Authorship Hash: $2a$10$h25nZm/q0tT54MJRnxan5uOoyrTFLcZAgMtpk7QD9mt.4n4u5mtt6

Introduction and Literature Review

Introduction

Throughout many different types of competitions people often feel like the opponent's team's success wasn't achieved purely through hard work, but instead due to the environment the team is located / based in. One key case is high school robotics competitions, where the access to funding, and industry experienced mentors influences a team's ability to perform. The goal of this study is to evaluate if the demographics & environment a FIRST robotics competition (FRC) team is based in significantly correlates with the team's performance. This study can also be used as a case study to look at how and when the demographic and environment a team is based in correlates with their performance in other contexts and competitions.

This study can act as a way to fact check the assumption that demographics play a major role in deciding a FRC team's performance. If we find that they are significantly correlated then it could assist in the accuracy of score prediction methods inside and outside of FRC. If we find that they are not significantly correlated then that could help reassure old and new teams that they are competing on an even playing field as their opponents.

The FIRST Robotics Competition is a high school robotics competition where teams from different schools all around the world compete in a new game. They are given 6 weeks to design, build, wire, and program a robot to efficiently complete different tasks (also called a build season). These tasks are completely different from year to year, requiring students to design completely different robots. These teams often have mentors who have experience in specific fields who can teach new students how to design / build / wire / program a robot. A major part of FRC teams is funding, these robots can range from $3500 to upwards of $10,000. In order to gather the funds for the build season teams are often sponsored by companies, ranging from a local pizza place to large companies such Boeing, NASA, and Microsoft.

Our initial hypothesis is that the demographics & environment a team is based in will correlate with the team's performance during their first several years. However as the team becomes more experienced their performance will be more related to how they performed in previous years. This will show up in the data as the error of an environmental / demographic based prediction model increasing relative to the age of the team or the consecutive years the team has played. In terms of how specific variables will affect performance we hypothesize that some statistically significant factors will be the school district's technology & extracurricular funding, the weather, and the number and payroll of industrial & manufacturing workers in the area (who could potentially become mentors for the team and / or who would encourage their kids to join the robotics team).

Literature Review

Assessment of Olympic performance in relation to economic, demographic, geographic, and social factors: quantile and Tobit approaches

2023, VOL. 36, NO. 1, 2080735

This dissertation utilized quantile and Tobit regression models in order to determine if a country's economic, demographic, geographic, and social factors correlate with the country's gold medal ranking performance. They found that some factors, such as inflation rate, and income rate appear to influence the country's medal ranking. They also found that other factors, such as GDP ranking, topography, and corruption level, do not appear to influence the medal rankings. This dissertation gives precedent that environmental factors can correlate and / or potentially influence sports performance. This also shows an example of how the scope of the athletes often matches with the scope of the environmental variables considered. Here, because the dissertation is looking at Olympic athletes who could come from many different backgrounds in the country, the dissertation is only considering economic, demographic, geographic factors related to the country, instead of the individual athletes. This study doesn't focus on a competitive environment where there are a large number of game-related or game specific statistics. This is because the Olympics is not just one sport or game where one can measure the number of shots taken, number of red cards, assists, etc (taking soccer as an example).

Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team

International Journal of Computer Theory and Engineering, Vol. 5, No. 5, October 2013

This study utilizes a Bayesian Network in order to predict the final result of Spanish league football matches. This study model does not utilize any variables that aren't influenced by the team's performance, with the exception of the weather at the time of the match. The Bayesian Network model in the study correctly predicted the match result 92% of the time. This shows that prediction models are able to accurately predict the performance of experienced teams without considering the environment the teams are located or trained in. However some environmental variables such as the weather can influence the outcome of a match.

An Integrated Framework for Team Formation and Winner Prediction in the FIRST Robotics Competition: Model, Algorithm, and Analysis

This study is utilizing the same domain (FIRST robotics competitions) as this study intends to use. Instead of looking at score prediction or the expected points added by each team they are evaluating the composition and formation of teams during the playoff portion of the game where high ranking teams select which teams they want on their 3 robot alliance. The goal of the study was to develop an analytical approach to developing teams based on prior data. This study is an example of another study utilizing FRC as a domain in order to evaluate statistical models or develop new ones.

Oftentimes when it comes to score prediction studies will be looking at Game-Related Statistics. It can be shown that by using purely game related statistics one can find some fairly accurate predictive models. However there is very little focused work surrounding whether or not environmental variables (meaning variables that aren't game related) can correlate with the performance of the teams. Oftentimes when studies look at environmental variables they are looking in domains that aren't competitive in nature such as the olympics, or don't consider solely environmental variables (this often shows up as how weather affects the results of a match). There is little to no academic research that we have found that specifically focuses on how environmental & demographic factors can correlate and predict a team's performance in a competitive context and how the accuracy of the predictions change over time.

Our study intends to analyze and fill in this gap within the domain of the FIRST robotics competitions. This is because every year FIRST releases an entirely new game that teams have to build a robot to complete tasks for. Meaning that most game related statistics are different from year to year and that team's can't train or practice game specific skills other than generally applicable skills. In most other sports a team can train to perform a specific aspect of the game, while in robotics gameplay specific skills don't often transfer from year to year. The main skills that transfer between years are skills relating to the construction of the robot.

The performance of a FIRST robotics team is related to the team's ability to build an effective, robust robot quickly during the 6 week build season before the competition season starts. Because a large portion of performance is determined over 6 weeks this also makes short term environmental variables like the weather on any given day have less of an effect on the team's performance over the season.

By utilizing this domain this study also serves a dual purpose. Oftentimes people feel like the location and school funding of other FIRST robotics teams gives them an advantage (which are factors that the FRC team cannot easily control). This study can potentially act as a way to check if that assumption has grounds and requires further research.

Method, Process, and Approach

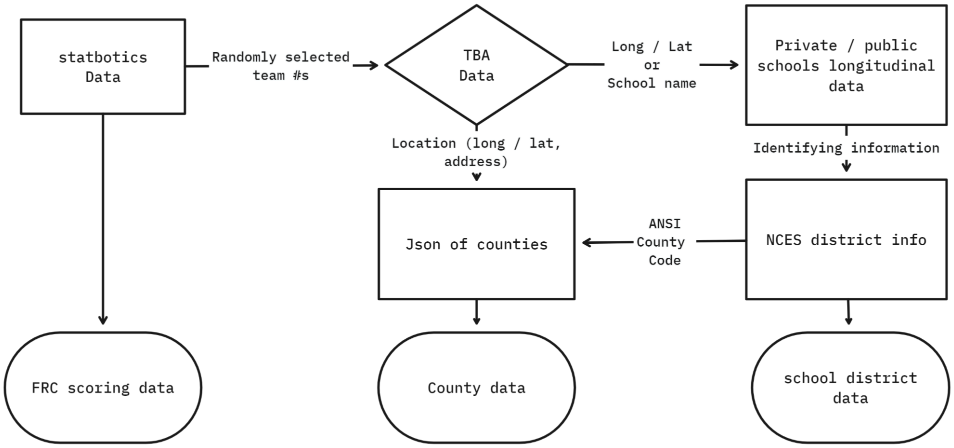
Research design

We chose our domain (FRC) because of the ease of access to data, the impact of environmental variables, and quantity of data that can be accessed. In FRC a large component of the games is determining if a team is high performing and is being under or over represented by the ranking system. This allows the team to form an alliance with 2 other robots that can best perform the task in concert with their robot robot. Because of this there is already a roughly standardized system for measuring a teams ability to score called EPA (Expected Points Added [to the alliance]).

Alongside this FIRST also published the total points each alliance accumulates which allows 3rd party apps (such as the commonly used app "The Blue Alliance") to calculate the EPA of the teams. This can then be associated with a normalized score which allows the performance of the team to be compared between different year's games (which often have different score distributions).

We are performing an observational study with mixed (qualitative and quantitative) analysis of FRC teams, environments and scores. This allows us to draw correlations between the environmental variables and the performances. We specifically avoided any data that is based directly on subjective opinions or that is likely to not be accurate. This would be data such as subjective scoring of a driver's performance, generally available scouting data (as it's often generated by a single team member who could accidentally put in inaccurate data). Instead we focus on data that is either derived from or is directly measured from the game / environment. Such as the total scores of each team (and in turn EPA), funding levels of associated schools, etc.

Data collection & processing



All of the data used is widely available from online sources.

Collection

The major datasets were the statbotics team score datasets for each year, The Blue Alliance team Dataset, the json of counties dataset, and the NCES district & geolocation info.
Every dataset, with the exception of the blue alliance, was able to be directly downloaded from their respective sites.
In order to download The Blue Alliance team dataset we coded a small python script in order to query the API and convert it to a CSV file. We directly discussed the process with the developers of the APIs to ensure that utilizing python was appropriate.

Processing

We utilized several jupyter notebooks in order to handle the data processing.
Our first step is to create an index file. This file contains the Team key (“frc” + team #), unique identifier for the School district, and a unique identifier for the county.

This was achieved through several interesting techniques.
There were 3 methods to get the district data from The Blue Alliance Dataset
We base it off the zip code, school name, and city name.
Each of these will generally give multiple results or, if they give one result, they are not as accurate (meaning they have a chance to return a null value).
From there, once we had the LEAID (the school district identifier) we got the county from the public school’s geolocation dataset.

In order to search all public schools for specific school names we utilize a “fuzzy” string searching algorithm against all public schools in the same state as the team.

It first looks at the zip codes for all valid teams (teams that do have a school name or have a zip code associated with them), then collects all schools in that zip code.
It then performs a similar operation based on the city. (using the fuzzy string search as this is user generated data which does occasionally contain small errors, such as an extra space or dash)
It then takes the intersection of the two groups of LEAIDs.

Then, if the school name isn’t null, then it’ll use the school name to get a list of LEAIDs. If any of them are present in the intersection of the above two groups it’ll default to that LEAID. If it isn’t present it’ll return the list of LEAIDs.

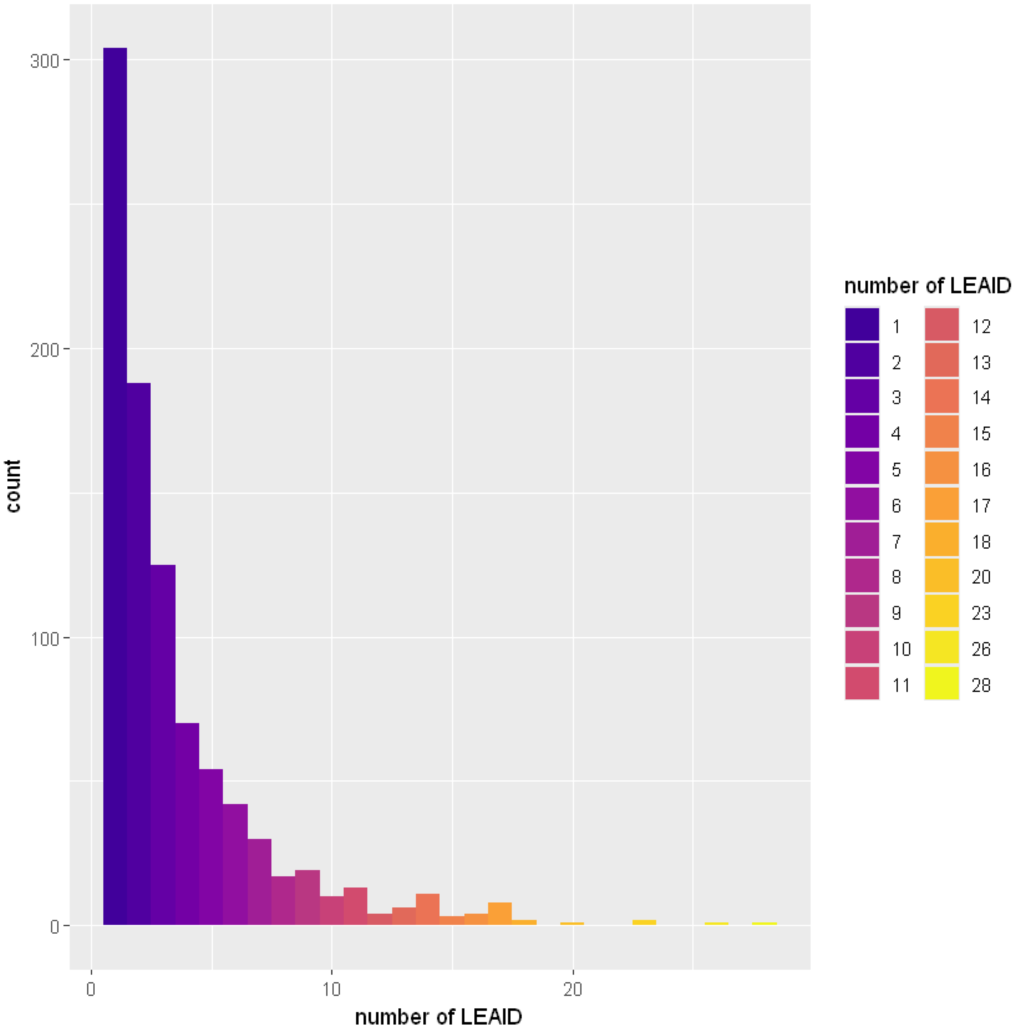
Once we have a list of potential LEAIDs we create a summary table where we convert the, potentially multiple, LEAIDs into the summary statistics (taking the average value of the teams with multiple LEAIDs) alongside the FRC key & county FIPs (the unique county ID)

Limitations

The main limitations of how this data is collected together removes certain subgroups that we weren’t considering. The main group that gets removed is the private schools, since they are not a part of a public school district.
When looking at school district or county statistics we can only look at teams associated with a public school & based in the US.

Another smaller limitation is the fact that, especially with more LEAIDs, the school district statistics would get less accurate.

However, due to the method we utilize in order to find the LEAID, all LEAIDs are going to be next to each other, meaning that environmental factors outside of the LEAIDs are likely to have similar statistics.



Analysis plan

We are planning to analyze the data through several tests:

- EPA V. Funding
  - Chi squared test
- EPA V. Age
  - Paired T-test
- Predictive model of EPA from environmental factors
  - Finding a predictive model for the EPA of the teams based purely on environmental factors
- Error Vs Streak
  - Anova test comparing the predictive model error & the consecutive years played
  - This test is very likely to fail the tests for a traditional ANOVA test, so if that is the case we plan on using a Kruskal-Wallis test (a non-parametric alternative to a ANOVA test)

EPA V Funding will be utilizing only the school district valid teams,

EPA V Age will be able to split the team summary EPA statistics into a row for every year the team participated in.

The predictive model will be using the school district teams & county data.

The Error Vs Streak will be comparing the year by year scoring data with the

Results, product or findings

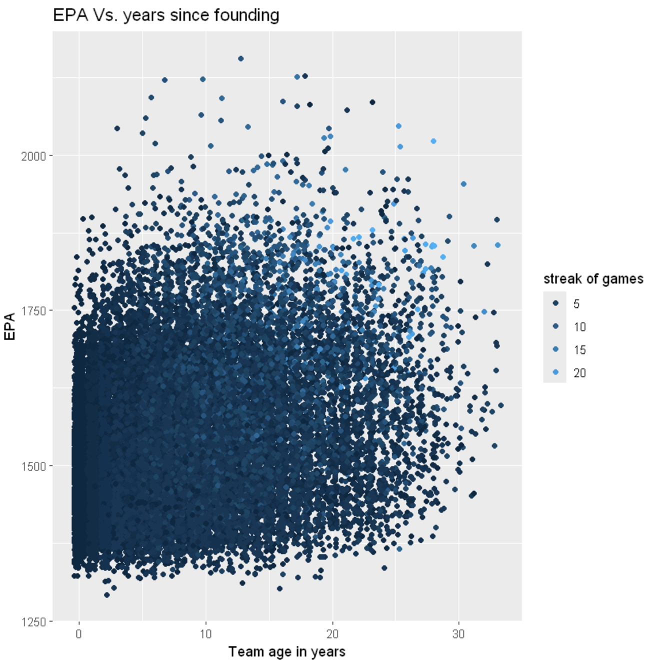
EPA VS Team Age

μ_P = the true population mean difference in EPA of the first 3 played years subtracted by the mean EPA of the rest of the year’s the team has played.

H_0: $\mu_P = 0$

H_0: $\mu_P \neq 0$

$\alpha = 0.05$



Checks

<u>Randomness</u>

The sample was selected from the dataset using R’s random sample functions.

<u>Independence</u>

The population size is less than 10% of the team’s whose data was easily accessible (8831)
$N * 0.1 > n$
$8831 * 0.1 > 200$
$883.1 > 200$

<u>Normality</u>

Because our sample size (n=200) is greater than 30 we can assume normality by CLT

Math

$t = \frac{\bar{x}_p - \mu}{\frac{s}{\sqrt{n}}} = \frac{12.33}{\frac{79.76}{\sqrt{200}}} = -2.1859,\ df = 199$
P-value = $P(t < -2.1859) = 0.02999$

Conclusion

Because our p value (0.03) is less than our alpha (0.05) we can reject the null hypothesis. This means there is convincing evidence that the true population mean difference in EPA of the first 3 played years is significantly different from the mean EPA of the rest of their FRC career.

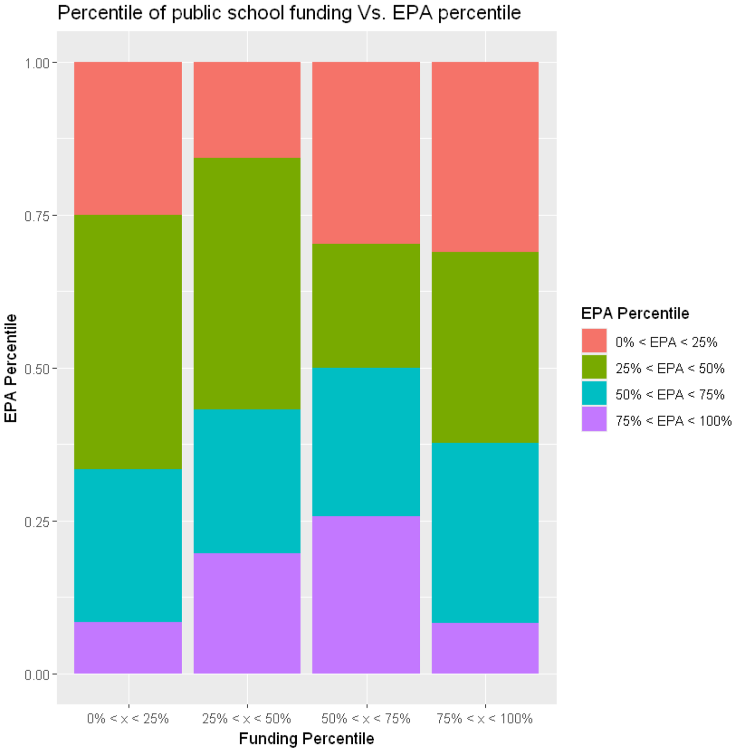
EPA VS Funding

χ^2 Test of independence

H_0: The percentile of public school funding & EPA percentile are independent
H_a: The percentile of public school funding & EPA percentile are not independent

$\alpha = 0.05$

Percentiles:
0 → 25%
25% → 50%
50% → 75%
75% → 100%



Checks

Sample was selected randomly using R's built in sampling method.
Expected cell counts are ≥ 5

Results

The chi squared test showed convincing evidence ($p = 0.02394 < \alpha = 0.05$) that the null hypothesis is false. This means there is convincing evidence that the percentile of the public school's funding & the associated robot's team's performance are not independent.

Predicting EPA from environmental variables

Variable	Variable Meaning
$\hat{Y}$	Predicted EPA
X_1	Total School District Revenue
X_2	Snow
X_3	Precipitation
X_4	Temperature
X_5	Number of educational employees in the county
X_6	Number of information employees in the county
X_7	Number of Scientific employees in the county
X_8	Latitude
X_9	Longitude
X_{10}	Poverty rate
X_{11}	Altitude

Model	Rsq	RsqAdj	Sigma	Fstat	F Prob
$\hat{Y} = \beta_1 X_1$	2.49E-05	-0.00503	110.032	0.005	9.44E-01
$\hat{Y} = \beta_1 X_1 + \beta_2 X_2$	2.97E-04	-0.00985	110.296	0.029	9.71E-01
$\hat{Y} = \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4$	8.70E-02	0.07307	105.67	6.229	4.63E-04
$\hat{Y} = \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4$	8.32E-02	0.06918	105.892	5.93	6.83E-04
$\hat{Y} = \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$	9.63E-02	0.07756	105.771	5.141	5.88E-04
$\hat{Y} = \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_6 X_6$	9.96E-02	0.08114	105.209	5.393	3.86E-04
$\hat{Y} = \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_7 X_7$	1.14E-01	0.09581	104.366	6.272	9.10E-05
$\hat{Y} = \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_8 X_8$	9.80E-02	0.07449	105.947	4.171	1.28E-03
$\hat{Y} = \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_9 X_9$	9.71E-02	0.07362	105.997	4.131	1.38E-03
$\hat{Y} = \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_{10} X_{10}$	1.66E-01	0.14447	101.863	7.653	1.00E-06

$\hat{Y} = \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_{10} X_{10}$	1.66E-01	0.14887	101.600	9.614	4.13E-07
$\hat{Y} = \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_{10} X_{10} + \beta_{11} X_{11}$	1.75E-01	0.15338	101.330	8.138	5.49E-07
$\hat{Y} = \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_7 X_7 + \beta_{10} X_{10} + \beta_{11} X_{11}$	1.85E-01	0.15907	100.990	7.211	5.98E-07

Process

We started with the schools total revenue (X_1) as our starting place, we then started adding environmental factors to reduce the standard deviations of the error. We added variables one by one in groups, first it was weather & temperature, then was the population statistics (such as the # of educational employees), next was location (longitude & latitude), etc.
Once we started getting diminishing returns by adding terms we looked at the summary statistics of the best model and found that the total revenue was not significantly affecting our predictions and, as such, we removed the term. We then added the scientific employees term back into the formula to arrive at our chosen model:
$\hat{Y} = \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_7 X_7 + \beta_{10} X_{10} + \beta_{11} X_{11}$

X_n	Intercept	X_3	X_4	X_5	X_7	X_{10}	X_{11}
β_n	1.789e+03	0.2773	-3.141	-4.373e-04	4.216e-04	-6.215	-4.798e-02

Rsq	RsqAdj	Sigma	Fstat	F Stat Prob
1.85E-01	0.15907382	100.9896	7.210918263	5.98E-07

F stat test

Chosen Model:

$$\hat{Y} = \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_7 X_7 + \beta_{10} X_{10} + \beta_{11} X_{11}$$

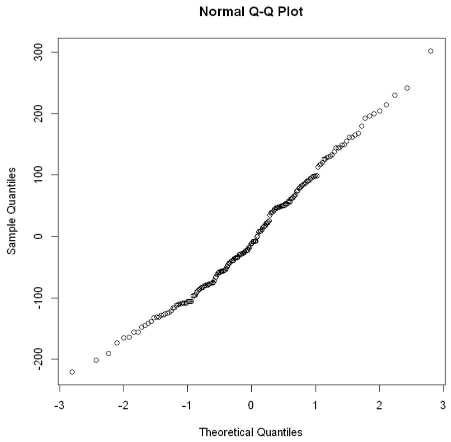
H_0: $\beta_3 = \beta_4 = \beta_5 = \ldots = \beta_{11} = 0$

H_a: at least one of the predictors β_3, β_4, β_5, ..., or β_{11} are not 0

$\alpha = 0.05$

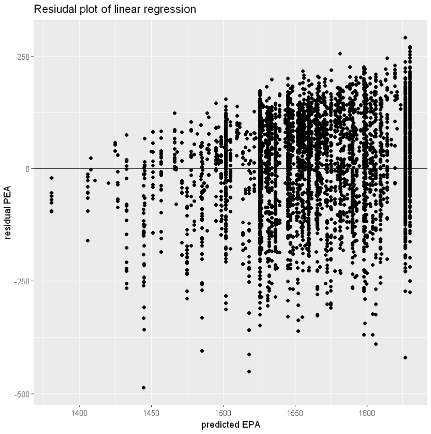
$F = \frac{MSTr}{MSE}$

LINE Checks



Normality

Because the normal probability plot (on the right) is approximately linear we can assume normality in the residuals of the model.



Equal variance

The residual plot on the left shows us that the residuals are approximately equally distributed around 0.
Although due to the way the error seems to be correlated with the predicted EPA we should proceed with caution.

Math

$$F = \frac{\frac{R^2}{k}}{\frac{(1-R^2)}{n-k-1}}$$

R^2 = 0.185

$k = 6$

$n = 198$

$$F = \frac{\frac{0.185}{6}}{\frac{(1-0.185)}{198-6-1}}$$

$F = 7.211$

$$df = \frac{k}{n-k-1}$$

$$df = \frac{6}{198-6-1}$$

$$\text{pvalue} = P(F > 7.211), df = \frac{6}{198-6-1}$$

$\text{pvalue} = 5.978 \cdot 10^{-7}$

Conclusion

Because our P-value ($\approx$ 0) is less than our alpha value (0.05) we can reject the null hypothesis. This means that there is enough evidence to suggest that at least one of the predictors is not 0, this means that there is a statistically significant relationship between at least one of the environmental predictors & the EPA of that team during that year.

Error compared to number of consecutive years played

Prelude

H_0: The true population mean squared residuals are equal for all streak lengths
(streak lengths being the number of years the team has played consecutively up to the year tested)

H_a: At least two of the true population mean squared residuals are different
$\alpha = 0.05$

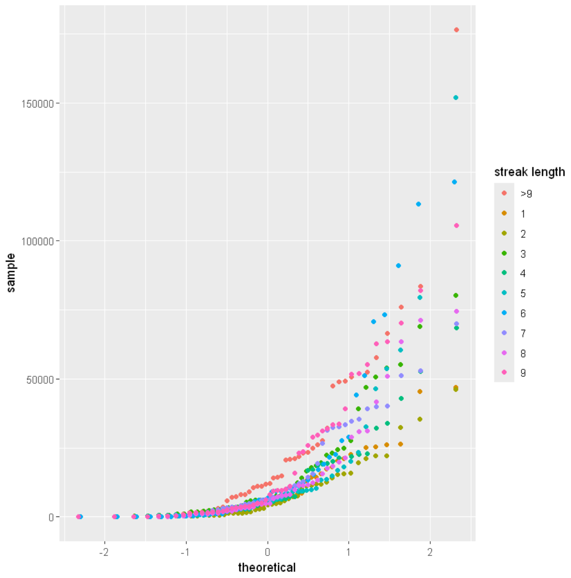
Conditions

Independence

We can assume independence

Normal distribution

As can be seen from the normal probability plot to the right none of the streak's appear to be linear, However we can assume normality since each distribution has a sample size of 30, which satisfies the conditions for CLT

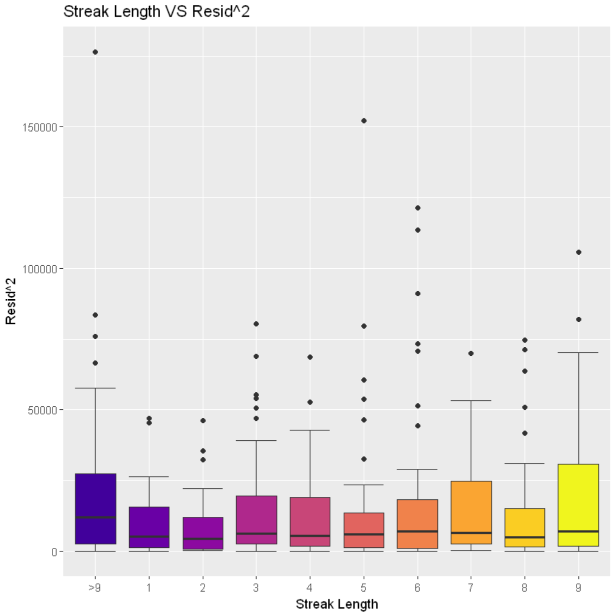


Homogeneity of variances

We performed a levene test ($\alpha = 0.05$) and because the p value (0.02424) is below our alpha value we reject the null hypothesis. This means that the distributions of squared residual cannot be assumed to have the same variances.

No significant outliers

Because of the presence of outliers we are unlikely to be able to use a normal anova test to compare the distributions.



We utilized a non parametric version of the anova test to determine if at least one of the true population mean squared errors are different depending on the streak length. (Kruskal-Wallis test)

Outcome

Because our P-value (0.1375) is greater than our alpha (0.05) we fail to reject the null hypothesis. This means that there is not convincing evidence that two of the true average population squared errors are different when comparing based on consecutive years played.

Discussion, analysis and Evaluation

Interpretation of results

EPA V Age

This test is used to demonstrate that a team's mean performance in their first handful (3) of years is statistically significantly different than their mean performance in the rest of their FRC career. This confirms the assumption that, the longer a team has been participating in FRC, their performance in the game will change. This gives precedent to further tests that would allow us to investigate further into what affects and contributes to their performance.

Funding V EPA

This test showed that the EPA percentile & the funding percentiles of a team are not independent. This shows us that these two variables are likely to not be independent of each other. This is an example of one of many environmental variables that are correlated with the team's performance.
This test directly tells us that a team's performance is likely related to how much funding they get. This warrants further investigation into what other environmental factors contribute to a team's performance.

Predictive model

We chose our predictive model through a refinement process looking at several different environmental variables
We looked at total school district revenue, amount of snow, amount of precipitation, average temperature, Number of educational employees in the county, Number of information employees in the county, Number of Scientific employees in the county, Latitude, Longitude, Poverty rate and Altitude.
We evaluated 13 models each different and selected the best model based on the highest R-squared value and lowest standard deviation values.
A summary table of all of the models we evaluated is on the following page

modelIndex	Rsq	RsqAdj	Sigma	Fstat	FstatProb
1	2.49E-05	-0.005025452	110.0317	0.004935615	9.44E-01
2	2.97E-04	-0.009852056	110.2956	0.029283943	9.71E-01
3	8.70E-02	0.073067743	105.6704	6.228890179	4.63E-04
4	8.32E-02	0.069181507	105.8916	5.930112615	6.83E-04
5	9.63E-02	0.077557461	105.7712	5.140859523	5.88E-04
6	9.96E-02	0.081141674	105.2091	5.39327608	3.86E-04
7	1.14E-01	0.095808027	104.3661	6.271501501	9.10E-05
8	9.80E-02	0.074488716	105.947	4.171063894	1.28E-03
9	9.71E-02	0.073623518	105.9965	4.13130425	1.38E-03
10	1.66E-01	0.144465485	101.863	7.653080638	1.40E-06
11	1.66E-01	0.148871288	101.6004	9.614338589	4.13E-07
12	1.75E-01	0.153382923	101.3307	8.138158849	5.49E-07
13	1.85E-01	0.15907382	100.9896	7.210918263	5.98E-07

We utilized model 13 as it seemed to be the most accurate.
This model had an R-squared of 0.18 which was best of all of the models. This means that the model was able to account for 18% of all variability in the EPA values. This may not seem significant, however we are dealing with a domain with a lot of variability and outside effectors that cannot be easily measured or accounted for, so getting ~20% of the variability in the expected performance of teams is extremely good.

The summary of the chosen model is on the next page:

```
lm(formula = rowEpa ~ prcp + temp + educationalEmployees +
scientificEmployees + poverty_rate + altitude)

Residuals:
    Min      1Q  Median      3Q     Max
-220.93  -77.51  -11.54   66.22  301.93

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.789e+03  7.698e+01  23.246  < 2e-16 ***
prcp                  2.773e-01  7.810e-01   0.355 0.722920
temp                 -3.141e+00  1.228e+00  -2.557 0.011322 *
educationalEmployees -4.373e-04  6.351e-04  -0.689 0.491921
scientificEmployees   4.216e-04  2.780e-04   1.516 0.131082
poverty_rate         -6.215e+00  1.696e+00  -3.664 0.000321 ***
altitude             -4.798e-02  3.330e-02  -1.441 0.151284
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 101 on 191 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.1847,     Adjusted R-squared:  0.1591
F-statistic: 7.211 on 6 and 191 DF,  p-value: 5.978e-07
```

Error V Streak

The comparison between the Standard error and the consecutive years played allows us to see if the number of consecutive years played, or more accurately, the number of years of experience a team has, correlates with a statistically significant difference in EPA performance. We found that this doesn't.
This has the interesting conclusion that the error when predicting the performance of a team purely based on environmental factors isn't significantly related to the number of consecutive years of experience the team has in FRC, which is surprising since it goes against the common assumption that the more experience the team has in the game, the better at the game they are. This directly is counter to our initial hypothesis that they would only be correlated for the first several years instead of over all the years they participated.

Implications & Limitations

Contributions

This study contributes to FRC as a way for people to visibly see that there is a chance that funding & other environmental factors are related / correlated with higher performance. This does also go to show that it might be beneficial to utilize environmental variables when predicting the true EPA of a team.

This likely isn't going to get a lot of use inside of FRC specifically since oftentimes teams will be looking at more game-specific statistics for alliance selection. Outside of FRC this does give some presidence to using environmental variables in score prediction which could help improve already existing models. Based on the literature review in the intro this avenue has been slightly explored but focusing on a match by match basis instead of team statistics.

Limitations & biases

There are some major limitations and biases that come into play with this study that prevent it from being generally applicable to all teams in FRC or even all US teams. The major bias is that we only looked at teams associated with a public school in the US. This prevents teams that are based out of private schools or were made by independent groups from being considered.

We also only looked at team environment variables, instead of looking at variables that changed depending on the season or match (such as time of day, weather during a specific match, etc). Another major question is whether the results here would be mirrored in other competitive environments, or if the impact of environmental variables is FRC specific.

Addressing research gaps

This study expands upon the research gap originally mentioned in the introduction. Here we show a specific competitive environment where the environment the teams were formed in significantly correlates with their performance regardless of their experience.

We first showed that as the distribution of performance changed as the team progressed through their career. We then showed that an environmental factor (total school district revenue) was not independent from the team's average performance.

We then created a predictive model to look at what environmental variables seemed to correlate the most with the performance of the teams. Once we had a relatively accurate predictive model (which accounted for 18% of the variability in the expected points added) we then took the residuals for each year and compared them to the number of consecutive years the team had played. This acted as a way for us to quantify the experience the team had with FRC at that point.

We found that the differences between the means of the squared residuals based on consecutive years played were not statistically significant. This effectively tells us that the amount of experience a team has doesn't seem to affect a purely environmental based prediction model.

Conclusion

Reflection & Impact

As stated in the intro, we were originally inspired to start this study because we saw several members of teams say that the reason certain teams were successful was because they got more funding or had more mentors. We found positive evidence that this may be true. We want to stress that this is an observational study and so we can only truly look at correlations. We cannot draw any causal conclusions about whether funding directly affects the performance or if teams get more funding because of their performance.

The reason we performed this study as an observational study is because we wanted to see if it would be viable to use environmental variables in score prediction models with FRC as a test case. We have found that, while it's not as accurate as a method based purely on previous year's performance, it could make those models more accurate. Especially in environments where outside forces such as funding or availability of specific people matter more.

This study isn't going to majorly affect the landscape of FRC competition, but it could impact how we look at score predictions and measurements of performance inside and outside of the First Robotics Competitions.

Future next steps

Further research

Here we will identify certain areas to look into the effect of environmental variables in other contexts.

We would recommend expanding this research into other competitive environments. Particularly ones where environmental factors will have a less clear connection to performance. Like soccer, football, and other sports that don't have large expenses.

We could also research using better methods. Here we used an observational study, but it is likely possible to control certain environmental variables to create an experimental study to research the cause & effect relationship between these environmental variables.

More research could be done into the domain of FRC teams, with more robust data collection & collation methods, which would allow one to refine this model to be even more accurate and perform more robust testing.

Applications

As mentioned before one application is using the environment based prediction model in addition to a performance based model, where the environmental based prediction model gives an initial estimate of the team's performance which the prediction model then refines into an even more accurate model than either could be individually.

There are real world application for an improvement to score prediction models, as an example it FRC teams currently do look at future competitions and try to predict what the competition will look like, this can affect decisions around the strategy their robot tries to play in order to appeal to specific alliances or to counter moves they've seen from teams in other competitions. By having a more refined EPA prediction model they are going to have better data backing up those decisions.

Works Cited

Shasha, Wang, et al. “Assessment of Olympic Performance in Relation to Economic, Demographic, Geographic, and Social Factors: Quantile and Tobit Approaches.” Economic Research-Ekonomska Istraživanja, vol. 36, no. 1, May 2022, doi:10.1080/1331677x.2022.2080735.

Owramipur, Farzin, et al. “Football Result Prediction With Bayesian Network in Spanish League-Barcelona Team.” International Journal of Computer Theory and Engineering, Jan. 2013, pp. 812–15, doi:10.7763/ijcte.2013.v5.802.

F. Galbiati, R. X. Gran, B. D. Jacques, S. J. Mulhern and C. -K. Ngan, "An Integrated Framework for Team Formation and Winner Prediction in the FIRST Robotics Competition: Model, Algorithm, and Analysis," 2023 IEEE International Conference on High Performance Computing & Communications, Data Science & Systems, Smart City & Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys), Melbourne, Australia, 2023, pp. 868-876, doi: 10.1109/HPCC-DSS-SmartCity-DependSys60770.2023.00126.

Soetewey, Antoine. “Two-way ANOVA in R.” *Stats and R*, 19 June 2023, https://statsandr.com/blog/two-way-anova-in-r/. Accessed 11 June 2025.