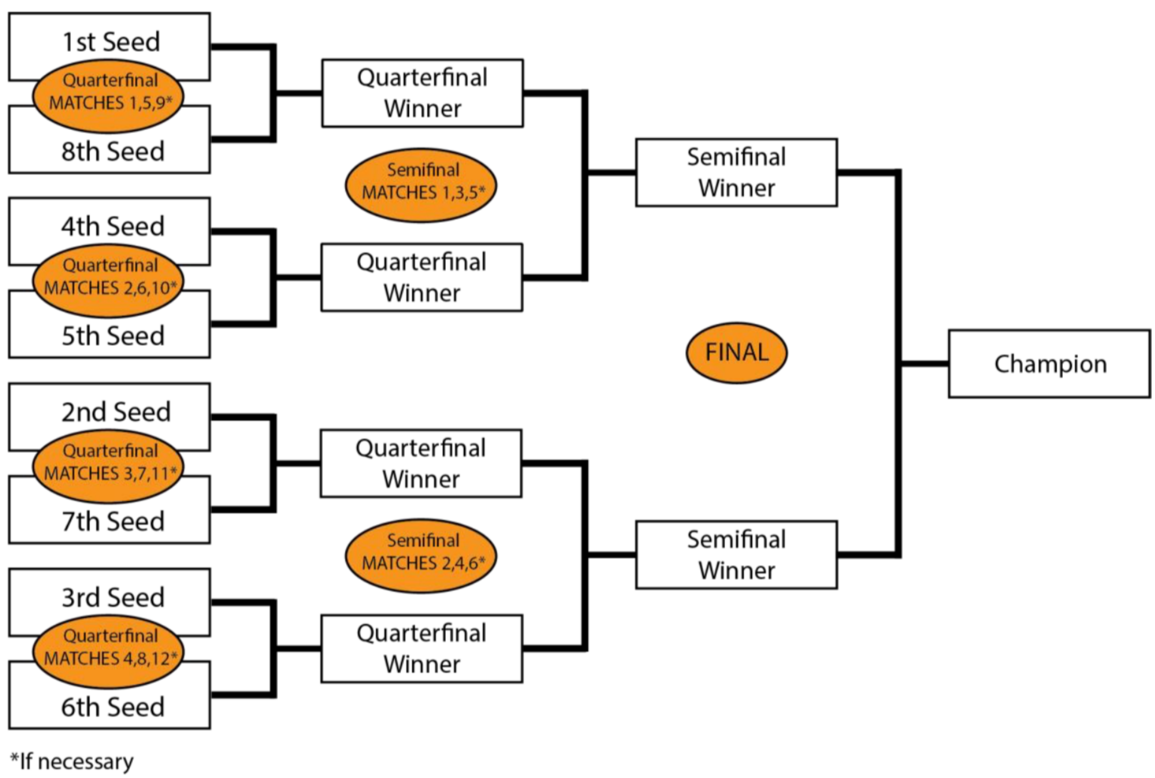
2019 FIRST Robotics Competition Predictors of Performance

By Michal Davis

Background

The 2019 FIRST Robotics Competition game had two teams of three, called the red and blue Alliances, competing to score the most points. For a succinct video overview of the rules and scoring, please watch this video: [2019 FIRST Robotics Competition Destination: Deep Space Game Animation](). In each qualification match, teams could earn Ranking Points by completing a rocket, scoring 15 points from climbing, tieing, or winning (for 2 points). Up to 6 Ranking Points could be earned each match, in theory. (This never actually happened). After all qualifying matches, teams were ranked (by ranking points : matches played ratio), and the top 8 ranked teams each picked two alliance partners to form 8 playoff alliances. The playoff alliances competed in a best 2 of three bracket like the one shown below:



(Source: FIRST)

In Ontario, each team attended two qualifying events and then 80 qualified for one of the district championship's two divisions (which were equivalent). From there, approximately 30 advanced to the world championship. A system called District Points was used to determine advancement. Teams earn District Points from five sources:

Team Age	First year teams: 10 points Second year teams: 5
Awards	Depends on award, most 5 points, some 10, others 15
Qualification Rank	Calculated by complicated formula, between 4 and 22 points for the events in this project
Alliance Selection	Alliance Captains: 17 - captain number (ex 14 for alliance 3 captain) Draft Acceptance: 17 - acceptance number (ex 12 for the team that is 5th to accept)
Playoffs	5 points for each playoff match won, up to 30 for winning the event.

Investigation and Hypothesis

This project asks "which single datapoint is the best predictor of performance?". To answer this question, a definition must be made. Performance can be subjective. To many teams, including our own, high performance weighs team experience over rank or another number. For other teams, the communication skills involved in awards submissions are highly valued. However, for the purposes of investigating the actual gameplay these sorts of concerns are irrelevant. For the purposes of this project, a team's performance was considered to be the sum of district points earned at each event in the qualification rank, alliance selection, and playoffs categories (omitting the team age and awards categories). This has several advantages: it uses metrics considered by FIRST itself to be important, and it also reflects the subtle aspects of an effective robot that something like raw score doesn't. Defense robots, for example, are usually found on every winning alliance but score very poorly without strong alliance partners. District Points also show performance beyond individual matches - they typically reward consistently excellent robots above all else.

I hypothesize that high-level cargo placement was probably the best single indicator of performance. Although placing cargo at high levels was not inherently very valuable, it only happened after high level hatch placement, which in turn usually followed low-level placements. Therefore, it seems reasonable that it would be an indicator of other valuable tasks and thus that it would correlate with performance.

Data Gathering and Methods

The initial intention for this project was to use the scouting data collected by the robotics team at our events in 2019. Unfortunately, as we switched scouting systems during that year, the data is incomplete. Solid data was only available for one event, the district championship, so I decided to use the less detailed publicly available match data published by FIRST. It covers every official and many offseason events. The datapoints available for each match include gamepiece placement, each robot's start and end levels, and foul points (though not the specific violations). Information about overall team performance, such as rank, district points, and alliances, is also available and was also used.

To access this data, I used the The Blue Alliance's (TBA) Read API, which allows download in a machine-readable format. I accessed the OPR calculations from a spreadsheet posted on an FRC forum, which also used TBA for its data. I then loaded all the data into a MySQL database for querying. To generate my graphs, I connected this data source to the data visualization software Tableau. To actually get the TBA data into MySQL, I used some tools I and other members of the robotics team had previously developed and which I updated for this project. They are written in JavaScript, and available from https://github.com/michal-davis/data-culmin.

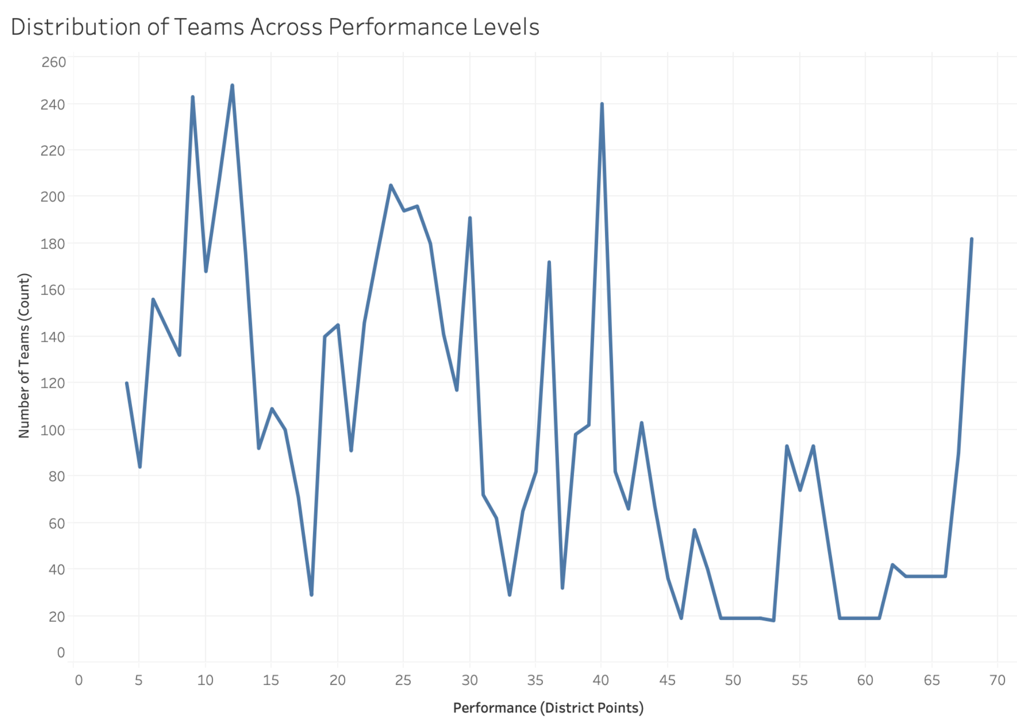
Data Preparation

My data, as initially gathered, was mainly hundreds of rows detailing specific scoring actions every alliance and team took. This was not immediately useful, requiring some calculations on it to be easily analyzed. In FRC, individual matches are too variable to use to represent a given team's ability, so all of a team's data must be grouped and measured together. Measures of central tendency are most often used for this, as well as minimum and maximum for some tasks. The software I was using and the size of my dataset made finding a median or mode extremely difficult, so I primarily used means in my analysis. I supplemented this with a considerably more advanced calculation, the Offensive Power Rating (OPR). Matches are played by three team alliances, making isolating an individual team's ability difficult. OPR uses algebra on matrices representing alliances and their match outcomes to calculated a single team's contribution to various scoring tasks. (For more on OPR, this article explains it well). The OPR calculator I used generated the number of *points* from a certain scoring task a team earned, rather than the number of times or degree to which that task was completed. For the 2019 game, the two are linearly related for most tasks so this is irrelevant but worth noting. Another aspect of OPR is that it can sometimes show a team that can, for example, only score hatch panels, as contributing some number of cargo per match. This is counterintuitive but makes sense - cargo can only be scored after hatch panels, so an excellent hatch panel robot allows its alliance partners to score more cargo. Similar situations can apply to other scoring tasks.

Analysis and Results

Once I gathered my data, I selected 10 datapoints, which together represent most aspects of the game:

- Cargo OPR: The number of cargo a team enables per match, calculated as an OPR
- Mean cargo: The mean number of cargo scored in each of a team's matches
- Hatch panel OPR
- Mean hatch panels
- Mean high level cargo: The mean number of cargo scored in the upper two levels of each rocket in each of a team's matches
- Mean high level hatch panels
- Mean team endgame level: The mean level the team climbs to each match
- Endgame OPR: The endgame points a team enables per match (level 3 is 12 points, 2 is 6 and 1 is 3)
- Fouls committed OPR: The number of foul points a team commits each match, calculated as an OPR
- Fouls drawn OPR: The number of foul points a team draws from the opposing alliance each match, calculated as an OPR



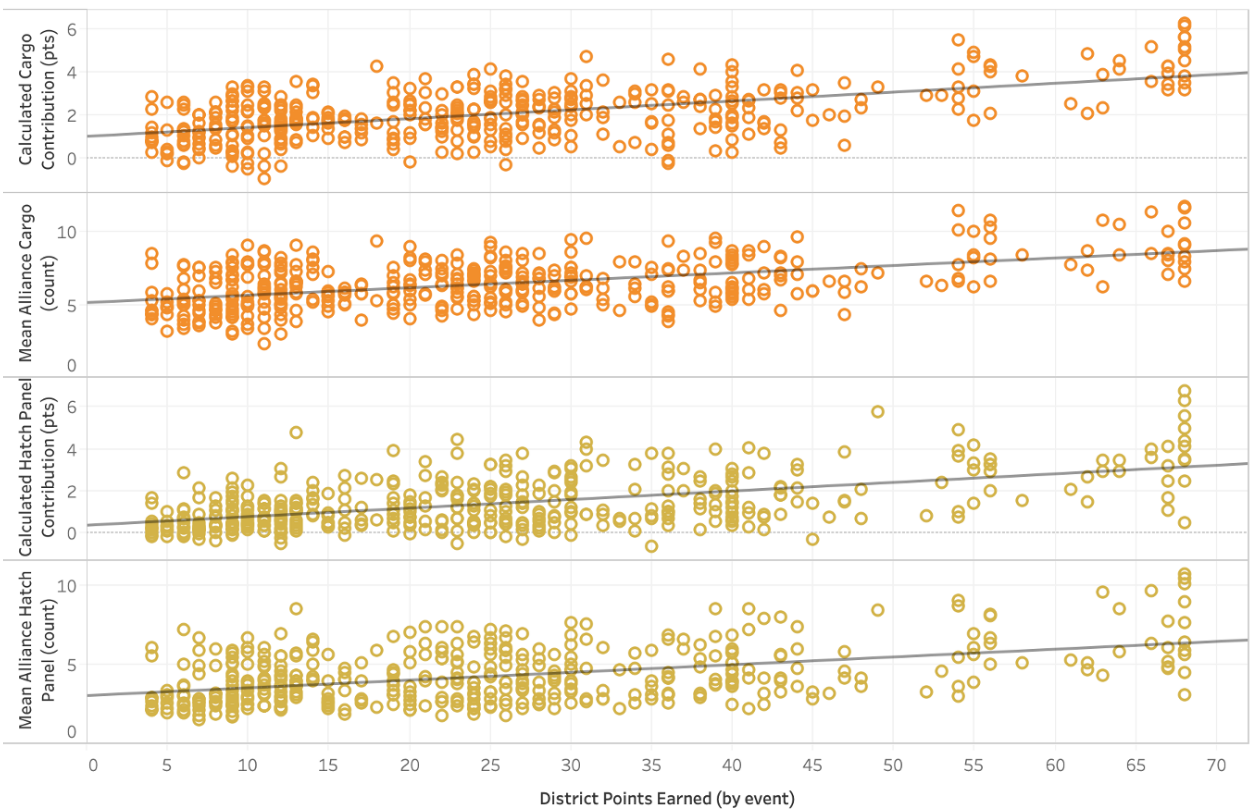
The trend of count of team_number (team_at_event) for performance_district_points.

I then created a series of graphs plotting performance against these datapoints. As I began analyzing, it became apparent there was another factor affecting the look of the graphs: the frequency of earning each possible number of district points at an event.

While fewer points are generally more common than more, the relation is far from linear. For the narrow purpose of finding datapoints which correlate with performance, this is an acceptable complication because it is consistent.

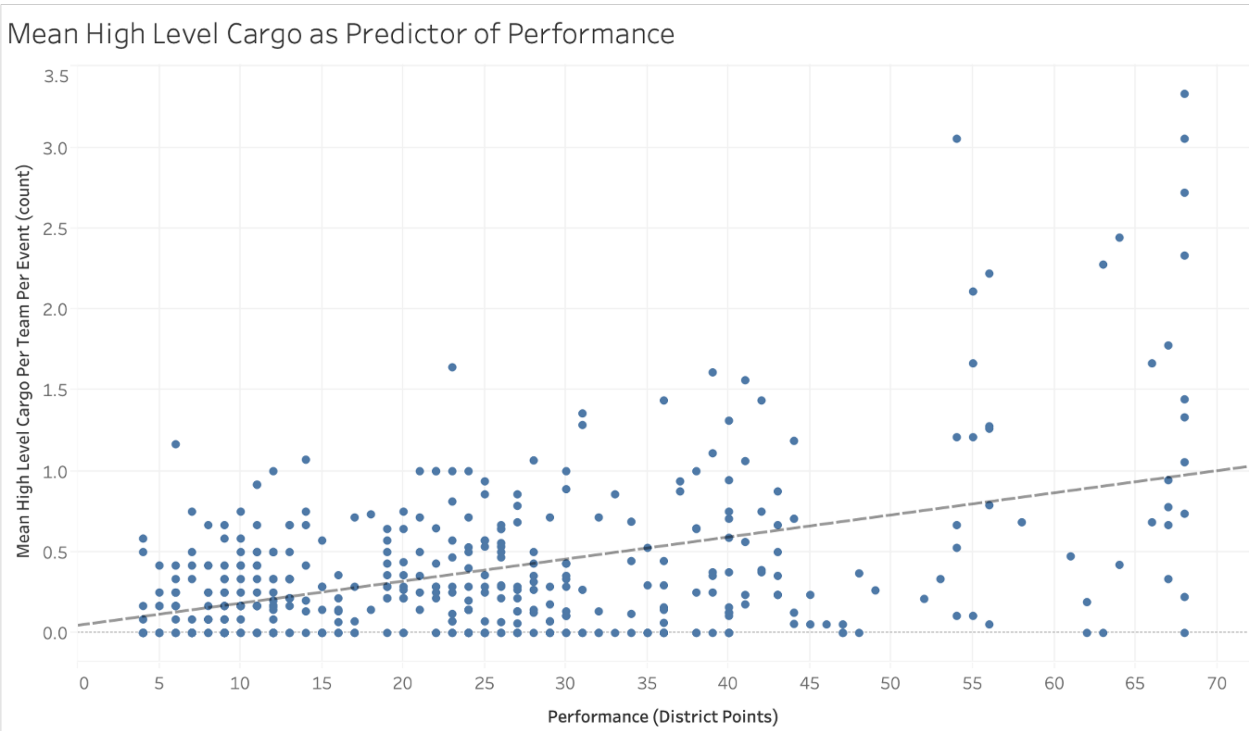
Keeping the uneven distribution of teams across performance in mind, each factor can be examined for its ability to predict performance. The most immediate result is that all factors examined are very likely to be significant at the given r values, with extremely small p values of <0.0001 for most factors. This makes sense, given that there is certainly a causal relationship between all the factors, except perhaps fouls, and performance. The only question is how strong the relationship is. Of course, it varies between factors with some being markedly better predictors than others. Each graph has interesting features warranting remark, and then all factors will be compared together.

Mean Gamepieces Compared to Gamepiece OPRs as Predictors of Performance



The plots of cargo_opr (team_opr_data1) as an attribute, average of cargo, hatch_opr (team_opr_data1) as an attribute and average of hatch for performance_district_points. Details are shown for team_number (team_at_event) and event_code (team_at_event).

This set compares the more advanced OPR calculation for gamepieces to the cruder mean values. From top to bottom, the trend lines have: $r = 0.55852, 0.51139, 0.54401$, and 0.46006. All have $p < 0.0001$, strongly indicating the null hypothesis can be rejected. Qualitatively examining the graphs, two clusters of high scoring teams appears at the top right, along the peaks in district points. The general shape is very similar for all the graphs. The leftmost section is more compressed for hatch panels than for cargo. This indicates that the lowest performing teams were closer to being equally bad at placing hatch panels than they were equally bad at placing cargo, in turn suggesting that for a weak team it was easier to perform relatively well in cargo than in hatch panels. For elite teams at the very rightmost end, however, more are above expected (above trend line) in hatch panels than in cargo. These elite teams are likely in large part elite because they did a difficult aspect well. The trend lines for mean cargo and hatch panel show the apparent difficulty difference between placing hatch panels and placing cargo, with cargo having a higher intercept and roughly equal slope.

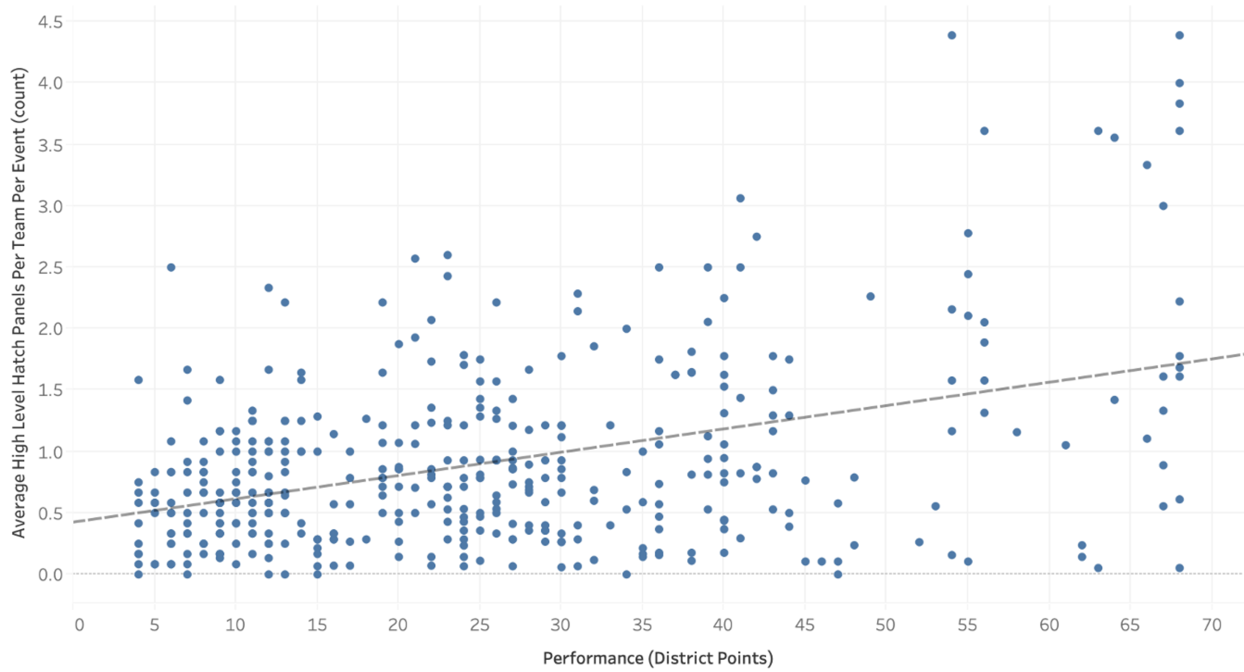


The trend of average of [mid_cargo]+[high_cargo] for performance_district_points. Details are shown for team_number (team_at_event) and event_code (team_at_event).

This graph tests my hypothesis. The trend line has r = 0.45487, a fairly weak correlation. An interesting aspect of this graph is that even at the highest levels of performance some teams were never in a match where cargo was scored on high levels. Although high level cargo is not a good overall predictor of performance, it's interesting to note that teams with very good high level cargo never achieved less than 52 District Points. High level cargo placement indicates high performance, but not the other way around.

The shape of the mean high level hatch panel graph is similar to that of high level cargo and has many of the same features. High level hatch placement also indicates high performance, but the correlation is fuzzier, with a few lower performing teams doing well in high level placement. The trend line has r = 0.41743, a slightly less strong correlation than for high level cargo. This brings up

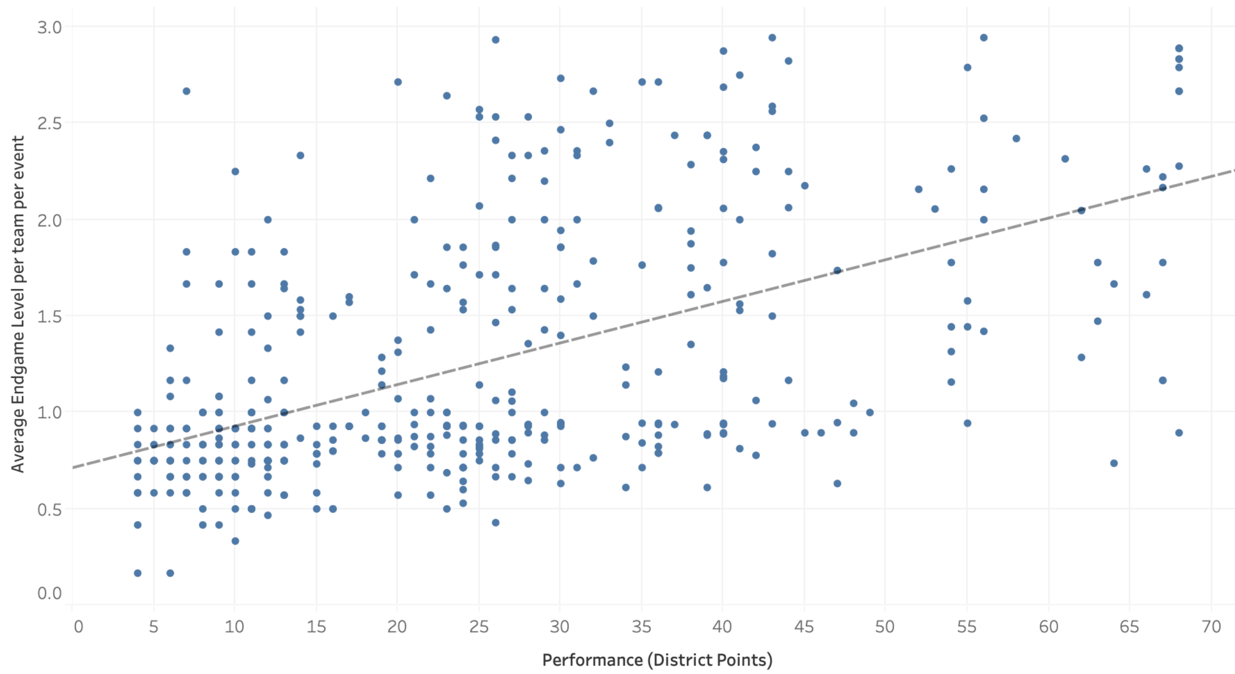
Mean High Level Hatch Panels as Predictor of Performance



The trend of average of [rocket_far_high_hatch]+[rocket_far_mid_hatch]+[rocket_near_h... for performance_district_points. Details are shown for team_number (team_at_event) and event_code (team_at_event).

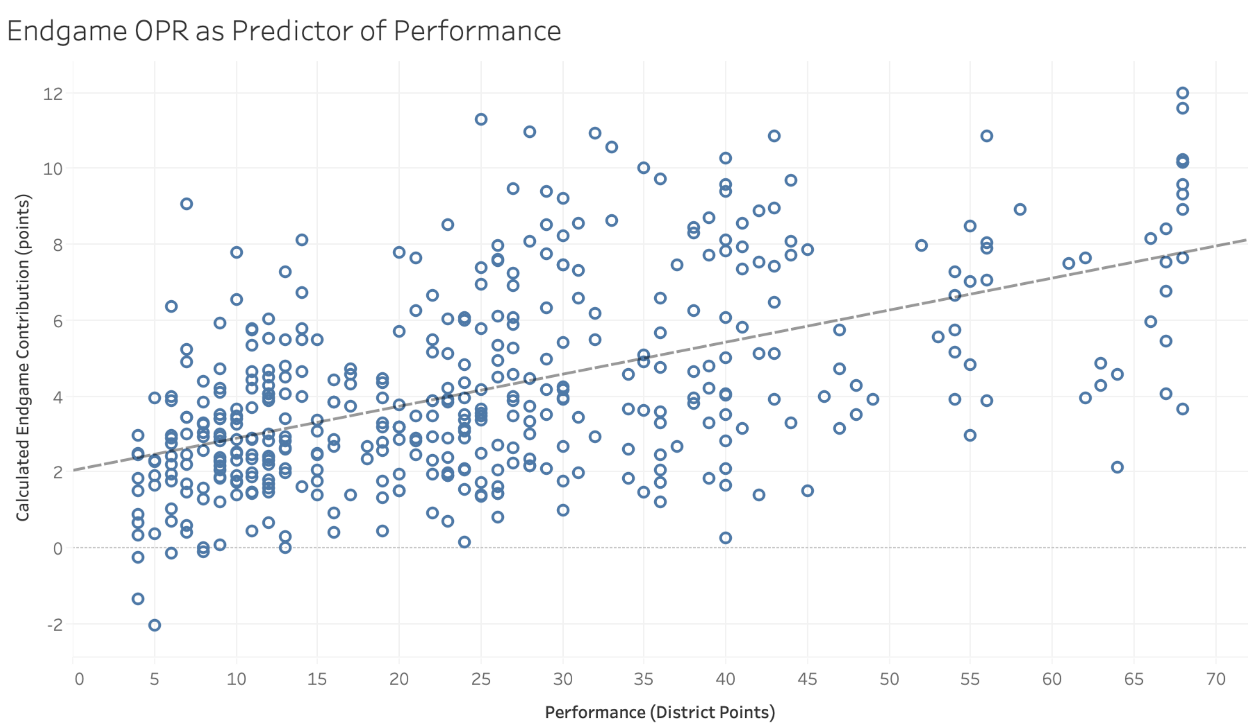
another interesting difference: though high level cargo is an overall stronger predictor, no high performing team got mean zero high level hatches. All got at least one in one match. This feature may just be noise, however: the number of high performing teams in this category is very small.

Mean Team Endgame Level as Predictor of Performance



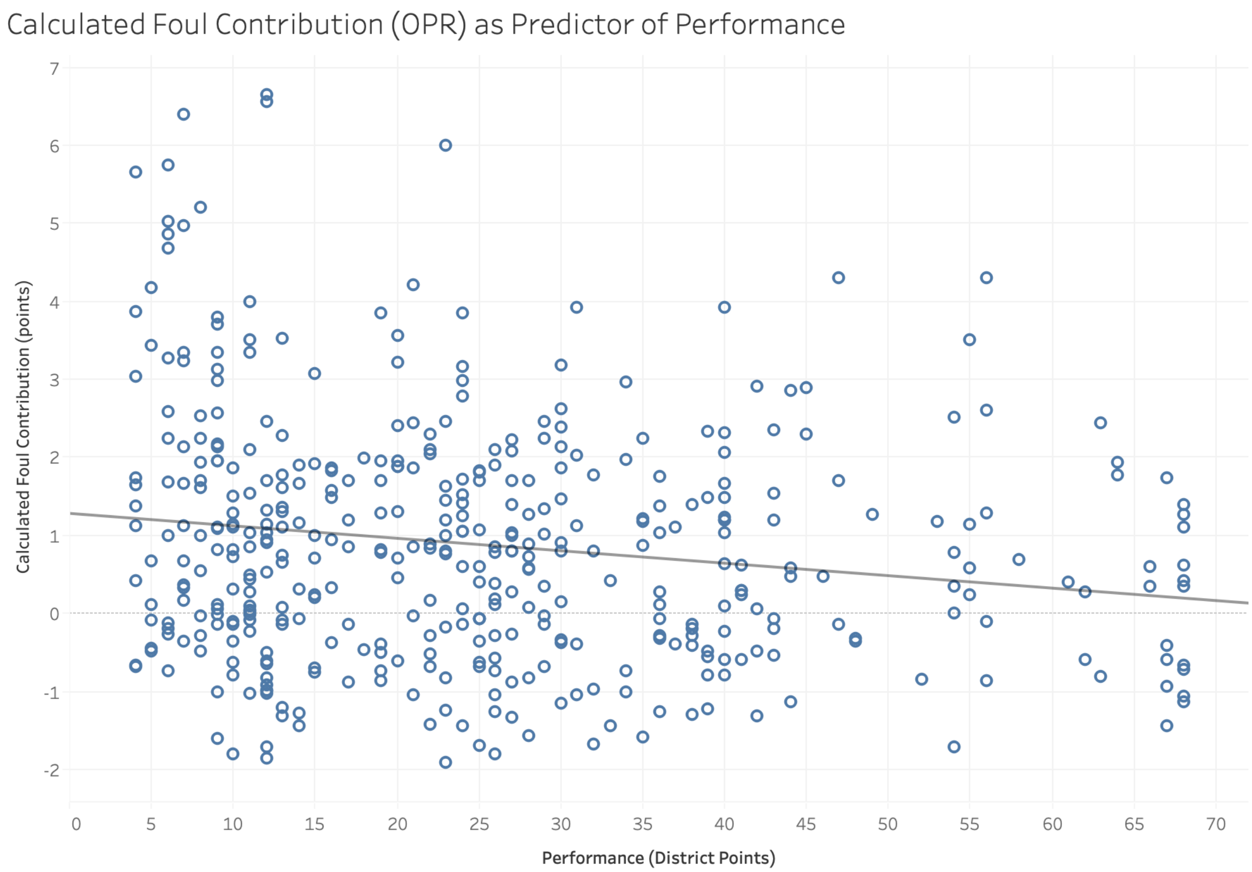
The trend of average of endgame_level for performance_district_points. Details are shown for team_number (team_at_event) and event_code (team_at_event).

The mean team endgame level is a unique factor in that it is the only datapoint that I had data for specifically about a team. While all the other datapoints I examined used a calculation to approximate a team's contribution, the dataset contained specific information on which teams climbed to which levels in which matches. This is both useful in that it is isolated, and a potential drawback in that teams climbs actually do affect each other. For example, most teams that can climb to level 3 can only do so alone - there isn't enough space on the platform for two robots, usually (this is not universally true). There appears to be a fairly strong correlation between performance and mean team endgame level, with r = 0.53955. Anecdotally, this makes sense - I didn't see a single winning or finalist alliance without a level three climb in the 2019 season, and no playoff alliance lacked at least a consistent level two, most often two of them. The graph suggests that even a fairly consistent climb doesn't guarantee high performance - above 25 district points, the upper range of climbs stays constant as district points increases (there's no upward trend). The lower range does increase appreciably. Another interesting feature is that a very large number of teams did not even consistently attain a level one climb, which is simply up a short and gentle slope. This may reflect strategy, in staying out of the way of climbers or finishing other scoring tasks right until the end buzzer, more than robot ability, but it is striking.



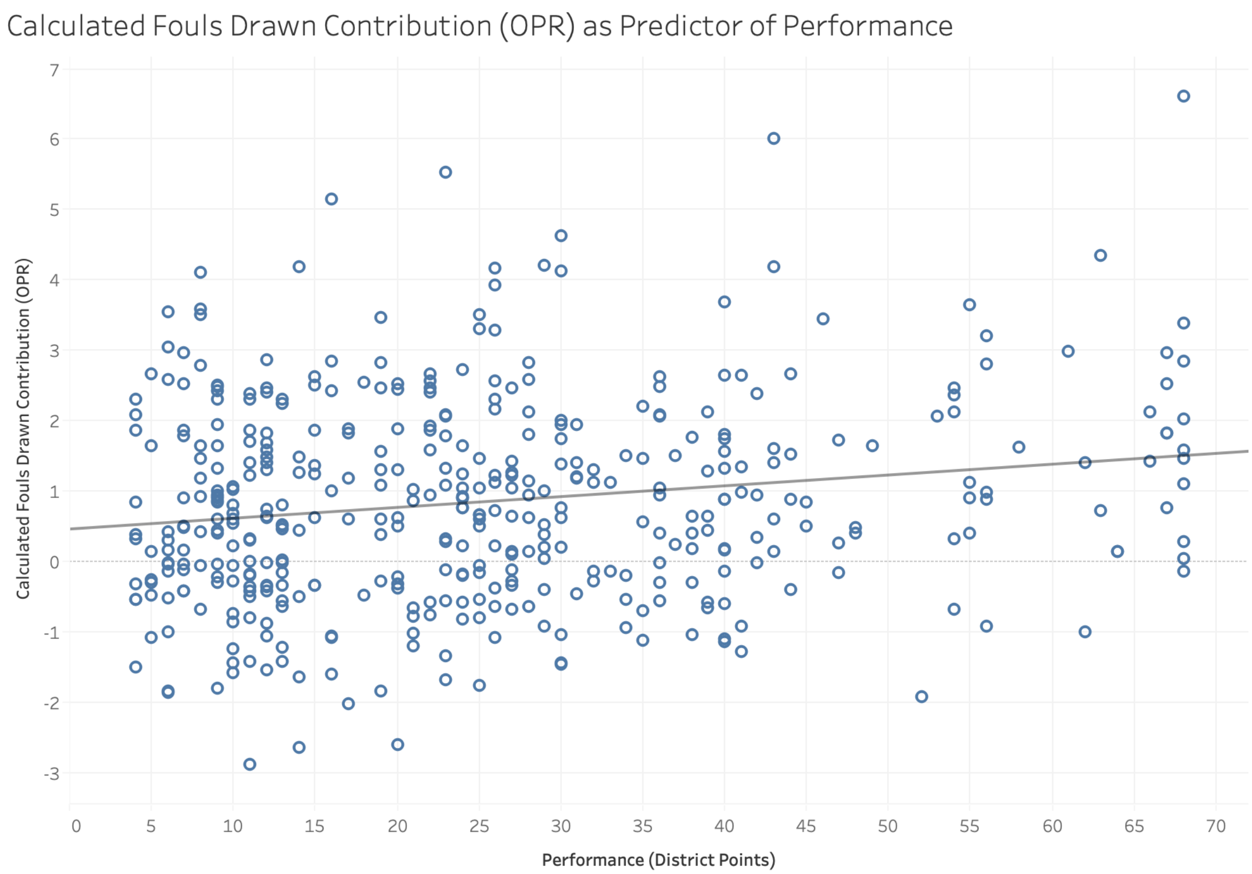
Performance_district_points vs. endgame_opr (team_opr_data1). Details are shown for team_number (team_at_event) and event_code (team_at_event).

Endgame OPR also correlates well with performance, with r = 0.53625. This is the first notable aspect of this factor: the OPR, typically the more sophisticated measure, does a (slightly) worse job of predicting performance than the mean. Two reasons are likely for this: OPR is most accurate when a scoring task is linear, which climb is not (level 1 is 3 points, 2 is 6 and 3 is 12), and the mean is a mean of a properly isolated datapoint, guaranteed to at least reflect what that robot does (though not necessarily what they would do on an arbitrary alliance). The OPR does have a striking insight: some teams contribute negative endgame points, meaning they interfere enough with partners' climbs to more than compensate for their own abilities. These negative points contributors only exist at the very bottom of performance, which makes good sense. Like the mean endgame points graph, this shows that the number of teams with a level three climb is roughly evenly distributed along performance, taking into account the frequencies of various District Point counts. It more clearly shows, however, that the "density" of level three climbs increases along with performance. Many teams in the 25-45 District Point range have level three climbs, but they are vastly outnumbered by teams in this range without level three climbs. At the 65+ district point level, most teams have level three climbs, and only a spare few lack at least a level 2 or have such an inconsistent climb it appears to be less than level 2.

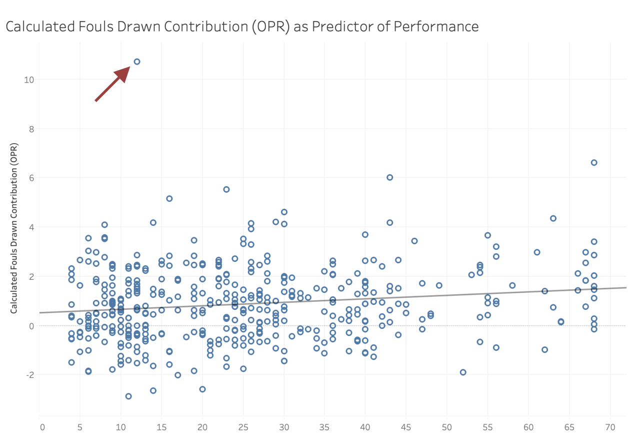


Performance_district_points vs. fouls_committed_opr (team_opr_data1). Details are shown for team_number (team_at_event) and event_code (team_at_event).

As a factor for performance, fouls are unique in that they're not supposed to happen, and are generally aimed at ensuring that robots are safe and the game can't be shut down by one or two robots aggressively defending. As this graph shows, there is a slight correlation between fouls a team commits and how they perform, but it's quite weak, with r = 0.17372. P is also high compared to the other factors, being 0.0004672 though still small enough to reject the null hypothesis. This weak correlation is explainable but still surprising. Considering that fouls are frequently the result of driver error, a good robot can't avoid them. Conversely, high performance can be achieved by a somewhat poor driver aided by a well-automated robot. However, a great deal of performance lies in getting picked for and playing in a playoff alliance. Many teams actively avoid picking even a great robot that draws many fouls, so it surprises me that there are so many fouls at high levels of performance. This could indicate that teams aren't very good at scouting fouls. Lastly, the data is quite noisy - especially above the trend line, there are quite high foul values scattered across performance.



Performance_district_points vs. foul_points_opr (team_opr_data1). Details are shown for team_number (team_at_event) and event_code (team_at_event). The view is filtered on Exclusions (event_code (team_at_event),foul_points_opr (team_opr_data1),performance_district_points,team_number (team_at_event)), which keeps 422 members.



Performance_district_points vs. foul_points_opr (team_opr_data1). Details are shown for team_number (team_at_event) and event_code (team_at_event).

Worth mentioning is that fouls drawn was the only factor where I had a clear outlier, indicated in red on the smaller graph and removed on the larger.
Fouls drawn is a similarly poor predictor of performance to fouls earned, with $r = 0.17372$ and $p = 0.0003365$. I suspect this correlation would vary widely between games, depending on how fouls can be drawn. In 2019, there was

limited opportunity for a strong drivetrain to push a defending robot into an illegal zone, for example. A somewhat odd feature of the graph is that a large number teams had a negative fouls drawn OPR, meaning that when they were on the field the opposing alliance played cleaner. This is challenging to explain with game mechanics, and the reason is probably in large part that the OPR calculation for one team is affected by that of another, even in completely different matches (though for this implementation not at different events). The teams that caused the opposing alliance to draw massive fouls raised the bar for this OPR and so many teams are negative. However, this isn't a completely satisfactory explanation.

With so many factors, to quantitatively address my question, I created this table, showing information about the trend lines for all factors. Though the null hypothesis can be strongly rejected in all cases, the r values vary widely.

Datapoint	*R*	*R^2*	*P*
Cargo OPR	0.55852	0.311942	<0.0001
Hatch panel OPR	0.54401	0.295949	<0.0001
Mean team endgame level	0.53955	0.291111	<0.0001
Endgame OPR	0.53625	0.287566	<0.0001
Mean cargo	0.51139	0.261522	<0.0001
Mean hatch panel	0.46006	0.211657	<0.0001
Mean high level cargo	0.45487	0.20691	<0.0001
Mean high level hatch panels	0.41743	0.174246	<0.0001
Fouls drawn OPR	0.17372	0.0301795	0.0003365
Fouls committed OPR	0.16939	0.0286937	0.0004672

Ranking the factors by their r values, cargo OPR seems to be the strongest predictor of performance. Although it does not match my hypothesis, this makes sense. As mentioned in my hypothesis, cargo can only be placed after hatch panels and so it can indicate success at both. Cargo OPR is also a very large composite datapoint, reflecting placing cargo at over a dozen locations on the field. As the amount of data in the datapoint approaches the amount of data for an entire match, it will of course become a better predictor. It is therefore interesting to note that mean team endgame level and endgame OPR, both measuring exactly one atomic thing, are such good predictors. The reason high level placement is such a relatively poor predictor is likely that it is relatively rare, and that its possible correlation with useful scoring tasks is outweighed by its own relative disutility.

Conclusions and Further Research

This project used quite crude methods to examine performance, lumping all three types of performance-related district points together and analyzing relatively few factor. Applying the same methods to performance broken up by district point source would give more insight, as would examining more factors. For time and brevity reasons, these two avenues were not explored. In further advancement, using machine learning to find which factor precisely is the best predictor of performance, or attempting to find how to predict match outcomes, would likely be extremely interesting. This project focused on largely theoretical aspects of FRC with no real likely applications, due to the emphasis on a such a high-level metric as District Points. Being awarded only at the conclusion of an event, they are of little interest to scouting teams. Ranking points, especially for bonus objectives, are frequently studied and predicted for scouting purposes, but I wanted to focus on a broader metric.

To conclusively answer my question of what datapoint best predicts performance as indicated by District Points, I have two answers: for the best absolutely single datapoint to predict performance, mean team endgame level cannot be beaten. Allowing composites, cargo OPR has a slight edge.

Sources

FIRST. (2019). *FRC 2019 Game and Season Manual*. FIRST. Retrieved from https://firstfrc.blob.core.windows.net/frc2019/Manual/2019FRCGameSeasonManual.pdf

Lim, R. (2019, February 26). *2019 Google Sheets Component OPR Calculator*. Chief Delphi. https://www.chiefdelphi.com/t/2019-google-sheets-component-opr-calculator/348196

Michal Davis, Moeez Muhammad, Jim Davis. (2020). Scouting Database. Retrieved from https://github.com/michal-davis/data-culmin

The Blue Alliance. (2019). Read API [JSON objects]. Retrieved from https://www.thebluealliance.com/apidocs/v3