged teams. We use examples from the two Outstanding papers, by Team 2332142 from Sun Yat-sen University and Team 2330646 from Wuhan University, to illustrate many of these points.

What was the Goal of the Model?

By far the most common issue for teams with the problem was identifying what their models should accomplish. For example, some teams took an approach modeled on how to determine used car prices. However, unlike some recent data problems in the MCM contest, prediction was only a part of the requirement. We discuss the issue further in the next section for the first requirement, a model for the prices of sailboats. In later requirements, involving the regions and Hong Kong, many teams struggled to clearly define the desired outcome, leading to models that were not particularly useful or results that were difficult to understand.

Common Problem-Specific Shortfalls

There were several items specifically requested in the problem prompt that were missed by many teams. In particular:

- Graphical displays in the report to the broker
- Models that explain prices
- Precision of estimates
- Practical and statistical significance

We discuss in this section these items, as well as some more-general issues that were noted by judges.

Report to the Broker

As with most MCM problems, this problem required teams to produce a report for a stakeholder. In this case, the report was meant to help a sailboat broker in Hong Kong. Teams that did not provide actionable and useful results in the report (and, similarly, the abstract) were not scored well.

Examples included simply describing the models used without actual output, or just giving the model accuracy. Rather, the analysis required that results have explanatory power for the broker.

Graphics

Unlike many problems, a specific requirement for the report was included in the problem prompt: including "a few well-chosen graphics." Some teams failed to provide a graphic and were penalized in judging. Most struggled to provide a graphic that was useful. Pictures of boats and maps were colorful but did not give brokers much insight about the sailboat market; in fact, many such "cute" pictures were a distraction and did not help the teams score well.

The most common graphics tended to depict either the factors most associated with price or the accuracy of models. The Outstanding team from Wuhan University included an example of the former, shown in **Figure 1**.

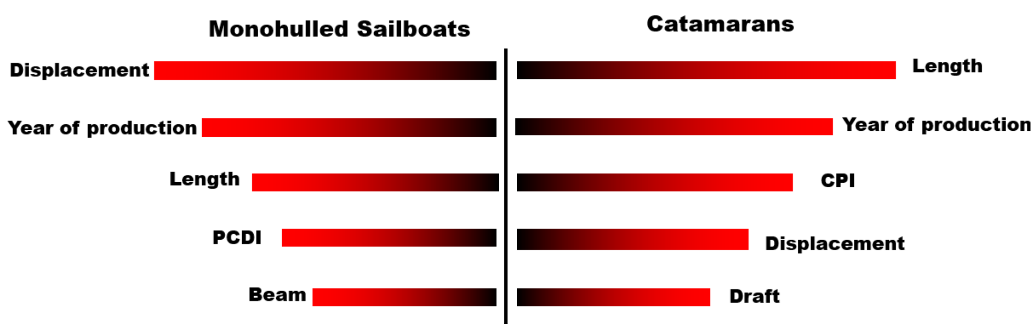

Figure 1. Graphic of top factors by boat type from the Wuhan University Outstanding team.

Even the Outstanding papers were not perfect, and this figure exemplifies a common shortcoming noted by judges. The factors and their relative importance, denoted here by the length of the bars, are provided. However, the direction of the relationship is not given. In other words, we do not know if increasing the factor causes the price to increase or to decrease. Note too that there are undefined variable names in the figure, another common pitfall.

Figure 2 is an example from Sun Yat-sen University of the second common graph giving an idea of the accuracy of the prediction model for prices.

Again, the team could enhance this figure. The regression equation is not likely understandable for a sailboat broker, for example. The team further needs to explain exactly what the graph depicts, the practical meaning, and why it is important in the report, in a way that a sailboat broker will understand.

Related to the quality of figures in reports was a tendency to produce reports that were colorful and novel. Teams should be careful not to use

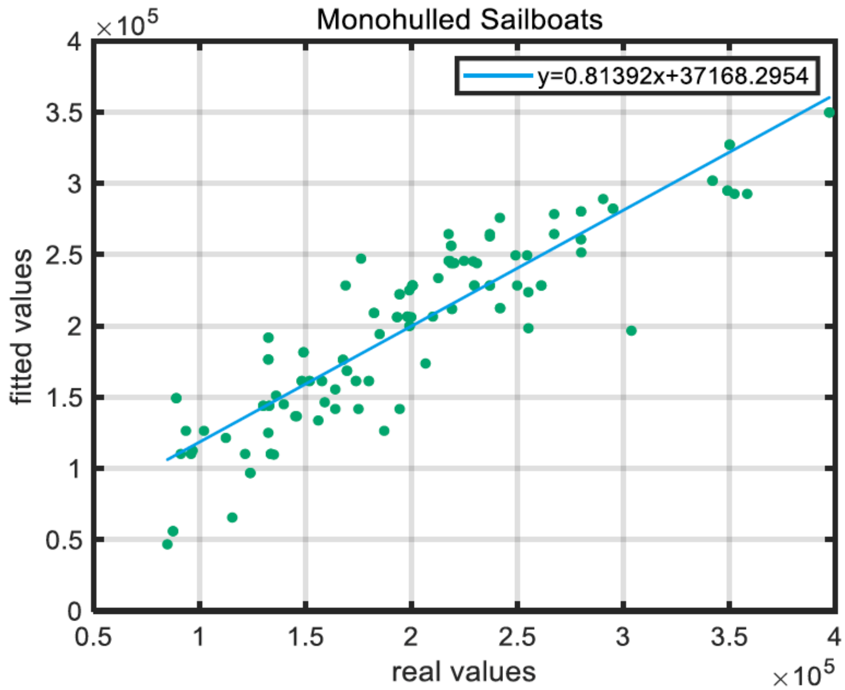

Figure 2. Predicted vs. actual prices graph from the Sun Yat-sen University Outstanding team.

fonts/colors that are hard to read (e.g., for colorblind readers), producing a report that skimps on content to be "eye-catching."

Assumptions

Almost all teams provided assumptions, although often they listed aspects that were not really important, or were not even actual assumptions but actually facts instead. A good assumption is one necessary to performing the analysis. The Outstanding paper from Sun Yat-sen University, as an example, pointed out that the team was ignoring seasonality in their analysis.

Often general assumptions are given, but no assumptions are provided related to the models themselves. As an example, linear regression significance tests assume errors that are normally distributed. Another assumption, which was critical in this problem, is independence of the predictor variables in regression models. Many teams failed to consider this assumption, leading to results that did not make sense due to multicollinearity. Further, without this recognition of multicollinearity, the accuracy of regression results can be greatly affected. Some teams utilized a random forest as a way to address this obstacle and rank importance of variables. Further, the Pearson correlation coefficient could be a useful tool to analyze the correlations among the independent variables, since this could help determine the proper regression method.

Both Outstanding papers recognized and addressed this concern. The

Sun Yat-sen University team utilized principal component analysis to create independent predictors, and the Wuhan University team selected only one predictor if there were two that were highly correlated.

Data Cleaning

The problem had several data cleaning issues that teams needed to address. The first was simply obtaining necessary data to augment the data provided. In so doing, teams need to clearly document the data source (the Sun Yat-sen University team example is in **Figure 3**). Additionally, new variables must be clearly defined; variable names are often insufficient descriptions. Poorly-defined variables compound the difficulty in interpreting the data. Many teams used linear regression, but this method cannot give the importance of the variables. Possibly a chi-squared method could help with rank importance.

Table 3: Data source collation

Source Names	Database Websites/Platforms
Beam and Length	https://www.sailboatlistings.com/ https://www.yachtworld.com/ https://www.boats.com/
Length of coastline	Google earth engine
Proportion of water area	Google earth engine
Tariff rate	https://www.usitc.gov/
Per capita GDP	https://www.census.gov/ https://www.worldbank.org/
Hong kong related data	https://www.oceanway.org/
COVID-19 related data	wx.wind.com.cn
Pictures used in the paper	Google

Figure 3. Sun Yat-sen University team data-source table.

Teams needed to address missing data. There were only three boats with missing data in the provided data set, and most teams just chose to remove them. However, most additional data had more missingness. Both Outstanding teams chose to reduce the total data set significantly. Such decisions should be documented and defended. The team from Sun Yat-sen University, for example, justified focusing on only the most-prevalent boat models.

Many teams identified outliers in the data and removed many, based on some rule of thumb such as more than three standard deviations from the mean. While there may be solid reasons for removing some outliers (not representative, clear data-entry errors, etc.), removing a large number in this manner loses potentially important information and should be done with caution. Other methods to handle the issue could be better; the Sun Yat-sen University team standardized variables, which might help. Transformations for skewed variables is another option.

A related data-wrangling action observed in many submissions was using the average price for boats of the same make/year. As with removing outliers, this may remove variability in the data that is actually important to understand, and doing so artificially improves precision estimates. The Outstanding teams were not perfect; the Wuhan University Outstanding paper unfortunately chose this approach.

Explanatory Model for Prices

We mentioned earlier that a key issue for participants was the *goal* of the modeling. In particular, the problem prompts developing a model that

- "explains the listing price" of sailboats (requirement 1);
- can "explain the effect, if any, of region" on prices (requirement 2); and
- will "[h]elp the broker in Hong Kong understand" if the model was "useful in Hong Kong."

Explaining Listing Price

A key aspect of the strong papers was to explain price variability, not just predict a range of values. Judges were keen to see a way to compute a price of a sailboat and also how the inputs were related. As previously mentioned, a good starting place could be to source the literature on statistical methods in determining used car prices. Another key aspect is to recognize the regional effects in determining a price. Specifically, in the Hong Kong SAR, a price interval needs to be articulated. While geographical factors are varied and complex, including changes among data points that are difficult to track over time, goodness of fit is an excellent way to help explain a listing price.

Thus, just producing a model that has high predictive ability is not sufficient. Many teams spent their time developing such predictive models and achieved decent results on various metrics for prediction—but never discussed how the model could explain prices or effects. Teams using machine-learning techniques, such as neural nets, struggled with this aspect of the problem. At best, most could provide only "variable importance" lists or graphs, similar to the one in **Figure 1**.

An improvement was to utilize "SHAP" (SHapley Additive exPlanations [Lundberg and Lee 2017]), which provides a direction (i.e., positive or negative) for the relationship of the important predictors to the outcome. The Wuhan University team utilized SHAP but—like the team from Sun

Yat-sen University—also performed a multiple linear regression. In both cases, the teams used methods to reduce the number of predictors in the model. The Wuhan University team used the rankings of predictors from their machine-learning (gradient-boosted) model and SHAP, while the Sun Yat-sen University team performed principal component analysis. Thus, both teams had models from which they could give explanations of prices. **Figure 4** gives an example using a table in the Wuhan University paper. The team was thus able to point out that increasing regional factors such as "Travel Income" and "CPI" would cause the price of catamarans to increase.

Factors	Coef(10^5)	Std Err(10^4)	t	P>|t|
Beam	1.34	5.28	4.033	0.000
Draft	1.345	2.3	6.145	0.000
Displacement	1.955	3.89	4.707	0.000
Sail Area	-2.38	4.57	-5.364	0.000
Sleeping Capacity	-0.7931	2	-4.301	0.000
Length	6.589	4.7	14.448	0.000
CPI	1.092	4.7	14.448	0.000
GDP Increase	-0.4179	4.7	14.448	0.000
Travel Income	0.6892	2.07	6.862	0.000
Year	3.043	1.74	17.293	0.000

Figure 4. Regression results for Catamaran prices from the Wuhan University Outstanding paper.

Explaining Regional Effect

For the second requirement, to explain regional effects, many teams used other methods more fully. The Outstanding papers added some additional analysis to the regression models. The Sun Yat-sen University team performed an ANOVA to help describe regional differences, and the team from Wuhan University used a cluster analysis. In both cases, these efforts gave insights about the actual regional differences.

Helping the Broker

Finally, the third requirement, helping the broker understand, proved very challenging for participants. Teams did not address the question or else did not provide meaningful results. The Sun Yat-sen University team utilized some of the more common methods that were most successful. They utilized their model, changing only inputs for regional factors, to predict Hong Kong boat prices and then compared the predictions to actual prices. A t-test was utilized to determine if differences were significant. From these results, a broker in Hong Kong could understand what factors influence price, and how.

Precision of Estimates

The first requirement was that teams "discuss how precise your estimate of the price of each sailboat variant is." Very few teams provided such a discussion. The failure to address precision at all was an important differentiator of papers that were considered for designations above Successful Participant. Better than nothing were teams that at least provided some measurements of the overall model precision using metrics such as R^2 and RMSE (root-mean-squared error); the two Outstanding papers included such measures. Judges would have preferred that teams then interpreted such measures in the context of the problem. In other words, what does a certain value of RMSE actually mean in terms of the price estimates?

The importance of importing additional data is reflected in the degree of accuracy needed in the results. While the initial data set was relatively small, importing data from other cited sources allowed the results to be more robust. This data set needed to include geographic data as well as sailboat characteristics. For example, with a small sample size, the task of multiple regression is made difficult. As previously noted, cleaning the data is critical to maintaining precision, especially when the original data set is supplemented with outside data.

The team from Sun Yat-sen University was one of only a handful to provide some idea of precision of the price estimate for each sailboat, with the graph shown in **Figure 5**. They gave estimates from their model of sailboat prices and included error bands for these estimates.

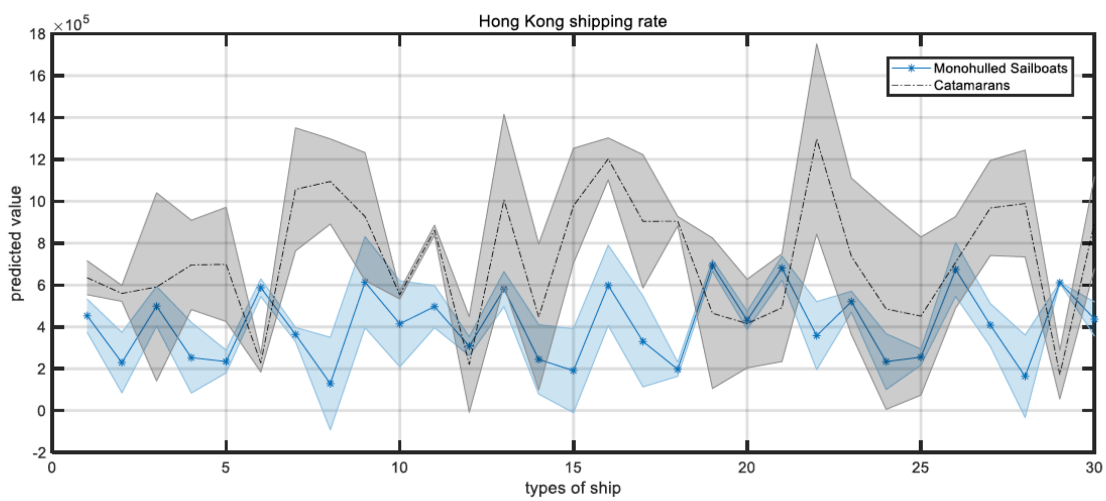

Figure 5. Price estimates with error bands from the Sun Yat-sen Outstanding paper.

Teams that utilized regression models could have done this with prediction intervals (which give the uncertainty for an individual boat price, rather than a confidence interval for the average of many boats). For other methods, some teams used RMSE to estimate basic bounds for their estimates. In general, however, such analysis was rare.

Practical and Statistical Significance

The second requirement of the problem prompt asked teams to "address the practical and statistical significance of any regional effects noted." Again, many teams failed to include one or the other (or both) of these items, instead just presenting differences by region that they noted, or factors that influenced differences.

The Outstanding paper from Wuhan University unfortunately did not provide statistical significance, using primarily cluster analysis for this requirement. However, the team did attempt to address the practical significance using plots such as the one in **Figure 6**. The plots illustrate how much the regional differences impact actual prices for a given sailboat type.

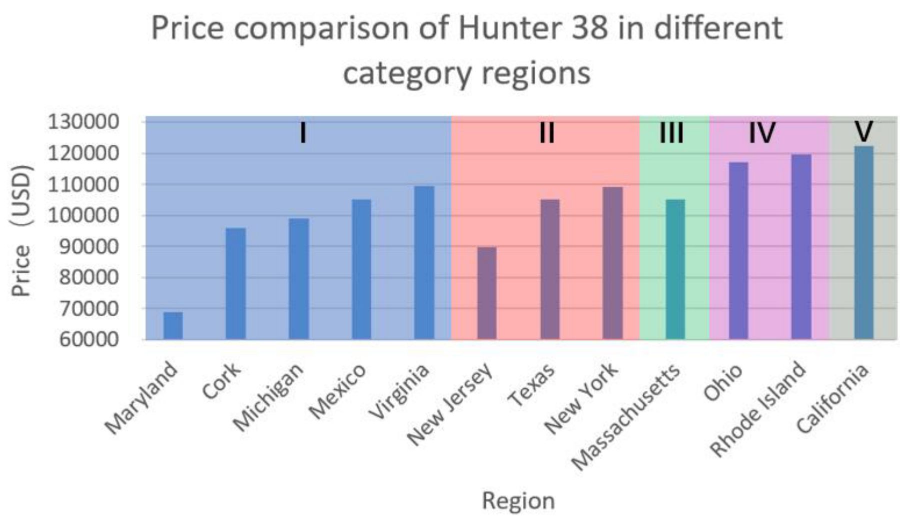

Figure 6. Price comparisons for a boat variant across regions (Wuhan University team).

The team from Sun Yat-sen University provided statistical significance using the p-value from an ANOVA for regions and t-tests for their Hong Kong analysis. For practical significance, they provided an estimate from the ANOVA, "eta-squared," which gives a measure of the strength of association. They could have improved on this by putting the estimates of the effects in some context in terms of actual price differences. However, they did include graphs such as **Figure 7**, with boxplots of prices by regions to accompany the ANOVA.

General Model Building Issues

In addition to the issues specific to this year's problem, there were other common pitfalls in solution papers, pitfalls that have been noted in most MCM data problems. In developing models, teams often failed to justify choices of parameters. Many machine-learning approaches were used in a "black box" fashion, with little discussion of why the model parameters were selected, or cross-validation or some sensitivity analysis at the end

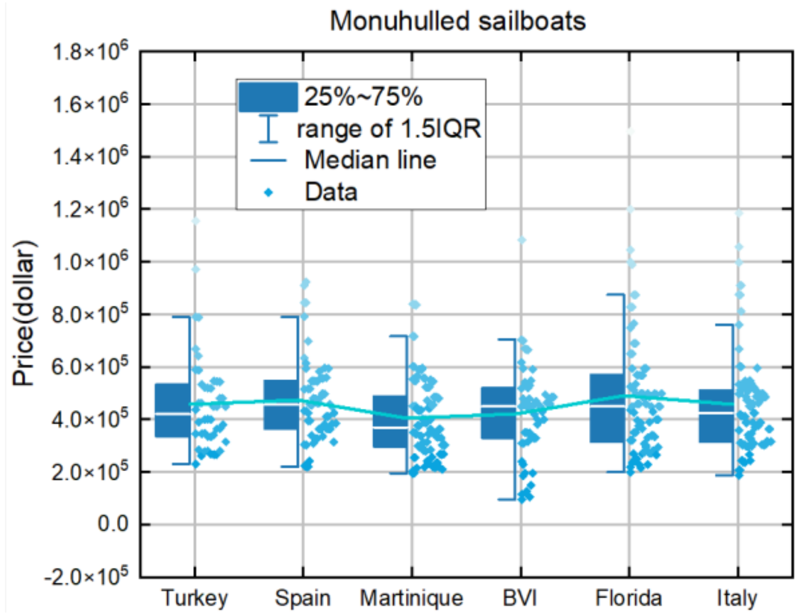

Figure 7. Sun Yat-sen University graph of prices by region.

to show the choices made sense. A simple example is for a tool such as K-means cluster analysis. The number K of clusters should be justified; the Wuhan University team, for example, used the "elbow method."

Assessing the appropriateness and adequacy of models is also often done poorly or not at all. Model assumptions should be discussed and ideally assessed. For example, linear models such as regression and ANOVA can be assessed with residual analysis. The team from Sun Yat-sen University included residual plots. Such analysis lends validity to results. Further, issues identified might provide important insights useful to decision makers.

What Made Papers Outstanding?

A paper that was rated strong included a clear explanation of results to a nontechnical audience, a Hong Kong boat broker. Explanatory graphs were required in the summary that were easily interpreted. Pictures of boats, while nice, did not help explain how a boat broker could determine a fair price for the sailboat in question. In addition, the variance of prices by geographic region needed to be addressed. For example, why are some prices higher for the same boat in a different geographic location?

Only two papers were judged as Outstanding, reflecting the difficulty that most teams had with many of the items discussed in the previous section. The Outstanding papers were not perfect, as noted. However, the largest distinguishing factor was that *both papers developed models that could truly be utilized to answer the questions posed in the requirements*. The teams both attempted to answer the question about "why" prices varied, overall and by region, and how that impacted brokers in Hong Kong.

As previously noted, the Outstanding papers also explained the results in a manner a nontechnical audience could understand and included graphs that were more than just pictures of boats but had useful explanatory power.

Conclusion

This problem proved very challenging, with very few teams earning Outstanding designation. Teams failed to address important items in the problem prompt: e.g., providing a useful graphic in the report, precision of price estimates, practical and statistical significance. The problem also highlighted the importance of understanding the goal of analysis and choosing modeling approaches accordingly. While predicting prices was part of the problem, the ability to explain the prices was critical. Machine-learning methods that many teams employed proved difficult to use in providing the necessary insights.

Finally, as with all MCM data problems, the judges rated higher papers that explained models clearly and provided useful results understandable by a decision maker.

Reference

Lundberg, Scott M., and Su-In Lee, S. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, edited by I. Guyon et al., 4765–4474. Red Hook, NY: Curran Associates. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

About the Authors

Rodney X. Sturdivant is Director of the Statistical Consulting Center and an Associate Professor in the Dept. of Statistical Science at Baylor University. He earned a B.S. from the US Military Academy, an M.S. in statistics and an M.S. in operations research from Stanford, and a Ph.D. in biostatistics from the University of Massachusetts–Amherst. He was previously Professor of Applied Statistics at Azusa Pacific University, Associate Professor and Chair of Biostatistics at Ohio State University, and an Academy Professor and Professor of Applied Statistics in the Dept. of Mathematical Sciences at the US Military Academy. He has been a final judge for both ICM and MCM for nearly 20 years, and was the assistant head judge for Problem Y.

Mark Arvidson is a professor of mathematics at Azusa Pacific University as well as an experienced teacher, educator, consultant, and presenter. He is also the Director of the Math Fellows program and trains both elementary and secondary math teachers. Dr. Arvidson did his undergraduate work at Wheaton College in mathematics; his graduate studies were completed at the Claremont Graduate University with an emphasis in mathematics education. As an applied mathematician, he has been a final judge for the MCM for five years.