# Notes on RPI

*Erik Forseth*

**The following provides a loose derivation (or at least motivation) for the standard RPI rating used in college basketball. Some generalizations of the rating are discussed and compared.**

## Derivation

Assume that we have season-to-date data for a league of $N$ teams as of a specified date. Denote by $R_i$ the RPI rating for the $i^{\text{th}}$ team. We are to calculate $R_i$ iteratively, so that $R_i$ will be obtained as

$$R_i = \lim_{j\to\infty} R_i^j, \tag{1}$$

for some sequence of approximations $\{R_i^j\}$. The first approximation, $R_i^0$, is simply the team's winning percentage:

$$R_i^0 = \mathrm{WP}_i. \tag{2}$$

In order to obtain the $1^{\text{st}}$-order approximation, we'll add to each team's $0^{\text{th}}$-order rating the average of its opponents' $0^{\text{th}}$-order ratings. We'll continue in this manner, adding to each team's $j^{\text{th}}$ rating the average over its opponents $j^{\text{th}}$ ratings.

This has the obvious disadvantage that it is generally divergent as $j \to \infty$. Are there circumstances under which could this possibly converge? And is there any benefit to the procedure in spite of its general divergence?

In general,

$$R_i^j = R_i^{j-1} + R_{\text{opp avg}}^{j-1}, \tag{3}$$

so convergence implies

$$\lim_{j\to\infty}\left(R_i^j - R_i^{j-1}\right) = \lim_{j\to\infty} R_{\text{opp avg}}^{j-1} = 0. \tag{4}$$

That is, we need the average of each team's opponents' ratings to eventually tend toward 0 as we iterate. This is obviously unsatisfactory.

Nevertheless, let's look in detail at some of these initial steps in order to illuminate what's happening. Say that team $i$ has played $O_i$ opponents. Then

$$\begin{aligned} R_i^1 &= R_i^0 + \frac{1}{O_i}\sum_n^{O_i} R_n^0 \\ &= \mathrm{WP}_i + \frac{1}{O_i}\sum_n^{O_i} \mathrm{WP}_n. \end{aligned} \tag{5}$$

Note that when computing opponent win percentages, we technically ought to exclude games against team $i$. Next,

$$\begin{aligned} R_i^2 &= R_i^1 + \frac{1}{O_i}\sum_n^{O_i} R_n^1 \\ &= \mathrm{WP}_i + \frac{1}{O_i}\sum_n^{O_i}\mathrm{WP}_n + \frac{1}{O_i}\sum_n^{O_i}\left(\mathrm{WP}_n + \frac{1}{O_n}\sum_{j\neq i}^{O_n}\mathrm{WP}_j\right) \\ &= \mathrm{WP}_i + \frac{2}{O_i}\sum_n^{O_i}\mathrm{WP}_n + \frac{1}{O_i}\sum_n^{O_i}\frac{1}{O_n}\sum_{j\neq i}^{O_n}\mathrm{WP}_j. \end{aligned} \tag{6}$$

Schematically, this is

$$R_i^2 = \mathrm{WP}_i + 2\ \mathrm{OWP} + \mathrm{OOWP}, \tag{7}$$

where we've introduced "OWP" to mean the the average of "opponents' win percentage," "OOWP" to be "opponents' opponents' win percentage," and so on. Note that the last term above (OOWP) is actually an average of averages.

Moreover, note that this is (up to an overall factor of 4) the standard formula for RPI. Let's continue.

Next,

$$\begin{aligned}
R_i^3 &= R_i^2 + \frac{1}{O_i}\sum_n^{O_i} R_n^2 \\
&= \mathrm{WP}_i + \frac{2}{O_i}\sum_n^{O_i}\mathrm{WP}_n + \frac{1}{O_i}\sum_n^{O_i}\frac{1}{O_n}\sum_{j\neq i}^{O_n}\mathrm{WP}_j \\
&+ \frac{1}{O_i}\sum_n^{O_i}\Bigg(\mathrm{WP}_n + \frac{2}{O_n}\sum_{j\neq i}^{O_n}\mathrm{WP}_j \\
&\quad + \frac{1}{O_n}\sum_{j\neq i}^{O_n}\frac{1}{O_j}\sum_{k\neq i}^{O_j}\mathrm{WP}_k\Bigg) \\
&= \mathrm{WP}_i + \frac{3}{O_i}\sum_n^{O_i}\mathrm{WP}_n + \frac{3}{O_i}\sum_n^{O_i}\frac{1}{O_n}\sum_{j\neq i}^{O_n}\mathrm{WP}_j \\
&\quad + \frac{1}{O_i}\sum_n^{O_i}\frac{1}{O_n}\sum_{j\neq i}^{O_n}\frac{1}{O_j}\sum_{k\neq i}^{O_j}\mathrm{WP}_k. \qquad (8)
\end{aligned}$$

Again, schematically, this is

$$R_i^3 = \mathrm{WP}_i + 3\ \mathrm{OWP} + 3\ \mathrm{OOWP} + \mathrm{OOOWP}. \quad (9)$$

We can begin to guess what the next terms will look like:

$$\begin{aligned}
R_i^4 =& \mathrm{WP}_i + 4\ \mathrm{OWP} + 6\ \mathrm{OOWP} + 4\ \mathrm{OOOWP} \\
&+ \mathrm{OOOOWP}, \qquad (10)
\end{aligned}$$

$$\begin{aligned}
R_i^5 =& \mathrm{WP}_i + 5\ \mathrm{OWP} + 10\ \mathrm{OOWP} + 10\ \mathrm{OOOWP} \\
&+ 5\ \mathrm{OOOOWP} + \mathrm{OOOOOWP}, \qquad (11)
\end{aligned}$$

$$\vdots$$

$$R_i^j = \sum_{n=0}^{j}\binom{j}{n}\big(\text{opps}^n \text{ win pcg.}\big), \qquad (12)$$

where $n = 0$ indicates the team itself, $n = 1$ is one level of opponent-averaging, etc. There are at least two issues with equation (12). For one, it is clearly divergent as $j \to \infty$, as we already knew. However, we can obtain finite ratings by truncating at finite $j$ (such as $j = 2$ for standard RPI). The second, and perhaps bigger issue, is that as $j$ increases, we shift higher and higher weight to deeper levels of opponent averaging. Conceptually, this may not be the right thing to do. In fact, it hardly makes sense to go beyond $j = 5$ to begin with, because the opponent average terms tend to achieve relative convergence to within $< 1\%$ after only about 5 iterations.

# Generalizations

Let's consider two convergent variations on the above outlined procedure. The first is due to Andrew Dolphin, and it removes the recursiveness from the equation. At $0^{\text{th}}$-order, each team's rating is

$$R_i^0 = \mathrm{WP}_i - \mathrm{WP}_{\text{league avg}}. \qquad (13)$$

At each step in the iteration, we add to (13) the average of the team's opponents' ratings. The key difference here is that as we iterate, we do not update the previous value in the routine, we update just the quantity (13). The end result is that each team's rating is (13) plus a term which encapsulates all the opponent corrections.

Alternatively, we might simply force the routine to converge by ensuring that successive corrections are decreasing in overall magnitude. Consider, for example, a rating constructed as in equation (3), but with the following modification:

$$R_i^j = R_i^{j-1} + (1/2)^j R_{\text{opp avg}}^{j-1}. \qquad (14)$$

The powers of 1/2 will tend to get mixed up against the binomial coefficients, but the first few iterations will go as:

$$R_i^1 = \mathrm{WP}_i + 1/2\ \mathrm{OWP}, \qquad (15)$$

$$R_i^2 = \mathrm{WP}_i + 3/4\ \mathrm{OWP} + 1/8\ \mathrm{OOWP}, \qquad (16)$$

$$R_i^3 = \mathrm{WP}_i + 7/8\ \mathrm{OWP} + 7/32\ \mathrm{OOWP} + 1/64\ \mathrm{OOOWP}, \qquad (17)$$

$$\begin{aligned}
R_i^4 =& \mathrm{WP}_i + 15/16\ \mathrm{OWP} + 35/128\ \mathrm{OOWP} \\
&+ 15/512\ \mathrm{OOOWP} + 1/1024\ \mathrm{OOOOWP}, \qquad (18)
\end{aligned}$$

etc.

How do these variations compare as predictors of game outcomes (win vs. loss, margin of victory)?

Using a dataset of about 16,000 non-neutral court games played over the past seven seasons, and using the same 0.8/0.2 training/test split, we fit linear regression formulas for predicting the margin of victory (MOV), using as input only the home/away teams' RPIs. We did this fitting for each of the RPI implementations described above. We find:

- Standard RPI (equation 7): 68.7% W/L, 12.13 RMSE

- Generalized $\text{RPI}_4$ (equation 10): 64.9% W/L, 13.05 RMSE
- Dolphin RPI: 71.2% W/L, 11.31 RMSE
- Geometric RPI (equation 14): 71.3% W/L, 11.27 RMSE

We draw several conclusions. The generalized binomial RPI performs much worse than the standard version. As we suspected, this probably has to do with injecting noise via overvaluing deeper levels of opponent averaging, and undervaluing the team's winning percentage as well as the first one or two levels of opponent averages. However, understanding standard RPI in the context of this procedure suggests a more efficient means of computing standard RPI. At face value, one needs to evaluate OWP and OOWP for every team. This is $\mathcal{O}(n^2)$ in the number of opponents $n$. Say we have 300 teams, and each team has played $\sim 20$ opponents, and we already know the winning percentages for each team. This means we need $\sim 400$ extra operations per team, or $120,000$ extra operations in total, in order to compute each team's RPI in the traditional way. In the iterative procedure, on the other hand, there are two successive steps in which we average over each team's opponents, for $2 \times 300 \times 20 = 12,000$ operations. In this example we have a factor ten speedup, but in general it will be $n/2$ in the average number of opponents $n$ each team has played. This can be quite useful when generating data using entire seasons for model-fitting. The two convergent RPI variations we've considered are clearly superior to the standard RPI, and boast essentially equal performance. These results have implications for schedule-adjusting other statistics besides win percentage, with some caveats (statistics like FG% are inherently different, and one likely wants to adjust such a thing using a corresponding measure of opponent defense).