

Introduction

This study had a three purposes: (1) to compare the scouting accuracy of some top Michigan teams; (2) to gauge the overall accuracy of scouting and assess the implications of accuracy on the value of scouting data; and finally (3) to provide recommendations on improving quantitative scouting data collection. There were a few base ways of comparing the teams, and the development of those statistics is explained in the methodology. The teams are presented as case studies of datasets, with only limited information on the methods of data collection and entry. Discussion of overall accuracy is based on the summary statistics and details of the case studies.

Methodology

5 teams' scouting databases for the 2014 Michigan State Championship (MSC) were collected. They were Team 3322 and four elite Michigan teams, let's codename them Red, Blue, Purple, Green, & Orange (don't try to read into the codenames, they're not emblematic, just easier than using numbers or letters).

To look at accuracy, I decided to focus on a single statistic to be representative of the dataset's overall accuracy. I chose Teleop High Goals Made (let's call it Shots), because it's well-defined, collected by all the teams, and visually pretty obvious.

At MSC, there were 128 matches with 6 teams each, or 64 teams with 12 matches each. Either way, that's 768 "team-matches" (e.g. 27 in match 100, 33 in match 6). Ideally, a team will have a Shots value for all 768, but realistically teams may not scout every match, may not enter every match, may enter the wrong team number, or may enter the wrong match numbers.

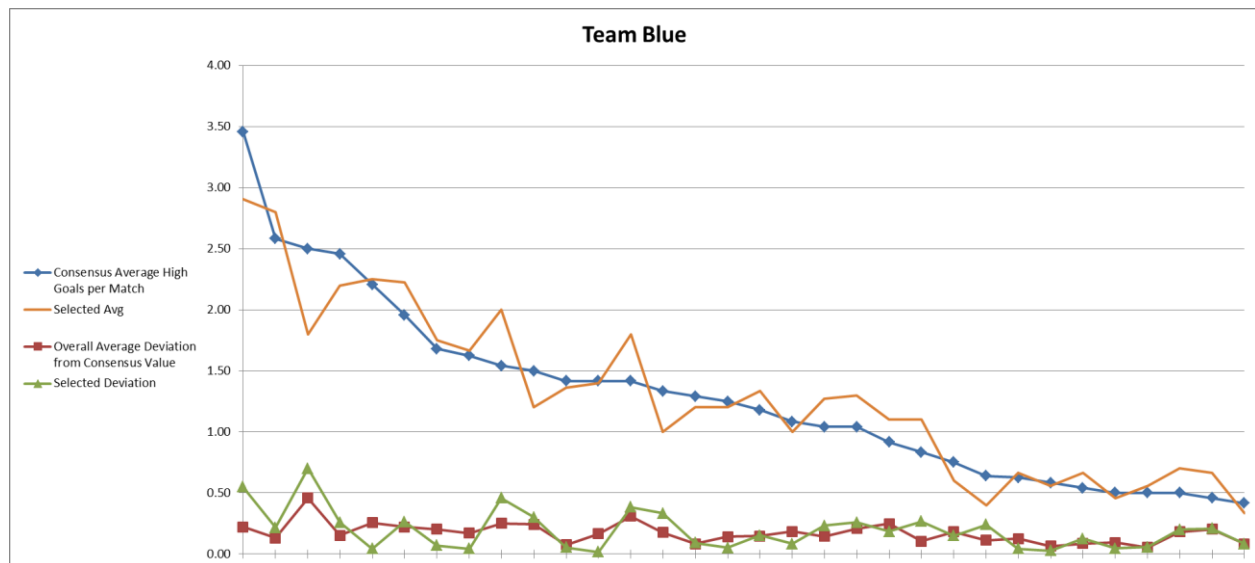
With the data available, I developed a consensus Shots value for each team-match. For the 94% of matches with a single mode, it was that single mode. In 86% of matches this was a majority opinion, with 3 or more teams agreed. For the remaining matches, the consensus was the average of the multiple modes. E.g. for one team-match, the values were 2 2 3 4 4, so the average of 2 and 4 (i.e. 3) was used as a "best guess." The consensus values are as close to the real values as can be had without collecting the data from the FiM videos (and triple-checking).

As a start, each dataset was compared to the consensus data to see how often their raw values varied. This is a decent overview, but it doesn't tell the whole story. The way a team will actually use their data is to look at team averages. So I further compared those averages to the averages that would come out of the consensus dataset.

Team Sections

Team Blue

The graph below shows team averages from Blue's dataset versus the consensus averages. It also shows Blue's deviation from average, versus the typical deviation from average for that team. To be clear, the y-axis is Average Shots per Match and the x-axis is the teams at MSC in descending order of Average Shots per Match as determined by consensus values (only the top 32 teams were taken).



Team Blue is about middle of the road of the teams surveyed. At times they deviate more than the average amount (for that 3rd team in particular), but for the most part they're right around that average deviation. A top 10 list developed by Team Blue would have most of the same teams as the consensus top 10, if in different orders.

Blue is interesting because they didn't collect scouting data on Saturday. With that constraint, their scouting looks pretty good! By MSC, most teams' production is stable, so the assumption that average Shots won't change is alright. However, with their scouting accuracy, Blue could've done much better. Blue had the consensus value 89% of the time, which was 2nd best, but their average deviation from average was 0.132, 2nd worst.

Their data was also hurt by data entry errors. Team Blue had data for 41 invalid team-matches, with 31 coming from invalid matches (e.g. 33 didn't play in match 45) and 10 coming from invalid teams (e.g. team 615 wasn't at MSC). While invalid matches are okay, since the data will still show up in team averages, invalid teams mean the data won't show up at all. The match with 615 entered as the team is data that won't get correctly attributed to team 815.

Team Red/3322

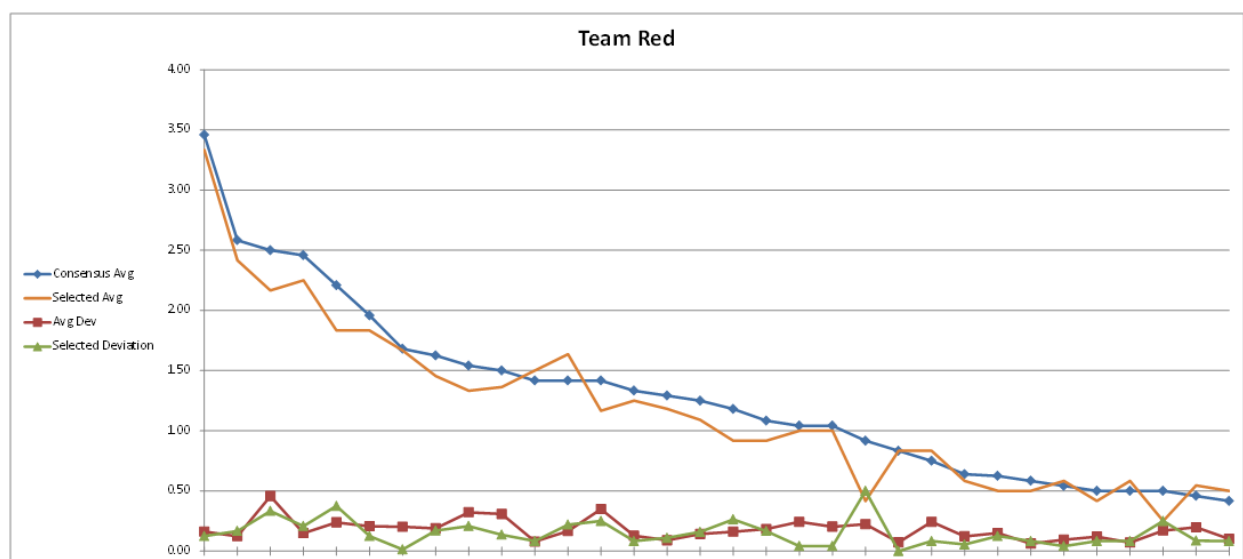
Team Red is the one I know the most about, and that's for a very good reason: it's my team. So I can share many more details about Team Red than any of the other teams. Altogether, I was really pleased with how we compared to the elite teams in this study (but we can and will get better).

The obvious trend from the graph is that we consistently underrated teams (one of the very few teams we overrated: ourselves. Embarrassing.. :P). I have no idea why this happened. Perhaps I need to work with my scouts on paying attention, since we came in last on value errors, only 83% correct, almost double the errors of the best team. We developed a new scouting system between our first event (week 3) and our second event (week 5), with additional changes from that 2nd event to MSC. As a result, many of the scouts weren't entirely familiar with the system. Hopefully, continuity alone will help.

Despite the common errors, our deviation from average wasn't bad, coming in second at 0.109 average deviation from average. Again, I can only guess at the reasons. We had the highest completion rate, with only 15 matches missed; unlike all the other teams, we entered data through the last qual match on Saturday. Additionally, we had 0 match number and team number errors, as a result of my data validation every few hours during the competition.

It's interesting that the % of errors doesn't correspond very well with the average deviation. It seems to indicate that there are two types of errors: those that are random, and those that are the result of (even subconscious) bias. Random errors would be the result of simple mistakes, and will occur similarly for everyone a team scouts. Our underrating must be the result of random bias, since it was so consistent among the teams we scouted. Our overrating of ourselves is a great example of a bias error, since it's extremely unlikely it was simply random (I could break out a p-test, but it's evident).

What I'm saying is, I think we performed well because we have a minimum of bias errors, which made up for the large random errors. Why? We haven't worked closely with other teams that much, particularly not the many elite teams at MSC. This is compared to the elite teams that I polled, who work with the other great teams (and each other) quite often. Perhaps I'm overplaying this, and it's simply all random, but it's one plausible explanation. Post if you have another theory.

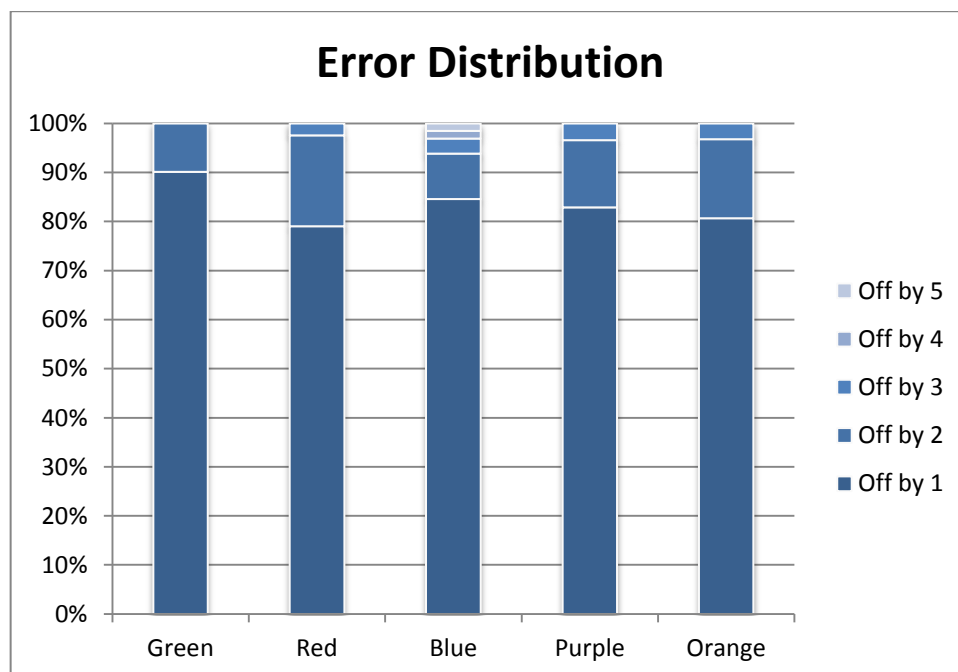
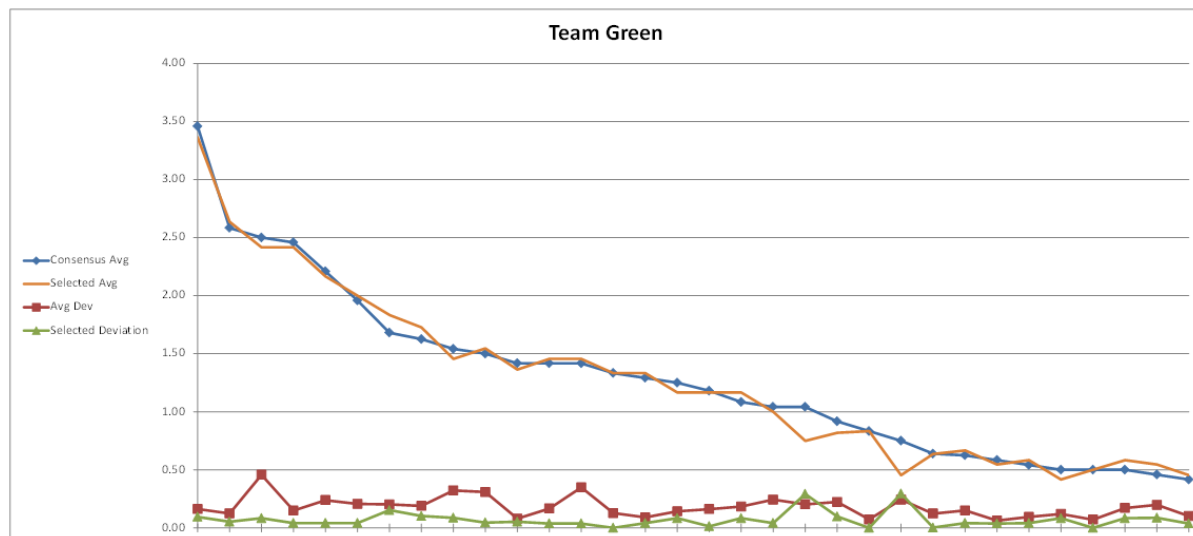


Team Green

There's not a whole lot to say about how well Team Green did. They were the best by far, in every meaningful metric. Only team with over 90% correctness. Other than skipping the last 4 matches of data, they only missed 7 team-matches of data, which was the least, with 0 team number or match number mistakes. When it comes to the most important metric, their average deviation from average was 0.056, which is pretty close to half the next best team. That they were the best is pretty clear looking at the graph below.

How they did this well is certainly worth discussing. There will always be some inherent inaccuracy in scouting, but Green took corrective action. The scouting leads do a "sanity check" before data is entered, and errant scouters have to re-do the scouting for that match using the FiM youtube videos

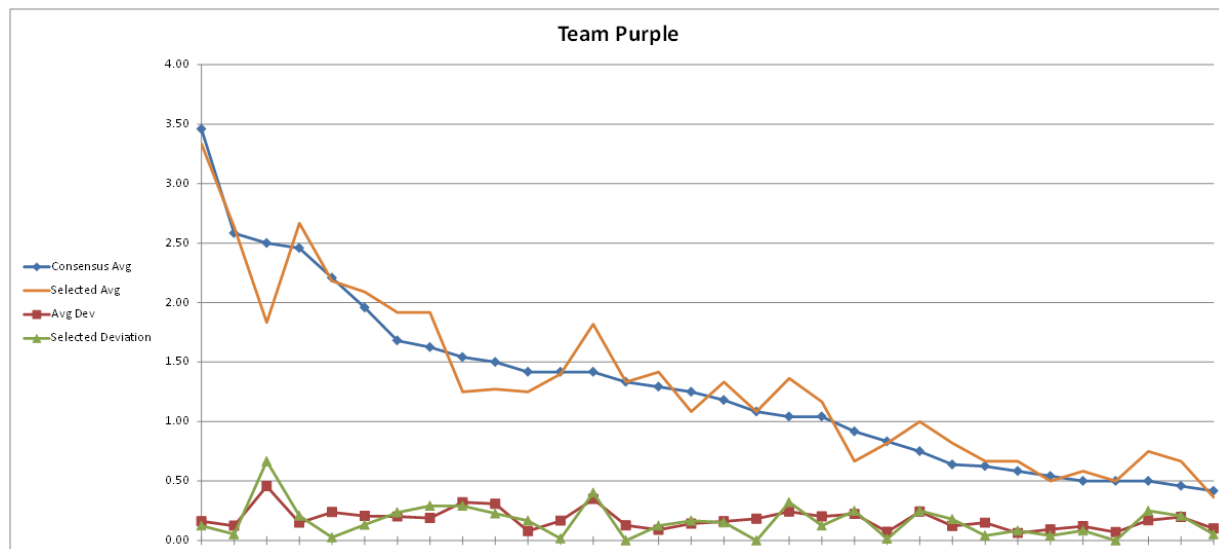
(which are available at all their events because Green is 2337). As a result, their average error size was the lowest of all the teams. Just over 90% of their errors were off by 1, and in zero cases were they off by 3, neither of which were true for any other team.



Team Purple

There isn't much to say about Team Purple either, but for a different reason. They were, statistically, extremely similar to Red/3322. They didn't enter data for the last two matches, and missed 10 other team-matches (though some of those were bad team #s or match #s). 84% correctness and 0.115 average deviation from average are both similar to the other teams and Red in particular. The difference between Purple and Red is that Purple's errors seem less random. Their rankings don't match up with

the consensus rankings particularly well (to cherry-pick, they were the only team that didn't have the consensus #2 team as their #2 team). Overall, Purple seems like a benchmark team: not the best, but not the worst either.

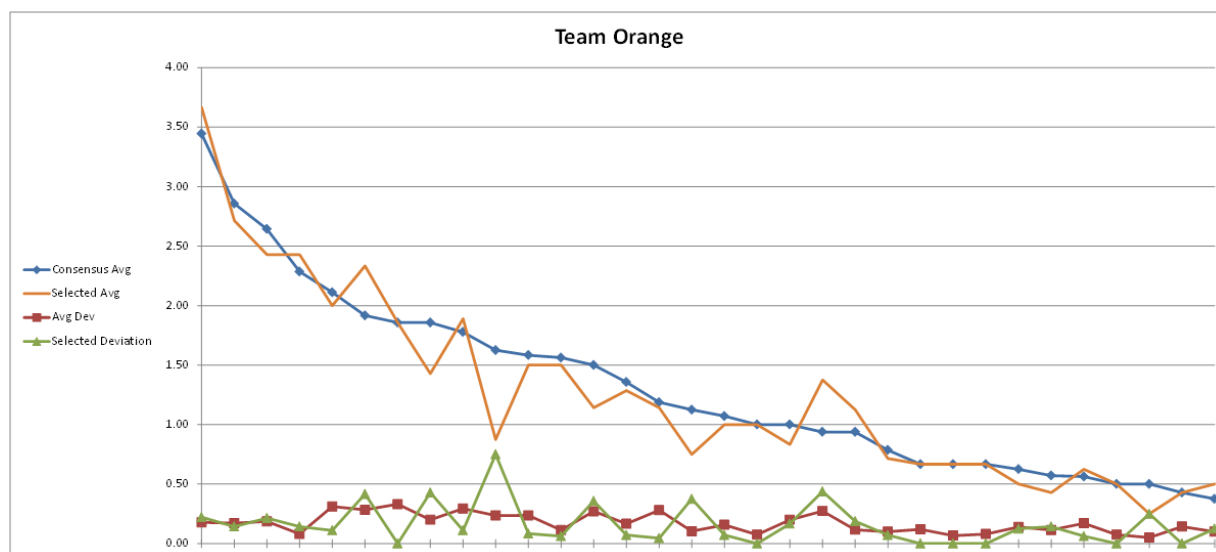


Team Orange

The first thing I have to say about Team Orange here is that they're getting graded on a "curve." Why do they get special treatment? Well, they do their scouting analysis on paper. They do enter their data, but only some of it got entered, with large parts getting left out (matches 81-101, 107-128). I guess I'm lucky they entered some at all! So, I'm playing along here, and working as if the only matches were the matches they entered.

And grading on that curve, they did pretty well! 87% correctness and 0.114 average deviation from average, which would have been good for 2nd and 3rd, respectively. The errors have some randomness, with a little bit of underration, like in Red's case. It's harder to tell with these larger deviations, as the size of the deviations is linked to the smaller sample size. Looking at the graph, there's a sense of there being teams they favoured and disfavoured, but the favoured teams aren't ones that they've historically worked with often, so perhaps it's just the random variation.

I have to say, and this is opinion, I'm personally not a fan of the paper-only system (and not just because it inconveniences my scouting study!). 3322's Week 3 system was paper-only, and while it worked, it forced more subjective analysis simply because really analysing the data was difficult. Yes, this game did lend itself to subjective analysis as opposed to stats-based analysis (as opposed to a 2013), but it's nice to have options. Data visualisation, comparison tools, there are many ways to leverage the computerized scouting system. Elite teams can get away with simpler scouting (and Orange here is a great example), but I think not going computerized is declining a potential advantage.



Conclusion

Findings

Three primary questions were asked here. Which team did the best didn't turn out to be much of a question, with Green/2337 coming in first by a long shot. The 4 remaining teams all ended up in the same rough area of accuracy.

The second question was simply about how good scouting was overall. Every team was between 83% and 91% correctness, so no team made twice as many errors as any other team. 83% is about 5/6, so even the worst team was only wrong for about one team each match. As a result, they were pretty close on the team averages, with an average deviation of about 0.1.

If a team was truly averaging 2.5 Shots per match, and your team's scouting said they were averaging 2.6 or 2.7 Shots, would it make that big a difference? Probably not. On the other hand, these are pretty elite teams, so the average team's scouting won't be quite as good. Altogether, I think teams that put an emphasis on scouting can be confident in the accuracy of their numbers.

Correctness		Avg Dev from Avg	
Green	90.4%	Green	0.056
Red	83.5%	Red	0.109
Blue	89.0%	Blue	0.132
Purple	84.3%	Purple	0.115
Orange	87.0%	Orange	0.114

Several of the meetings mentioned to me that they did not use quantitative data this season to the extent that they have in the past. This makes sense, of course. We played a game where teams aren't performing the same tasks every match and even for the same task, strategy plays a major role in volume and difficulty. For those reasons, this wasn't a good year to do this study. 2013 would have been

perfect (every good team shot many discs in every match), but unfortunately I didn't think of this then. If the 2015 game works better, I may do a Part II.

I also wanted to throw in a sidebar on a common question. Do teams overrate themselves? Sometimes. Of the 5 teams, 2 significantly overrated themselves, 2 significantly underrated, and 1 was pretty close. Not sure I can draw a conclusion here. Maybe scouters aren't biased, maybe some scouters don't scout their own team's matches (underrating) while others do and are biased (overrating), maybe just elite teams aren't so susceptible. Can't say for sure.

Recommendations

If teams want to be really good at scouting, it goes farther than the common things people discuss (picking which qualities to measure, defining objective criteria, etc.). Emphasize to your scouters the importance of scouting (which includes making it clear that you are in fact using the data). Validate that team numbers and match numbers are valid and entered correctly, using the official match schedule. Have educated scouting leads sanity check data as it's being entered. As possible, re-scout matches with clear errors, or at minimum remove that data. These steps will significantly improve the accuracy of any dataset.

Keep in mind that these are recommendations for improving quantitative data accuracy. Altogether, a great scouting system consists of strong data collection and strong data utilization for both qualitative and quantitative data. This study only addresses one of these four aspects, while how to do the other three are left up to the reader at this point.