Creating a Monte Carlo Model to Forecast DCI Scores

Evan Murray

Last Updated: 07/13/2017

The model uses historical data and caption scores to forecast DCI scores for any data. The model has 4 main steps, all implemented in R:

1. Read and format caption scores for every DCI show
2. Fit an exponential curve to the caption scores for each corps
3. Evaluate the curves, accounting for uncertainty, to determine Finals week scores
4. Repeat Step 3 10,000 times to develop a percentage-based forecast

All historical data was provided by corpsreps.com. Without them, this project would not be possible. Anything that was determined based on historical data was based on overall (not caption) scores provided by them from the 1995 season through the 2016 season.

Scores are recorded for each show on a per-caption basis. There are three captions – GE, Visual, and Music. In theory, scores can be forecasted for sub-captions which would allow for prediction of which corps win the caption awards. However, there are numerous issues which make this impractical, chief among them the use of incomplete judging panels for most shows in the early and middle parts of the season.

The model compiles a list of scores for a corps and the days of the season they came on. From there, it fits an exponential curve of the form $y=a+x^b$, where the independent variable is the day of the season and the dependent variable is the caption score. The form of this exponential is based on historical data, and the exponent b is always less than 1. The curve is fit using the iterated reweighted least squares method, which is a robust statistical method for fitting both linear and non-linear curves to data. This exponential is then evaluated for Finals week scores.

There are two types of uncertainty in the model we need to keep track of. The first one is the uncertainty of a and b in the exponential curve fitted to the data. The model accounts for this by drawing random a and b coefficients from a normal distribution. The distribution has a mean set by the best fit coefficients and the standard deviation is set by their respective standard errors. That means a corps will have more variation in its predictions for a corps if their scores don't fit the curve very well. When the model predicts scores, it assumes the deviation from the best-fit coefficients for all corps are uncorrelated. This is not the case for the second type of uncertainty, which we'll get to soon.

The model evaluates the exponentials for Prelims, Semifinals, and Finals. For Open Class corps, it also evaluates the exponential for Open Class Finals, which is two days before Prelims. In theory, the model can predict the scores for any show, but it will focus only on Finals week.

At this point, the model has predictions for the 3 (or 4, for Open Class) Finals week shows for each caption. At this point, we need to account for the second type of uncertainty, which is random variation from show to show. This exists because drum corps scores are noisy, and the "theoretical" score for a corps, set by the model, is rarely realized.

The noise is added to each caption score by drawing a number from a normal distribution with a mean of 0. The standard deviation of the normal distribution, which was determined based on historical data, is 0.09 for the GE caption and 0.05 for the Visual and Music captions. Unlike the first type of certainty, these errors tend to be correlated between corps. The correlation, based on historical data, has a correlation coefficient of 0.3 for all three captions.

The model draws a series of random numbers – one for each caption and each corps – which are correlated and adds them to the "base" scores determined by the exponentials. This results in the final caption score for that show. All the captions are added together to determine each corps' scores.

The model predicts scores for the Finals week shows 10000 times. Each time, it predicts each corps' caption scores by drawing *a* and *b* coefficients for the corps' exponential curve. Once each corps has a set of base caption scores, it adds random error, which is correlated from corps to corps, to all the caption scores. From there, it adds the caption scores to form its final prediction.

Once there is a set of 10000 predictions, it analyzes them to create the forecast. It forecasts these basic pieces of information:

- Corps ranking
- average Prelims score
- percent odds of making Semifinals
- average Semifinals score
- percent odds of making Finals
- average Finals score
- percent odds of getting Bronze medal
- percent odds of getting Silver medal
- percent odds of winning Founders trophy

The model will only return information for corps for which it is relevant. That means a corps with a 0% chance of making Semifinals will only get the first two columns of information. Likewise, a corps must have a greater than 0% chance of making Finals to get the last 4 pieces of information. The model returns 0 for all irrelevant pieces of information.

Finally, there are conditions which, if not met, will cause the model to ignore a corps when it makes its predictions. First, a corps must have performed at least 6 shows. Second, the curve fitting algorithm needs to be able to converge to stable coefficients. If it doesn't, the model, in essence, decides to ignore that corps until more data comes in to produce a stable result. This is rare, as the model has pretty

relatively relaxed tolerances for instability. When a model ignores a corps, it does so entirely. It's as if that corps doesn't exist.