Groger Ranks 2019-20 Methodology Changes

Steven Liu

November 11, 2019

Contents

1 Introduction

First thing's first: if you've somehow managed to come across this document without knowing what Groger Ranks is or what we do, please check out our website for more details.

Though this isn't meant to be a formal paper or anything like that, it is a fairly deep dive into the inner workings of Groger Ranks. Less technical summaries of many aspects of this are available in a number of analysis articles we've posted over the last year or so.

Throughout this document, the abbreviations P/G and PPB are used as abbreviations for powers per game (i.e. 20 tossups heard) and points per bonus, respectively. In addition, aP/G and aPPB are used to refer to the adjusted versions of those stats.

2 Previous Methodology

This section outlines how adjustments and rankings were calculated by the end of the 2018-9 season, along with some extra commentary and detail that I don't think has been articulated as well as it should. It obviously evolved over the course of that year, but this document will for the purposes of brevity gloss over the changes we made and simply focus on the "final" methodology of the season.

2.1 Power Adjustments

The calculation for power adjustments was fairly simple. The first step is to select a baseline[1] to mark the zero point for adjustments. Every other set is judged based on how teams did in comparison to that baseline.

To take a concrete example, suppose we have a set A we want to compare to our baseline B. We have data on three teams X, Y, and Z, who got the power stats

Table 1: Dummy power stats

Team	A P/G	B P/G	P/G diff[2]
X	4.5	5.3	+0.8
Y	6.1		
Z	3.7	3.5	-0.2

Note that Y did not play the baseline set, so we only consider the performances of X and Z to calculate the adjustment. The adjustment value is simply the average of the difference in P/G between the two sets amongst the teams who played both. In this case, Z found B to be slightly harder than A while X found the opposite, but when taken together, we find that on average A was harder to power than B by about 0.3 P/G. Thus, its power adjustment is +0.3.

This approach can be generalized to any number of sets which share any number of teams, with the one restriction that the baseline must remain fixed.

2.2 Bonus Adjustments

The way we calculate bonus adjustments has the same underlying logic, but with an added layer of math and complexity.

2.2.1 Motivation

Consider the following stats for a number of teams on three sets:

Obviously, this data is incredibly contrived, but it illustrates in an exaggerated manner issues which do arise in real stats.

Suppose by some bad luck that we decided to use B as our baseline[3]. The adjustment with A of +2.4 is pretty robust. C's adjustment of 0 is less robust with just one team to look at.

However, we can pretty easily get some better insight into what C's adjustment would be with more data. Specifically, notice that we have some pretty robust evidence that C is 0.8 PPB easier than A. Combining this with our A adjustment, we can conclude that C should really be something like +1.6 rather than 0, as the naive method suggests.

Or consider these stats for four sets:

[1] Note that the choice of baseline is not arbitrary — the rankings can and do change significantly if you select different baselines!

[2] The stat on B minus the stat on A. This may seem backwards at first glance, but it has the effect of making harder sets have larger positive adjustments.

[3] Why can't we just change the baseline? There are a couple of potential reasons not to, but perhaps the biggest one is that changing the baseline can significantly change the relative adjustments of other sets, making the scores that each team gets and consequently the rankings shift around much more than is necessary.

Table 2: Dummy bonus stats

Team	A PPB	B PPB	C PPB
T_1	20	20	20
T_2	15	18	
T_3	16	19	
T_4	17	20	
T_5	18	21	
T_6	15		16
T_7	16		17
T_8	17		18
T_9	18		19

Table 3: Dummy bonus stats

Team	A PPB	B PPB	C PPB	D PPB
T_1	20	20		
T_2		20	20	
T_3			20	20
T_4	20			20

No matter which set is selected as the baseline, there is one other set which shares no teams with it and thus cannot be adjusted. But notice that, using similar reasoning to before, we can quite sensibly claim that all the sets are equal in difficulty.

All of this data, again, is designed to make my points clear. Regardless, these types of situations are common, and ad hoc logic can only get so far. The task then is to formalize this notion that we should look at every pair of sets to get a complete picture of how hard they are.

2.2.2 "Multiple Baseline"[4] Adjustments

This section is pretty mathematical; a full understanding probably requires at least some knowledge of calculus, although you can certainly get a general understanding of it without any calculus.

Suppose we have $n+1$ sets with adjustments $a_0 = 0, a_1, a_2, ..., a_n$, a_0 being the baseline set[5]. We can calculate how good of a 'fit' these adjustments are to the data by comparing how closely they match the adjustments between pairs of sets[6].

Let the adjustment of set j with i as the baseline, using methods similar to the power adjustment, be a_{ij}. We also have, based on our variables, that the predicted adjustment $\hat{a}_{ij} = a_j - a_i$. Then, the residual for this pair of sets is equal to $a_{ij} + a_i - a_j$. We can square the residuals, and weight based on the size of the shared team population, to get an expression to minimize. Letting p_{ij} be this population, we get

$$R = \sum_{i,j} p_{ij}(a_{ij} + a_i - a_j)^2$$

In order for this to be minimized, the partial derivative with respect to each variable must be equal to

[4]This doesn't really have an official name as far as I'm concerned (hence the quotes), although people seem to have caught on to calling them "Liu adjustments" above anything else. Multiple Baseline is probably more evocative, although it admittedly isn't entirely accurate. If anyone has suggestions for a better name, get in touch!

[5]Unlike before, this baseline is actually (almost) completely arbitrary. The main operative thing is the relative adjustments between sets.

[6]This is in principle similar to least-squares regression.

zero[7]. Neglecting a_0, since that is fixed to be 0, the partial derivative with respect to a_i is

$$\frac{\partial R}{\partial a_i} = \sum_j 2p_{ij}(a_{ij} + a_i - a_j) - \sum_j 2p_{ji}(a_{ji} + a_j - a_i) = 0$$

Because of symmetries[8], this simplifies to

$$\sum_j p_{ij}(a_j - a_i) = \sum_j p_{ij}a_{ij}$$

This gives n equations, one for each of the variables $a_1, a_2, ..., a_n$. Those n equations, combined with the constraint $a_0 = 0$, provide a solvable system of $n + 1$ variables and $n + 1$ equations.

The above derivation assumes that adjustments between sets are purely additive. This appears to be roughly true for bonuses[9], but it certainly is not for powers, which is why this method was only implemented for bonuses and not powers.

2.3 Score Calculation

The calculation of the so-called "Groger Score" from adjusted P/G and PPB comes in two parts. For each team, their performance on each set that they've played is given a score. Then, up to the top three performances are considered together in a weighted average.

2.3.1 Set Performance

Perhaps the most natural interpretation of our old formula is that adjusted powers and bonuses are weighted such that 8 aP/G and 24 aPPB both give 100 points, scores which are averaged together. The exact coefficients can be computed from that, and then negs are accounted for, giving the formula

$$\text{Score} = \frac{25}{4}\text{aP/G} + \frac{25}{12}\text{aPPB} - \text{N/G}$$

2.3.2 Weightings

For teams with multiple stats available, we consider up to three scores given to them. For teams with two scores, their top score is weighted twice as much as the next score. For teams with three scores or more, the top, second, and third scores are weighted 3 to 2 to 1. These average scores are then sorted to give the published rankings.

2.4 Shortcomings

Our methodology as it stood did alright, but there are clear areas for improvement. Two of the biggest are addressed in this season's update. First of all, the coefficients used to calculate a team's performance from their adjusted stats is actually pretty arbitrary, even if the listed criteria for 100 points is aesthetically pleasing. Second, ideally it would be good to bring the power methodology more in line with the bonus methodology, since the flaws of traditional adjustments which brought about the bonus methodology in the first place are still present.

[7] This is equivalent to setting the gradient to the zero vector. This is technically a necessary but not sufficient condition for minimizing R, but either using the trivial inequality or the second derivative test it can be shown that this is in fact a global minimum.

[8] $p_{ij} = p_{ji}$ and $a_{ij} = -a_{ji}$

[9] Insofar as we do not have enough data to conclusively say that bonuses do *not* behave this way, and that the current adjustment method empirically works pretty well.

3 2019-20 Season Changes

Ideas like these have been toyed with internally in the past, but the impetus for implementing them came from two completely unrelated moments. The first, oddly enough, came from a brief conversation in the HSQB discord during which I made some plots of last year's data, which became the basis of some of the graphs in this document. The second was a practical issue, where not enough sets had intersections at the beginning of the 2019-20 season to properly adjust them as we did before.

3.1 Power Adjustments

As mentioned above, an implementation of our bonus methodology onto powers directly is not reasonable[10], since powers are flagrantly not additive. So, the first step is to force powers to behave like bonuses.

3.1.1 Exponent Determination

For some arbitrary function f, we can use f to re-express our data, giving the general form

$$\text{aP/G} = f(f^{-1}(\text{P/G}) + \text{adj})$$

Note that, if $f(x) = x$, this gives our standard adjustments. Furthermore, if $f(x) = e^x$,[11] then we get multiplicative adjustments[12] which has been experimented with before.

Now we need to determine what function to use to make P/G additive. Fortunately, we already have PPBs, whose adjustments we already assume to be additive. Thus, using PPB as a proxy for some abstract notion of "skill," we can determine P/G as a function of skill, which gives us f.

Figure 1 shows every raw statline collected last season, and it's clear that P/G is not linearly related to skill. To determine exactly how they are related, we can plot the data on a log-log plot like in Figure 2. The line of best fit has a slope of approximately 3.5,[13] so $f(x) \approx x^{3.5}$. As it turns out, the exact value of the exponent is not particularly important[14], so for the sake of clarity an exponent of 3 was taken, making $f(x) = x^3$. Then, the power adjustment formula becomes

$$\text{aP/G} = (\sqrt[3]{\text{P/G}} + \text{adj})^3$$

What's more, by replacing PPB with $\sqrt[3]{\text{P/G}}$, the bonus adjustment methodology can be reused to calculate the power adjustments.

[10] It actually does still somewhat improve the accuracy of the rankings, but it outputs some qualitatively very strange results and isn't as good as the final methodology.

[11] Or an exponent with a different base; in this case it doesn't matter.

[12] $\text{aP/G} = e^{\ln \text{P/G} + \text{adj}} = e^{\text{adj}}\text{P/G}$

[13] 3.476 with a standard error of 0.062.

[14] In terms of correlation with PACE finishes, the range of R^2 values for exponents between about 1.3 and 5 is something like 0.003. Noticeable, but small compared to the difference between an exponent of 1 and 3, and especially compared to the impact changing the coefficients has. Given the variance of a single tournament, even one as large as PACE, such small differences may just be random flukes.

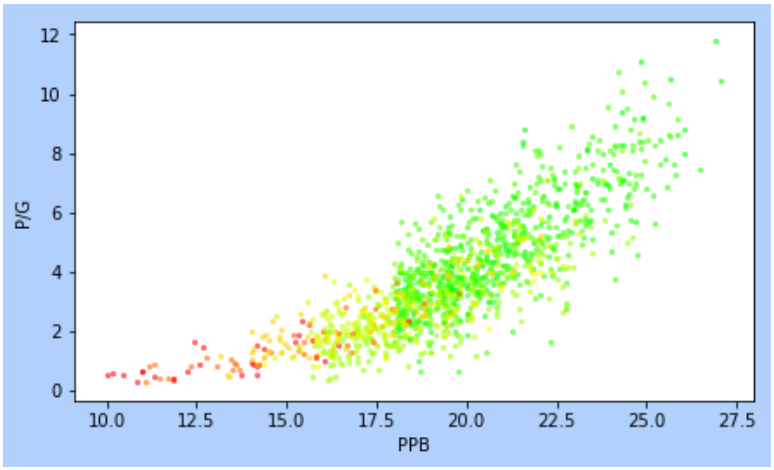

Figure 1: All raw stats collected in the 2018-19 season. Red points are harder sets (e.g. Penn Bowl), while green points are easier sets (e.g. CAST)

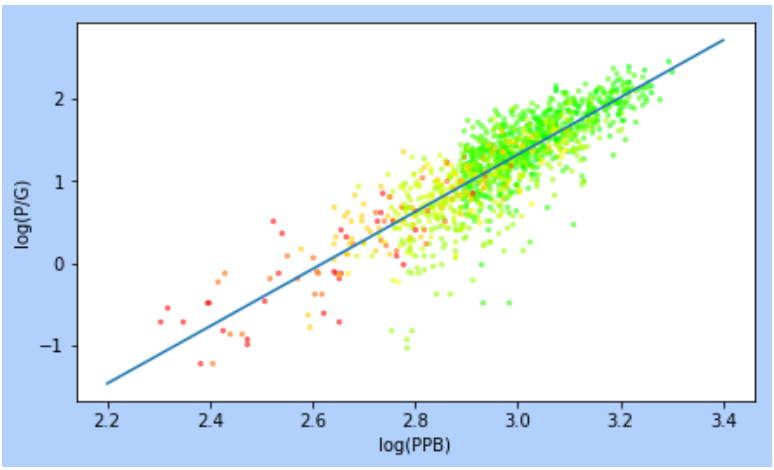

Figure 2: Log-log plot of 2018-19 season stats.

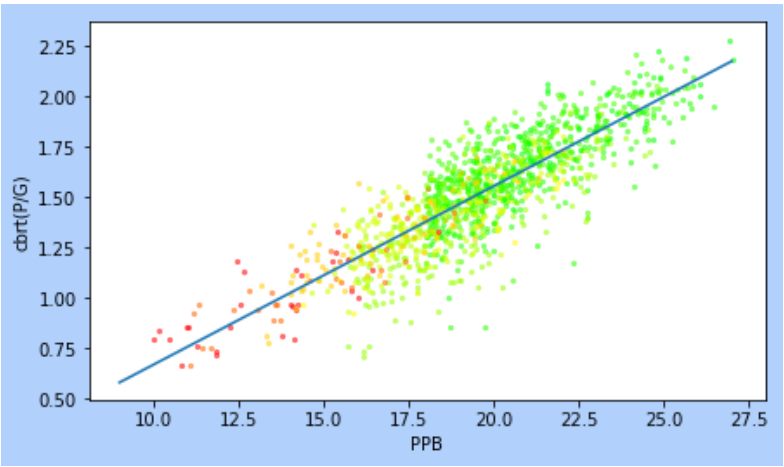

Figure 3: Plot of 2018-19 season stats, with the cube root of P/G on the y-axis.

3.1.2 Inter-Season Comparisons

The output of the new methodology is on a different scale from the old output, almost by an order of magnitude. For the sake of being able to compare this year's adjustments to last year's, I ran the program on last year's data and compared the new power adjustments to the old.

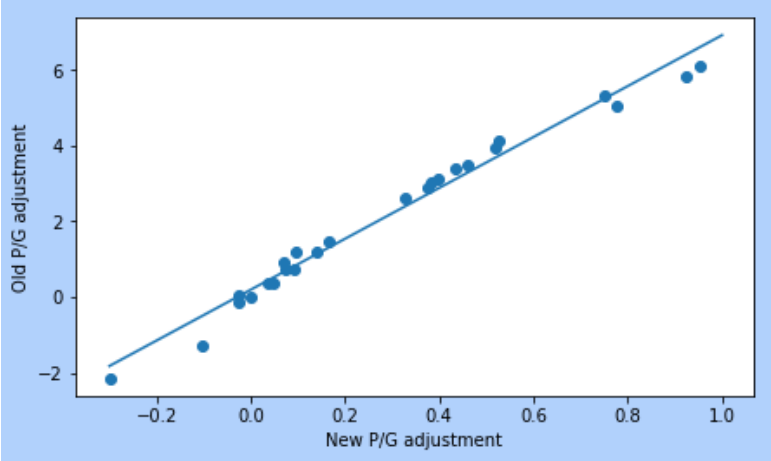

Figure 4: Old and new adjustments for 2018-19 sets.

There probably isn't some deep meaning behind the results. That said, it does give a way to compare adjustments between seasons. Explicitly, the adjustments from the old method are about 6.7 times larger than those in the new method (e.g. a set with +0.1 power adjustment this year is comparable to a +0.67 adjustment set from last year).

3.2 Score Calculation

As previously mentioned, the coefficients for P/G and PPB are somewhat arbitrary, so using the new power adjustment methodology, we can check how changing up the coefficients affects how well PACE NSC finishes are predicted[15].

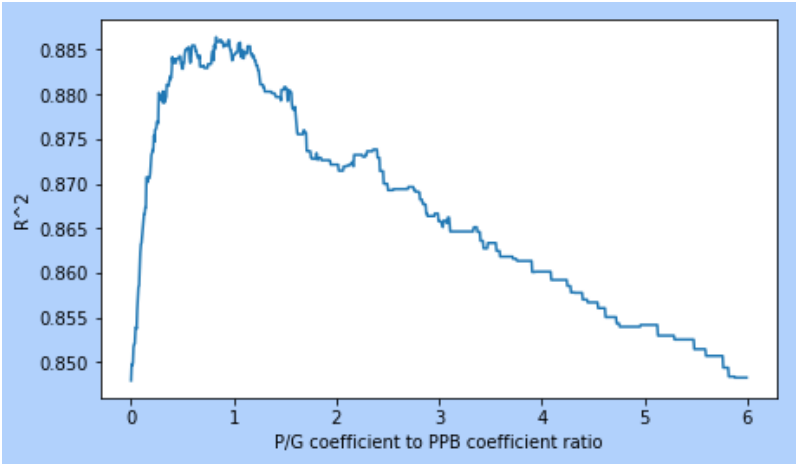

Figure 5: PACE and predicted finish R^2 values vs. coefficients.

Our current methodology in 5 is at a ratio of 3, and considering PPB alone is a ratio of 0. There's a clear indication that decreasing how much powers are weighted makes the ranks predict finishes considerably

[15]The R^2 in the graph is the R^2 between predicted and actual relative ranks, rather than scores and the like.

better.

To keep things comparable with last year, the coefficients are scaled such that a team receiving 8 aP/G and 24 aPPB will get 100 points before negs are counted. The final formula, then, is

$$\text{Score} = \frac{100}{32}((\sqrt[3]{\text{P/G}} + \text{P/Gadj})^3 + (\text{PPB} + \text{PPBadj})) - \text{N/G}$$

3.3 "Superadjusting"[16]

One last tweak we made is that we are now including formerly unusable stats (e.g. PPBs that contain bouncebacks or sets that are not powermarked) for the sake of adjustments. This is fairly straightforward and doesn't involve any big methodological changes. Regardless, we may as well use the stats that are usable, even if the overall reports aren't.

Perhaps the most obvious result of this is the addition of a PPB adjustment for ACF Fall, which a number of high school teams play. Obviously none of these results can be used to generate a score due to a lack of powers, but for curiosity's sake they have been calculated.

4 Possible Future Changes

So, what now? Well, there are a few things we might want to take a look at in the future, which I've briefly commented on below. I don't anticipate doing too much while the season is ongoing; at the earliest there may be a midseason methodology update if things appear to be very out of whack. At any rate, we're collecting even more data than we did last year, so we'll see what fruit that can bear for us for 2020-21.

4.1 Revisiting Bonuses

Throughout this document, there's been the underlying assumption that bonuses are nicely additive. There is, of course, no reason this necessarily must be true. The data we gathered last year (consisting only of teams with 18+ aPPB) is, unfortunately, insufficient to get a full picture of the bonus landscape. Now that we've greatly expanded the scope of data collection, we should be able to pick out what, if anything, the bonus linearization should be.

Of course, if we do end up changing the bonuses, the powers will have to be adjusted as well, since the power analysis is intricately tied to bonus performance.

4.2 Scoring

There are a few aspects to scoring that are still untested, but were less major considerations for this overhaul. If anything does get changed during the season, it will probably be something in this section.

4.2.1 Weighting

The system of weighting we have right now is, like a lot of things, arbitrary but aesthetically pleasing. I'm frankly not sure what form an alternate system would take, or whether it would even stay as a weighted average or change to some other averaging method.

4.2.2 Powers

Something I wrestled with was whether to use the aP/G we have now or the cube root of that in our adjustments. I didn't actually test it, mostly because I'd have to do another round of analysis to find the best coefficients and partly to keep things similar to last year, but there's at least some theoretical value in using the linearized statistic to rank rather than what it is now.

[16] This also doesn't really have a name. Superadjusting is the term we've been using internally, even if it probably makes more sense as a name for Multiple Baseline. Oh well.

4.2.3 Negs

We account for negs in our rankings which I think is probably a good thing overall. Again, though, I'd like to see if the coefficient could be tweaked to improve our model. Maybe we should be punishing negs more or less, or maybe it basically doesn't matter in the end.

4.3 HSNCT Accuracy

I only analyzed how our model did compared to PACE. HSNCT wasn't considered due both to the large amount of data entry that would have to be done for unclear gain and HSNCT having very large ties making the correlation a bit funky. I may return to this in the future, though it isn't necessarily a priority.

5 Final Thoughts

This methodology is hardly perfect, and there's certainly still plenty of space to improve. If you have suggestions on approaches for us to take that I haven't considered[17], feel free to get in touch with me or any of the Groger Ranks staff.

Congratulations on getting to the end of this document. Thanks so much for supporting Groger Ranks, and may you have luck in the rest of the season!

[17] If you're tempted to yell "Machine learning!" at us without any details on how we would implement it, please don't. That said, if you have some expertise in the area and concrete ideas that you can communicate to us clearly, we'd certainly welcome it.