RANKING METHODS FOR OLYMPIC SPORTS: A CASE STUDY BY THE U.S. OLYMPIC COMMITTEE AND THE COLLEGE OF CHARLESTON

PETER GREENE, STEPHEN GORMAN, ANDREW PASSARELLO[1], BRYCE PRUITT [2], JOHN SUSSINGHAM, AMY N. LANGVILLE[*], AND PETER VINT[£]

Abstract.

This paper develops a method to determine if the United States' Olympic sports program is improving on an individual sport basis. In addition, the probabilistic likelihood of a U.S. Olympic sports team reaching the medal podium was investigated. The research presented in this paper focuses on U.S. Men's Ice Hockey (a head-to-head sport), with additional research done on Women's Alpine Skiing (a multi-competitor sport).

Three ranking methodologies were employed for men's ice hockey: Massey, Elo, and Microsoft's Trueskill. Within these methodologies, information accumulated from Olympic tournaments and world championships (2004 to 2013) were applied. The subsequent results allowed a definite determination in regards to improvement of the U.S. team in respect to other competitors.

A Monte Carlo simulation was run for ice hockey in the format of the 2014 Olympic Games. Comparison of a known distribution (normal and uniform) to the generated ratings resulted in a predictive measure for the U.S. to medal. Overall, it was found that ranking methodologies employed throughout this research were considerably accurate when compared to actual results of the 2014 Games.

Key words: Massey method, Elo method, TrueSkill method, ranking vector, rating vector, Monte Carlo simulation

1. INTRODUCTION

Dr. Peter Vint, Senior Director of the U.S. Olympic Committee's Center for Competitive Analysis, Research, & Innovation provided the following project to the problem-based learning class, Operations Research, at the College of Charleston.

- *On a sport by sport basis, how competitive is the United States at any point in time? Is the U.S. getting better, staying the same, or getting worse over time?*
- *What is the likelihood that the U.S. will reach the medal podium in any given competition?*

While an Olympic athlete or team may improve with respect to competition-independent measures (e.g., improved distance thrown in the shot put or improved team gymnastics scores), in order to medal, the athlete or team must improve relative to its competitors. Thus, in our analysis we focused on improvements in U.S. **rankings** relative to the competition. A ranking is an ordered list where the ordering signifies the skill levels of the various competitors. Specifically, each

[1] Presenting author and can take any questions directed towards the paper.

[2] Presenting author and can take any questions directed towards the paper.

[*] Department of Mathematics, College of Charleston, 175 Calhoun Street-RSS 339, Charleston, SC 29401, USA. All authors but the last are affiliated with the College of Charleston. The research of A.N.L is supported in part by NSF grant CISE-CCF-AF-1116963. email: langvillea@cofc.edu

[£] United States Olympic Committee, Competitive Analysis, Research & Innovation, 1 Olympic Plaza, Colorado Springs, CO 80909 competitor is given a rating, or numerical score, based on that athlete's performance relative to its competitors. After all the athletes have been scored, these ratings are ordered, creating an overall ranking. Many methods exist for rating athletes or teams, each generating its own ranking. Which ranking method is best becomes a philosophical question: best by what measure? Accuracy with respect to prediction of future events? Accuracy with respect to past events? Accuracy with respect to some other measure? We provide some answers to these questions at the end of Section 2.

Dr. Vint and the U.S. Olympic Committee provided us with data on several sports; however, to narrow the focus, in this paper, we only present our results for men's ice hockey and women's downhill skiing. These two sports are representative of the analysis that can be conducted on other sports because they cover the two main classes of Olympic sports: (1) head-to-head sports such as team sports and (2) multi-competitor sports such as track and field and skiing.

Ice hockey is a head-to-head sport because each game takes place between two teams, whose overall placement in an Olympic event is decided upon a series of binary games played against opponents. Contrast this with multi-competitor sports in which the speed at which one races, or the length one throws or any number of metrics for success, is directly compared to every other individual competing in the same event. There are many methods for ranking head-to-head sports. See, for example, those in [1]. On the other hand, there are few methods for ranking multi-competitor sports.

We present line graphs of the rank of U.S. over time to answer the first question: is the U.S. getting better in a particular sport? Head-to-head sports are covered in Section 2 and multi-competitor sports are covered in Section 3. In Section 4, we present the results of Monte Carlo simulations to determine the probability that the U.S. will medal in a particular sport.

2. RANKING HEAD-TO-HEAD SPORTS

This section presents the three methods that we analyzed for ranking head-to-head Olympic sports: the Massey method, the Elo method, and the head-to-head variant of the TrueSkill method.

2.1. The Massey Method. The Massey method for ranking items was created by Kenneth Massey in 1997 [2]. Massey had in mind traditional sports teams, such as basketball and football, when he developed his model. But his method has since been applied also to individual sports such as tennis and bowling. Thus, the Massey method applies naturally to head-to-head Olympic sports. Massey's model revolves around the rule that the difference in the ratings of two teams *i* and *j*, denoted $r_i - r_j$, represents the point differential in a matchup of these two teams. The Massey method can be succinctly summarized with one linear system

$$\mathbf{Mr} = \mathbf{p} \qquad \textbf{(1)}$$

where the Massey coefficient matrix **M** is defined

$$\mathbf{M}_{ij} = \begin{cases} t_i & i = j, \\ -n_{ij} & i \neq j, \end{cases} \qquad \textbf{(2)}$$

where t_i is the total number of games team *i* played and n_{ij} is the number of games teams *i* and *j* played against each other. The Massey right-hand side vector **p** is a vector of cumulative point differentials. That is, p_i is the total number of points team *i* scored on all opponents minus the total number of points opponents scored against team *i*. It can be proven that **M** is singular since *rank*(**M**) = *n* − 1. As a result, an adjustment is made to ensure nonsingularity. Any row of **M** is replaced with a row of all 1s and the corresponding entry in **p** is set to 0. This new constraint forces the ratings to sum to 0. Following Massey's advice, we applied this nonsingularity adjustment to the last row, creating an adjusted linear system which we denote $\mathbf{Mr}^{-} = \mathbf{p}^{-}$.

Figure 2.1 shows the Massey rankings of the U.S. plus five other countries[1] as well as the actual placement of the U.S. from 2004 to 2013. These line graphs illustrate how well the U.S. is doing relative to the 'best' competition and can be easily produced for any Olympic sport.

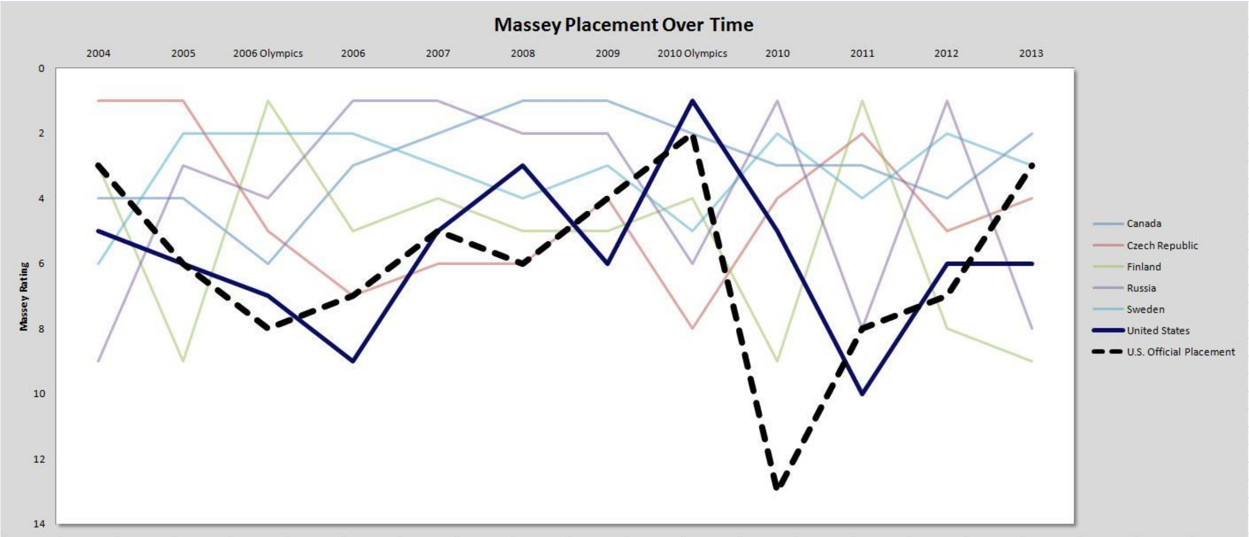


Fig. 2.1. *Line graph of the Massey rank of U.S. Mens Hockey vs. competitors over 2004-2013*

2.2. The Elo Method. The Elo Method, as discussed in [1], is a ranking methodology characterized by the following equation:

$$\mathbf{r_i(new) = r_i(old) + k(S_{i,j}} - \mu_{\mathbf{i,j}}) \qquad \mathbf{(3)}$$

For each competitor with a subheading *i* in the characteristic equation, one's rating upon playing against an opponent *j* is updated by adding a multiple of the difference from the binary result of the game and the expected result. In this ranking methodology,

$$\mu_{\mathbf{i,j}} = \frac{\mathbf{1}}{\mathbf{1} + \mathbf{10}^{\frac{-\mathbf{d_{i,j}}}{\xi}}} \qquad \mathbf{(4)}$$

$$\mathbf{d_{i,j} = r_i(old) - r_j(old)} \qquad \mathbf{(5)}$$

are determined before the game starts. $D_{i,j}$ refers to the difference between the pre-match ratings of *I* and *j*, while $\mu_{i,j}$ is an estimate of the probability of team *I* winning a match against team *j*. The variable

$$S_{i,j} = \begin{cases} 1 & i \text{ beats } j, \\ 1/2 & i \text{ and } j \text{ tie}, \\ 0 & j \text{ beats } i. \end{cases} \qquad \mathbf{(6)}$$

[1] We compared the U.S. to all other competitors, however to avoid cluttering the graph, we only included countries that won a gold medal in a world championship during the period under consideration.

takes on a different value depending on the result of the match in question. Finally, ξ is a constant determining the spread of the ratings, and the value $\mathbf{k}$ determines the weight of an individual game. Both of these values are determined by the user and are only important in relation to each other. For instance, their ratio should be manipulated in order to not artificially inflate a team's rating with respect to the average. The values ξ and $\mathbf{k}$ are usually set and then forgotten about, with one notable example: if one possesses sufficient evidence that victories in certain circumstances are more valuable than others. This allows $\mathbf{k}$ to be manipulated so to award more points to those particular wins.

Elo focuses on rewarding a team based on the opponent's skill level. Teams will rise in rank more quickly if they beat teams with a higher Elo rating. Similarly, teams will fall in rank faster if losing to an opponent with a far worse Elo rating. In general, each team is given an arbitrary starting point, usually at rating zero. The main benefit of the Elo method is being able to update the ratings after each individual match, making it ideal for following along with rankings on a day-to-day basis. This is one reason why Elo is used by FIDE (Federation Internationale des chees) for chess and FIFA (Federation Internationale de Football Association) to soccer.

The following graph shows the United States ranked in men's ice hockey over time along with the same five other countries. The results from Figure 2.2 were found by using a constant $\mathbf{k}$ value. However, several other results were procured by changing $\mathbf{k}$ throughout a single tournament. For more information on these results, feel free to contact the authors.

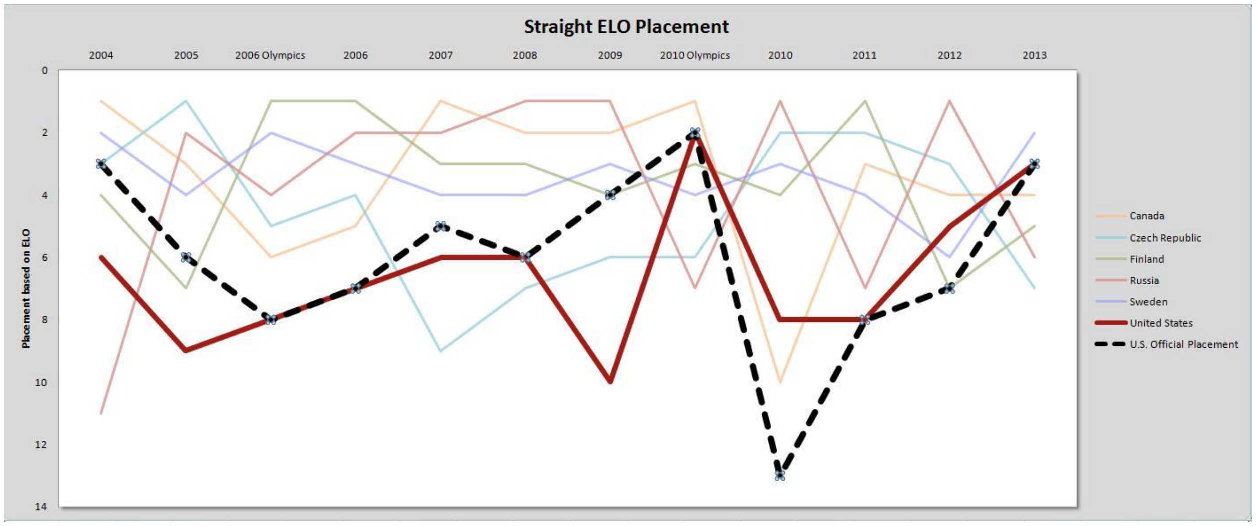


Fig. 2.2. *Line graph of the ELO rank of U.S. Men's Hockey vs. competitors over 2004-2013*

2.3. The TrueSkill Method. Microsoft TrueSkill refers to the proprietary, patented system of ranking players of multi-competitor computer games. It is used by the XBOX Live leaderboards. For more information on Trueskill, see [3]. TrueSkill is a Bayesian method. Parameters of the model are estimated using two values: a mean, μ, and a variance, σ^2. Contrast this with the traditional statistical approach of using one value, the mean. Thus, this Bayesian approach accounts for the variability of the sampling itself, which is helpful for modeling sports because the skill level of an individual's performance typically varies from game to game. The tradeoff is that the implementation of the Bayesian TrueSkill method is more complicated than that of the Massey or Elo methods. While the TrueSkill method was originally devised for and shines in multi-competitor situations, it can also be applied to head-to-head situations. Head-to-head events are simply multi-competitor events with just two competitors. However, the fundamental philosophy between the two are different. The advantage of using TrueSkill for head-to-head events is the additional information obtained about the variability of each rating.

In the TrueSkill method, each competitor starts with the same mean and standard deviation (e.g., μ = 25, σ = 8.3). Similar to the Elo method, TrueSkill updates after every match according to the following set of update equations for both the winner w and loser l:

$$\begin{aligned}
\mu_w &= \mu_w + \frac{\sigma_w^2}{c} * v(\frac{\mu_w - \mu_l}{c}, \frac{\epsilon}{c}) \\
\mu_l &= \mu_l - \frac{\sigma_l^2}{c} * v(\frac{\mu_w - \mu_l}{c}, \frac{\epsilon}{c}) \\
\sigma_w^2 &= \sigma_w^2 * [1 - \frac{\sigma_w^2}{c} * w(\frac{\mu_w - \mu_l}{c}, \frac{\epsilon}{c})] \\
\sigma_l^2 &= \sigma_l^2 * [1 - \frac{\sigma_l^2}{c} * w(\frac{\mu_w - \mu_l}{c}, \frac{\epsilon}{c})] \\
c^2 &= 2\beta^2 + \sigma_w^2 + \sigma_l^2 .
\end{aligned} \quad \textbf{(7)}$$

where

$$v(t, \alpha) = \frac{N(t - \alpha)}{\Phi(t - \alpha)}, \; w(t, \alpha) = v(t - \alpha) * [v(t - \alpha) - (t - \alpha)], \quad \textbf{(8)}$$

and $N(t - \alpha)$ refers to the standard normal probability density function evaluated at $t - \alpha$, and $\Phi(t - \alpha)$ refers to the standard normal cumulative density function evaluated at $t - \alpha$. Each player's skill can be thought of as having a normal distribution itself. The Empirical Rule implies that for any skill level represented by $Normal(\mu_i,\sigma_i^2)$, 68% of the individuals performances will fall within the region $(\mu_i - \sigma_i, \mu_i + \sigma_i)$, 95% of the performances will fall in the region $(\mu_i - 2\sigma_i, \mu_i + 2\sigma_i)$, and 99% of the performances will fall in the region $(\mu_i - 3\sigma_i, \mu_i + 3\sigma_i)$. In other words, 99% of the individual's performances will occur at a skill level above $\mu_i - 3\sigma$. Microsoft uses this value, called a 'Conservative Ranking', as their Xbox Live leaderboard rating for a particular player [3].

Figure 2.3 shows Trueskill rankings for the United States Men's Ice Hockey team over time along with the same five other countries used in the Massey and Elo methods.

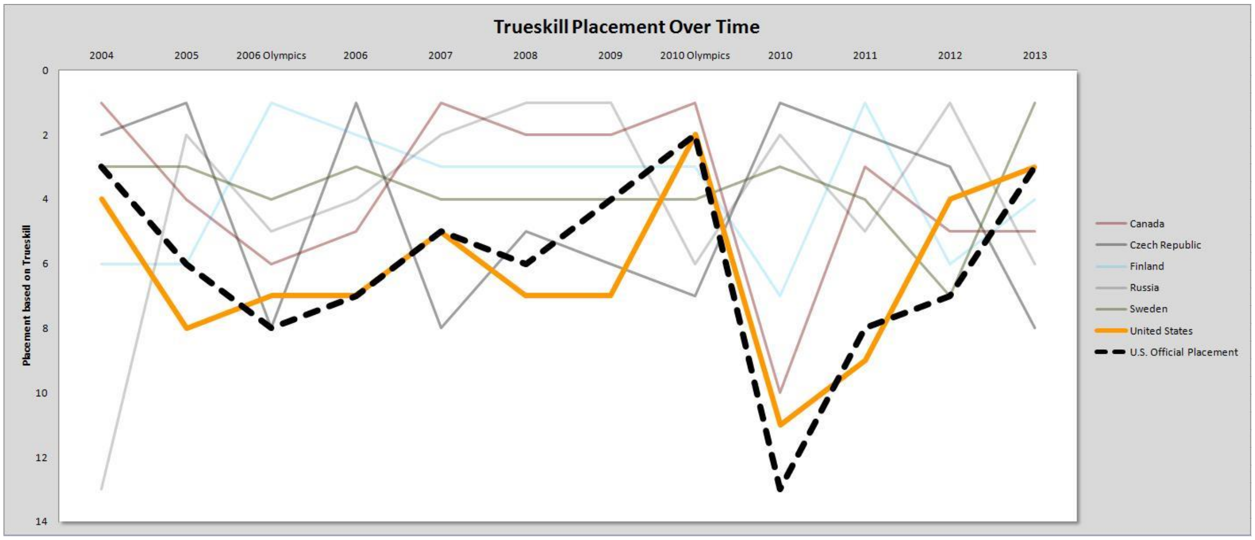


Fig. 2.3. *Line graph of the Trueskill rank of U.S. Men's Hockey vs. competitors over 2004-2013*

As seen below, Figure 2.4 displays all three methods simultaneously along with the official U.S. placement. One can observe that the lines are not drastically different from the official placement. The most one could say is that the Trueskill rankings appear to resemble the official placements the most, but a simple visual observation cannot definitively determine if any method is superior to the others.

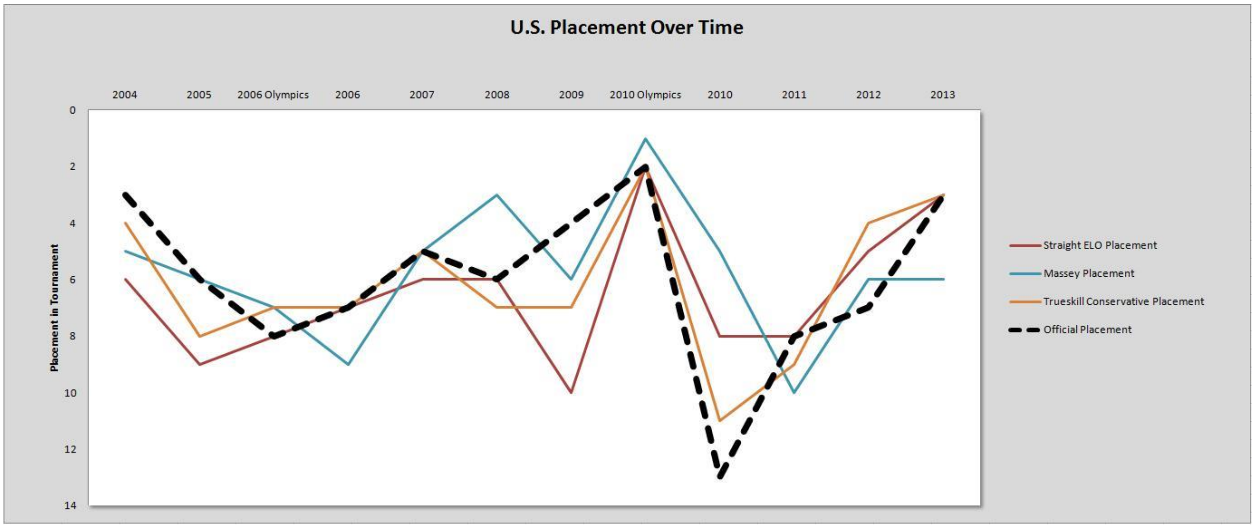
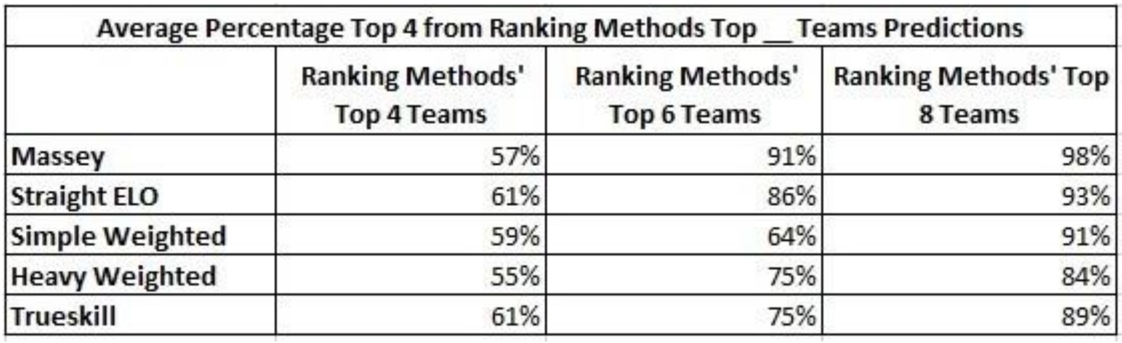


Fig. 2.4. *Comparison of all three methods*

2.4. Comparison of the Methods.

2.4.1. Direct Prediction Using Prior Results. In order to find some accuracy for which teams would advance to the semi-final of a tournament, we focused on a team's success from the previous year. The thought of doing such measure was two-fold: first, by seeing how successful a team is to advance based from the previous year served as a decent metric for the accuracy of each ranking system. Secondly, and more importantly, we wanted to find a threshold below which we no longer had to consider teams as a threat to medal. Figure 2.5 shows the average predictive power for each method and threshold. Note that 'Simple Weighted' and 'Heavy Weighted' are two forms of Elo found by not using a constant **k** value.



Average Percentage Top 4 from Ranking Methods Top __ Teams Predictions

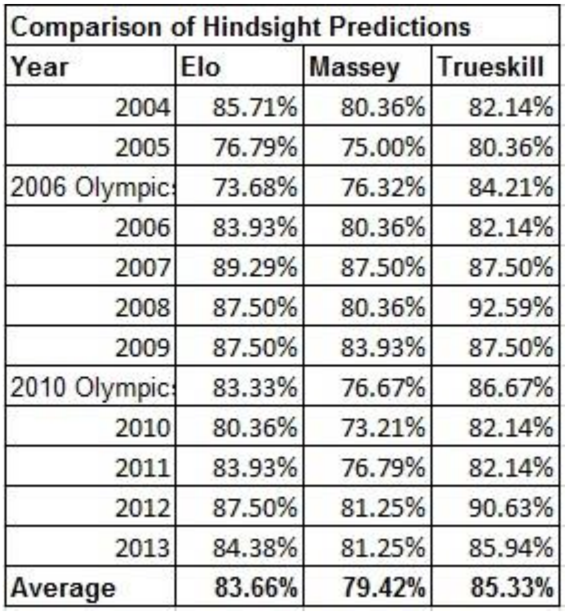
	Ranking Methods' Top 4 Teams	Ranking Methods' Top 6 Teams	Ranking Methods' Top 8 Teams
Massey	57%	91%	98%
Straight ELO	61%	86%	93%
Simple Weighted	59%	64%	91%
Heavy Weighted	55%	75%	84%
Trueskill	61%	75%	89%

Fig. 2.5. *Top 4 Predictions*

For men's ice hockey, these results show that by using the top six teams from the previous year on average results in a 91% level of accuracy (based on the Massey method) at finding the top four teams in the current year. Also, the Elo method applied with a constant **k** value (Straight Elo) fairs better than when not holding **k** constant. If one considers just Straight Elo, along with Massey and Trueskill, one will observe that Trueskill produces the worse predictions for using the top six teams in the previous year to predict the top four in the following year. This is particularly interesting since the line graphs for Trueskill appeared to reflect the official placement for the U.S. the most. This reiterates the point that visual observations can be misleading. Even so, when the threshold is extended to eight teams, one will see that the prediction results are moderately better. This leads us to believe that looking at the top six teams from the previous year is sufficient for this sport. Therefore, Figure 2.5 puts an emphasis on consistency as a standard for achievement. If a team consistently places in the top six, it stands to reason that for the following year, that team will be competitive for medaling.

2.4.2. Hindsight Predictions. One of the later goals in this project was to quantify the accuracy of each ranking method in order to verify its efficiency. To achieve this goal, the ratings from these ranking methods are used to determine the outcome of all the games associated with the rankings. The higher rated team would be deemed the winner. These outcomes were then compared to the actual results. For our research, the ratings accumulated at the end of a year would be used to determine the outcomes of those games. This method is known as hindsight accuracy [1].

Figure 2.6 displays a comparison of all three methods. From this table, Trueskill has the best hindsight predictions with an average percent of 85.33%. Thus, the year by year ratings on average from Trueskill are better than Elo and Massey for predicting the actual outcomes of games for individual years. Also, the minimum percent correct for Trueskill is 80.36%, whereas for Elo and Massey, they are 73.68% and 73.21%, respectively. Since we are focusing on the Olympics, Trueskill also has the highest percentages for the 2006 and 2010 Olympic Games. This could be due to the structure of Trueskill by accounting for the variability a team's skill level.

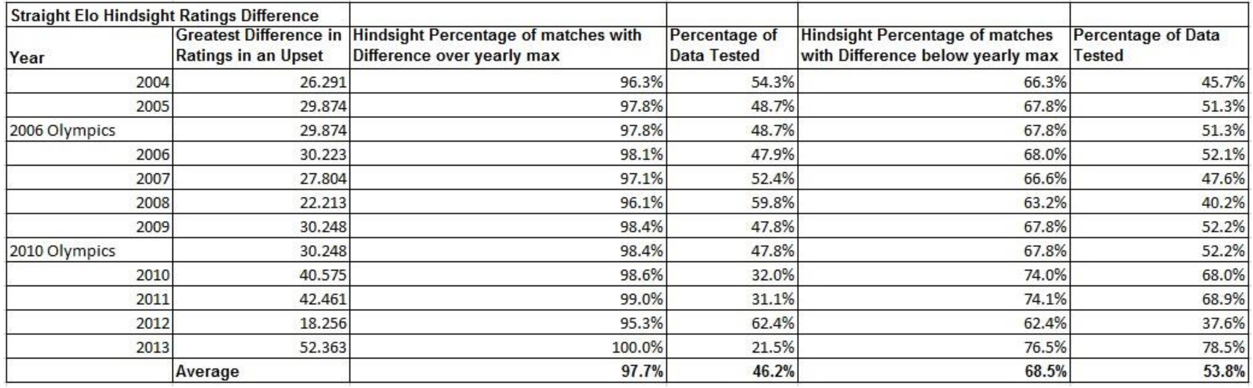


Comparison of Hindsight Predictions			
Year	Elo	Massey	Trueskill
2004	85.71%	80.36%	82.14%
2005	76.79%	75.00%	80.36%
2006 Olympic	73.68%	76.32%	84.21%
2006	83.93%	80.36%	82.14%
2007	89.29%	87.50%	87.50%
2008	87.50%	80.36%	92.59%
2009	87.50%	83.93%	87.50%
2010 Olympic	83.33%	76.67%	86.67%
2010	80.36%	73.21%	82.14%
2011	83.93%	76.79%	82.14%
2012	87.50%	81.25%	90.63%
2013	84.38%	81.25%	85.94%
Average	**83.66%**	**79.42%**	**85.33%**

Fig. 2.6. *Comparison of all three methods hindsight accuracy*

Figure 2.7 below gives one access to data found by using the Elo method with a constant **k** value. This table illustrates the biggest upset (based on the difference of ratings) from each season, which characterizes a threshold above which the ranking method is statistically perfect. We identified the max differences from each year and compared these to the rest of the data. This was done to see how much of the entirety of the data rested on each side of these differences. The smallest difference was

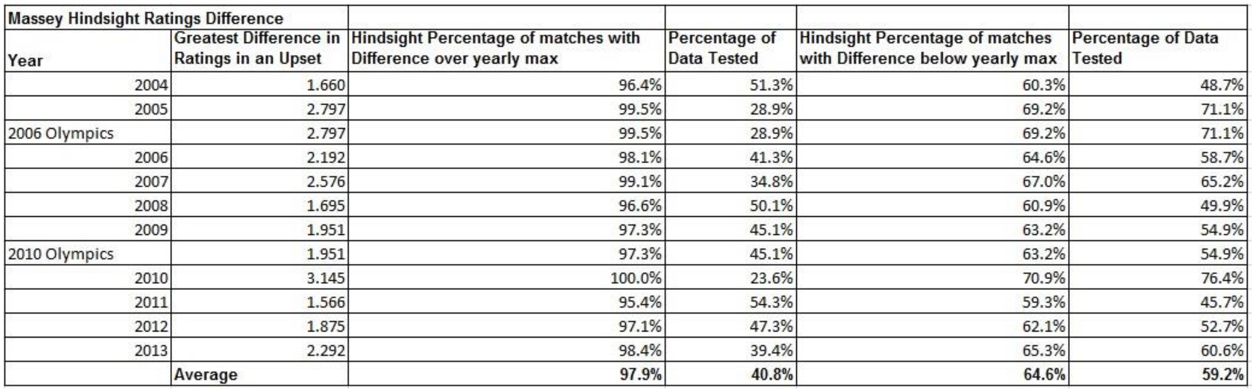
18.256, which 62.4% of the data had a greater difference of the ratings. If one takes the hindsight accuracy of just this data, one obtains a value of 95.3%. Therefore, if the difference of the ratings is above 18.256, the higher ranked team will have a 95.3% chance of winning.



Straight Elo Hindsight Ratings Difference					
Year	Greatest Difference in Ratings in an Upset	Hindsight Percentage of matches with Difference over yearly max	Percentage of Data Tested	Hindsight Percentage of matches with Difference below yearly max	Percentage of Data Tested
2004	26.291	96.3%	54.3%	66.3%	45.7%
2005	29.874	97.8%	48.7%	67.8%	51.3%
2006 Olympics	29.874	97.8%	48.7%	67.8%	51.3%
2006	30.223	98.1%	47.9%	68.0%	52.1%
2007	27.804	97.1%	52.4%	66.6%	47.6%
2008	22.213	96.1%	59.8%	63.2%	40.2%
2009	30.248	98.4%	47.8%	67.8%	52.2%
2010 Olympics	30.248	98.4%	47.8%	67.8%	52.2%
2010	40.575	98.6%	32.0%	74.0%	68.0%
2011	42.461	99.0%	31.1%	74.1%	68.9%
2012	18.256	95.3%	62.4%	62.4%	37.6%
2013	52.363	100.0%	21.5%	76.5%	78.5%
	Average	**97.7%**	**46.2%**	**68.5%**	**53.8%**

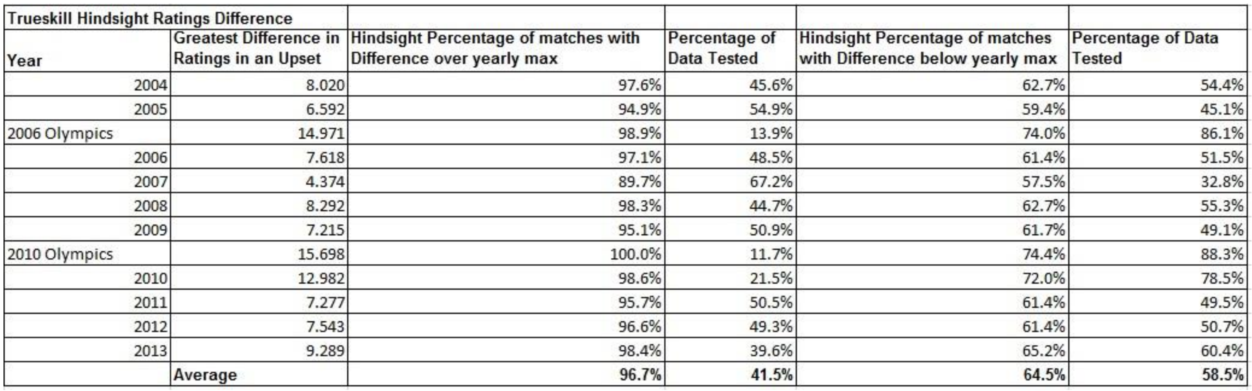
Fig. 2.7. *Hindsight with Straight ELO*

Each ranking method has its own range of values for ratings to figuring out the rankings. Consequently, the smallest difference in ratings for Elo, for instance, is much higher than that of Massey and Trueskill. Therefore, it is impractical to compare the smallest differences from all three methods. Figures 2.8 and 2.9 below show the results for Massey and Trueskill, respectively. Thus, the most one could say is how well do the *rankings* match up to the actual results.



Massey Hindsight Ratings Difference					
Year	Greatest Difference in Ratings in an Upset	Hindsight Percentage of matches with Difference over yearly max	Percentage of Data Tested	Hindsight Percentage of matches with Difference below yearly max	Percentage of Data Tested
2004	1.660	96.4%	51.3%	60.3%	48.7%
2005	2.797	99.5%	28.9%	69.2%	71.1%
2006 Olympics	2.797	99.5%	28.9%	69.2%	71.1%
2006	2.192	98.1%	41.3%	64.6%	58.7%
2007	2.576	99.1%	34.8%	67.0%	65.2%
2008	1.695	96.6%	50.1%	60.9%	49.9%
2009	1.951	97.3%	45.1%	63.2%	54.9%
2010 Olympics	1.951	97.3%	45.1%	63.2%	54.9%
2010	3.145	100.0%	23.6%	70.9%	76.4%
2011	1.566	95.4%	54.3%	59.3%	45.7%
2012	1.875	97.1%	47.3%	62.1%	52.7%
2013	2.292	98.4%	39.4%	65.3%	60.6%
	Average	**97.9%**	**40.8%**	**64.6%**	**59.2%**

Fig. 2.8. *Hindsight with Massey*



Trueskill Hindsight Ratings Difference					
Year	Greatest Difference in Ratings in an Upset	Hindsight Percentage of matches with Difference over yearly max	Percentage of Data Tested	Hindsight Percentage of matches with Difference below yearly max	Percentage of Data Tested
2004	8.020	97.6%	45.6%	62.7%	54.4%
2005	6.592	94.9%	54.9%	59.4%	45.1%
2006 Olympics	14.971	98.9%	13.9%	74.0%	86.1%
2006	7.618	97.1%	48.5%	61.4%	51.5%
2007	4.374	89.7%	67.2%	57.5%	32.8%
2008	8.292	98.3%	44.7%	62.7%	55.3%
2009	7.215	95.1%	50.9%	61.7%	49.1%
2010 Olympics	15.698	100.0%	11.7%	74.4%	88.3%
2010	12.982	98.6%	21.5%	72.0%	78.5%
2011	7.277	95.7%	50.5%	61.4%	49.5%
2012	7.543	96.6%	49.3%	61.4%	50.7%
2013	9.289	98.4%	39.6%	65.2%	60.4%
	Average	**96.7%**	**41.5%**	**64.5%**	**58.5%**

Fig. 2.9. *Hindsight with Trueskill*

3. RANKING MULT-COMPETITOR OLYMPIC SPORTS

In this section, we transition from methods that rank head-to-head competitions to methods that rank multi-competitor competitions. The data provided for the multi-competitor portion of this project came in the form of placements. Consider a three-competitor race with placements, A, first, B, second, and C, third. One can transform these placements into a series of head-to-head events by forming all pairwise combinations of athletes, resulting in the following head-to-head data: A beats B, A beats C, and B beats C. Thus, a multi-competitor event with n competitors creates $\binom{n}{2}$ head-to-head matches. Since methods for ranking head-to-head data exist, this is a very useful transformation.

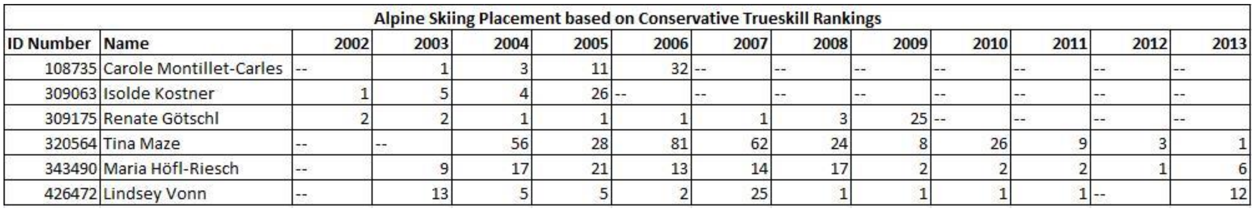
For this multi-competitor to head-to-head transformation, the point differential data of the Massey method is helpful. One can think of the difference in the rank placement as a scoring differential and can create the matrices the same way as presented in the head-to-head section. This implementation provides easily interpretable results, although the computation time is longer than running other methods.

Unfortunately, applying a similar transformation to the Elo and the (head-to-head variant of the) Trueskill methods is problematic due to the updates required after every match. Given that a multi-competitor match is broken up into $\binom{n}{2}$ individual matches, how does one choose the order in which to update the matches? Our experiments showed that the order does indeed affect the results.

Consequently, we focused on the multi-competitor variant of the Trueskill method to handle multi-competitor sports. In particular, we decided to use multi-competitor data from Women's Alpine skiing. Due to the large number of athletes within our dataset, we limited our visual representation (see Figure 3.1 below) to only include those whom have been ranked as number one according to Trueskill for at least one year. The data ranged from 2002 to 2013.

We chose to display the results in a table since implementing the same type of line graphs done for head-to-head data with multi-competitor data can be deemed impractical. This is because it is difficult to compare athletes from different time periods. For example, a notable trend from Figure 3.1 is that athletes will no longer compete after a few years of reaching their peak performance (highest ranking). Contrast this with ice hockey where teams can replace individual athletes, and this results in a much more practical use of line graphs for a wider range of years. National teams exist as long as the country qualifies for the competitions.

Although one could still compare the skiers in the years they overlap, the table is misleading in its representation of the data as a whole. Tina Maze, Maria Hofl-Riesch, and Lindsey Vonn participated much more often than the other skiers included in the data. The average time a skier competed was four years. The three skiers mentioned above have each been participating for at least ten years, and the longevity of these women's careers is uncommon. This makes for fairly sparse data and creates a much higher level of uncertainty due to a low retention rate of athletes from year to year.



Alpine Skiing Placement based on Conservative Trueskill Rankings

ID Number	Name	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
108735	Carole Montillet-Carles	--	1	3	11	32	--	--	--	--	--	--	--
309063	Isolde Kostner	1	5	4	26	--	--	--	--	--	--	--	--
309175	Renate Götschl	2	2	1	1	1	1	3	25	--	--	--	--
320564	Tina Maze	--	--	56	28	81	62	24	8	26	9	3	1
343490	Maria Höfl-Riesch	--	9	17	21	13	14	17	2	2	2	1	6
426472	Lindsey Vonn	--	13	5	5	2	25	1	1	1	1	--	12

Fig. 3.1. *Table of Trueskill Rankings for Women's Alpine Skiing*

To further complicate matters, actual ratings show very little variation. For example, Figure 3.2 shows the ratings for 2005 for the previously mentioned six skiers. Using the interpretation of TrueSkill ratings as parameters to a normal distribution of the athlete's performance, Figure 3.2 displays 95% confidence intervals of the six skiers' performance distribution. One can see that Renate

Gotschl's and Lindsey Vonn's intervals are much higher than the other four skiers' intervals. This suggests that a predictive model based off these ratings would eliminate any possible outcomes where any of the four skiers finish before either Gotschl or Vonn. However, the nature of the sport dictates that this is not something that could be predicicted with absolute certainty, which implies that we may have designed a model too restrictive to yield accurate predictions.

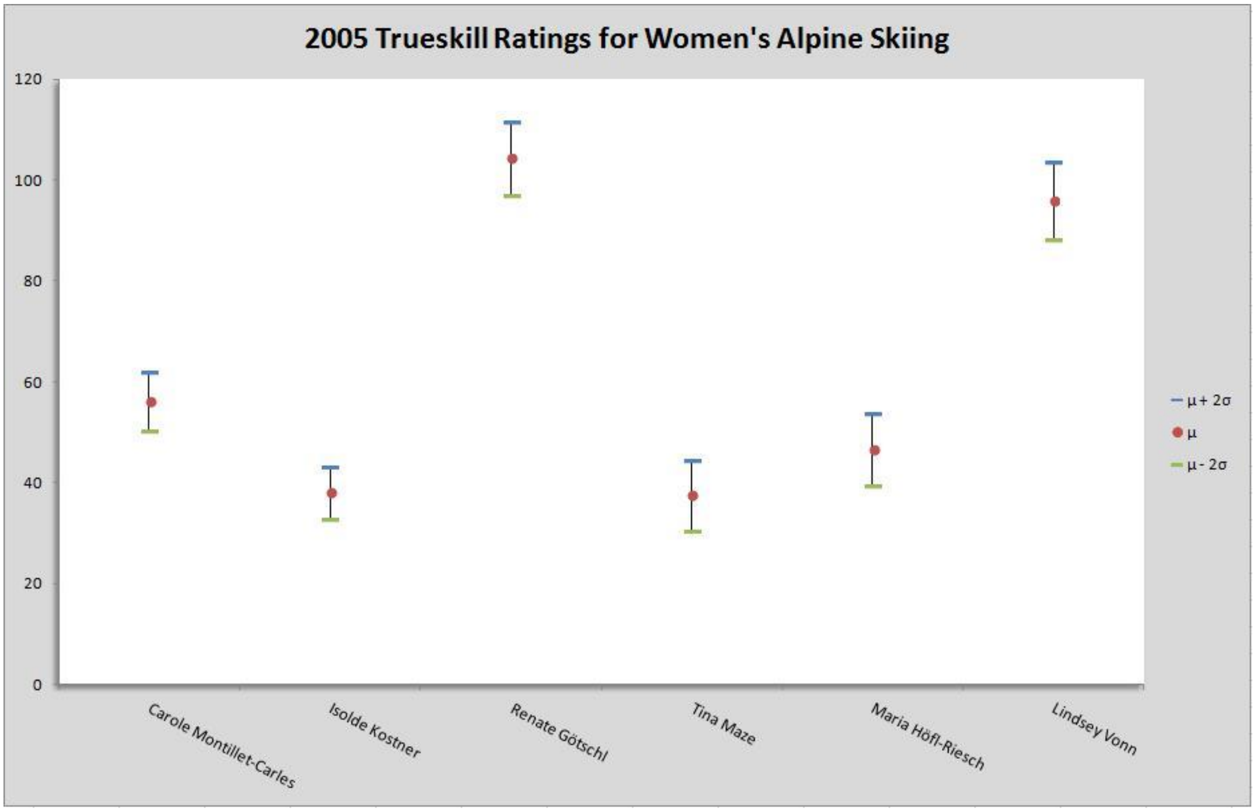


Fig. 3.2. *Graph of 2005 Trueskill Rankings for Women's Alpine Skiing*

These results led us to conclude that multi-competitor sports cannot be analyzed in the same ways we approached head-to-head sports. Further research will need to be conducted in order to find an efficient method of ranking for multi-competitor sports dealing with data only including placements.

4. MONTE CARLO SIMULATIONS TO PREDICT MEDALING

Monte Carlo is a general term used to describe any situation in which random sampling obtains a numerical result. A statistical random variable is used in combination with a win probability given by the ranking methods in order to determine the winner of a hypothetical contest between two teams. This is iteratively repeated in the form of the Olympic tournament in question, and this procedure is done until the results converged to a statistical probability giving what place each team would finish.

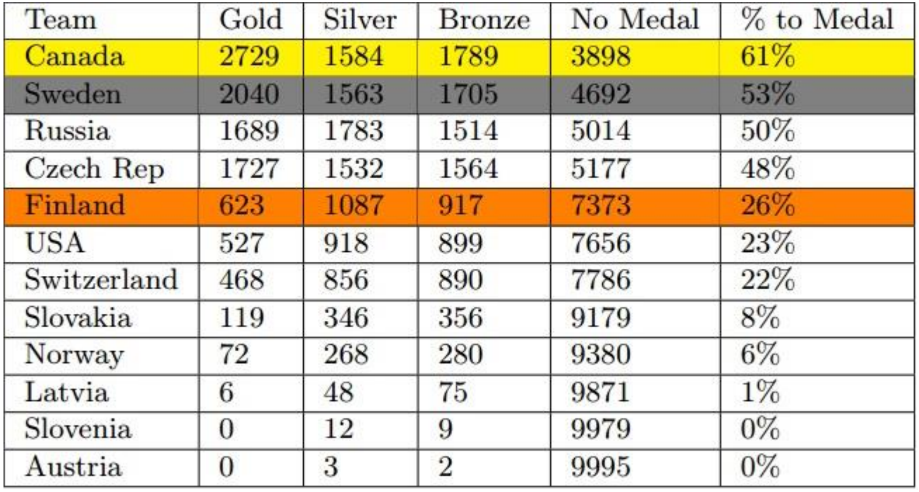
For the Massey method, the probability that team *i* beats team *j* is $\Phi(\frac{r_i - r_j}{\sigma})$, where Φ is the standard normal cumulative distribution function and σ is the standard deviation of all the Massey ratings. The normal cumulative distribution function gives a value between 0 and 1, which allows one to compare the win probability to a $X \sim Unif(0{,}1)$ random variable. Thus, this comparison results in the ability to pick the winner of a hypothetical match between teams *i* and *j*.

The Elo method produces the value $\mu_{\mathbf{i,j}}$, which is the probability that team *i* beats team *j*. Thus, one can create a random variable with a $\mu_{i,j}$ probability of being in one region and a $1 - \mu_{i,j} = \mu_{j,i}$ probability of being in another region. To do so, take the value $X \sim Unif(0,1)$ and compare it to the $\mu_{i,j}$ value. If $X < \mu_{i,j}$ then the hypothetical contest results in a win for team *i*. Otherwise, team *j* wins the hypothetical contest.

Since Trueskill treats skill as two variables, mean μ and standard deviation σ, one can use these two variables to create a distribution for the skill of each player. Consequently, the skill for player *i* can be described by $S_i \sim N(\mu_i,\sigma_i^2)$ while the skill for player *j* can be described by $S_j \sim N(\mu_j,\sigma_j^2)$. Drawing random values from each of these distributions results in the ability to directly compare the hypothetical skill levels between players *i* and *j* in a given match.

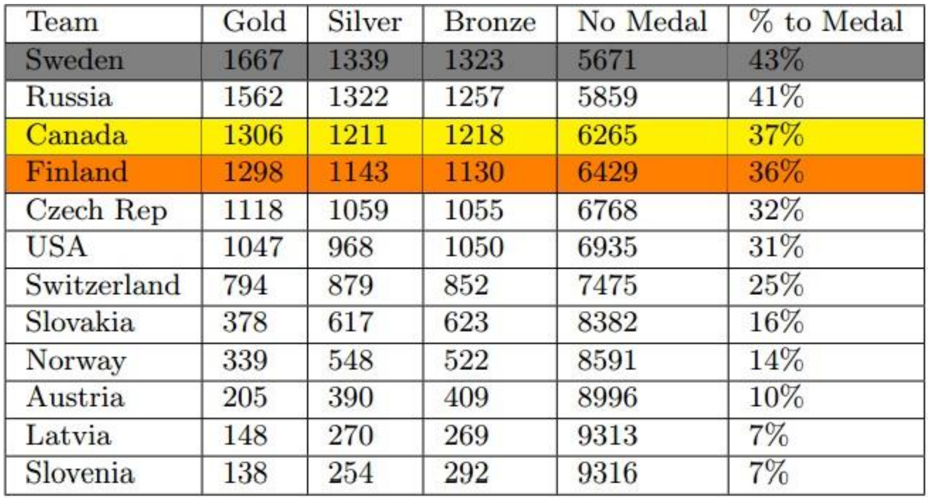
With these methods for finding the winner of a hypothetical match between any two competitors, the rest of the predictive Monte Carlo model is derived from the Olympic sport in question. For men's ice hockey, we followed the format of the tournament for that sport and thus began by simulating a group stage involving three groups of four teams by using the teams that qualified for the Olympics. This provided the order in which the teams are placed into a knockout-style competition. The tournament was simulated until a winner was confirmed. For our simulation, this process is repeated multiple times, in this case ten-thousand, until we reached the results in the following tables. The colored rows indicate which countries actually medaled in the 2014 Olympics.

Table 4.1
2014 Prediction Results - Massey



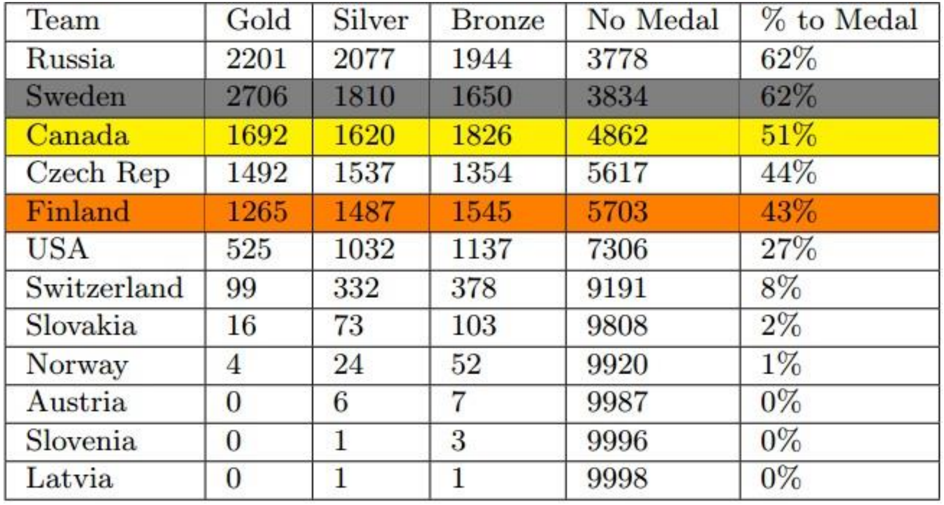
Team	Gold	Silver	Bronze	No Medal	% to Medal
Canada	2729	1584	1789	3898	61%
Sweden	2040	1563	1705	4692	53%
Russia	1689	1783	1514	5014	50%
Czech Rep	1727	1532	1564	5177	48%
Finland	623	1087	917	7373	26%
USA	527	918	899	7656	23%
Switzerland	468	856	890	7786	22%
Slovakia	119	346	356	9179	8%
Norway	72	268	280	9380	6%
Latvia	6	48	75	9871	1%
Slovenia	0	12	9	9979	0%
Austria	0	3	2	9995	0%

Table 4.2
2014 Prediction Results - Elo



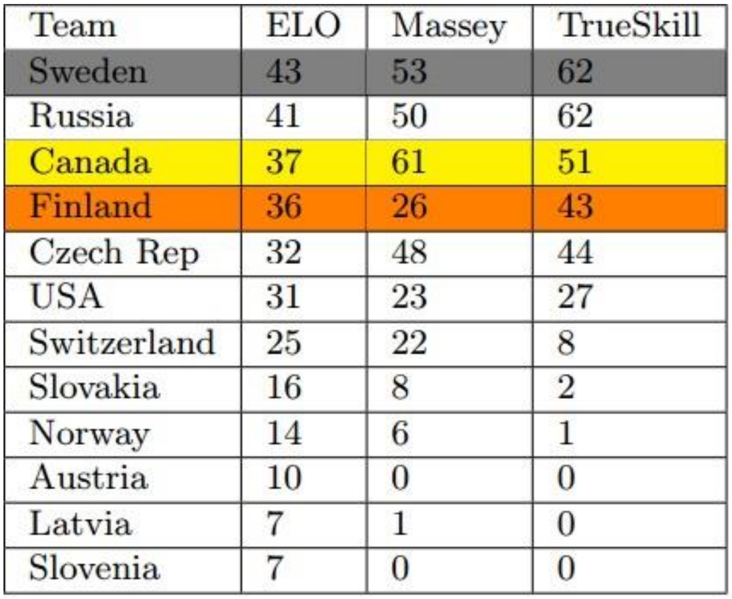
Team	Gold	Silver	Bronze	No Medal	% to Medal
Sweden	1667	1339	1323	5671	43%
Russia	1562	1322	1257	5859	41%
Canada	1306	1211	1218	6265	37%
Finland	1298	1143	1130	6429	36%
Czech Rep	1118	1059	1055	6768	32%
USA	1047	968	1050	6935	31%
Switzerland	794	879	852	7475	25%
Slovakia	378	617	623	8382	16%
Norway	339	548	522	8591	14%
Austria	205	390	409	8996	10%
Latvia	148	270	269	9313	7%
Slovenia	138	254	292	9316	7%

Table 4.3
2014 Prediction Results - Trueskill (β = 40)



Team	Gold	Silver	Bronze	No Medal	% to Medal
Russia	2201	2077	1944	3778	62%
Sweden	2706	1810	1650	3834	62%
Canada	1692	1620	1826	4862	51%
Czech Rep	1492	1537	1354	5617	44%
Finland	1265	1487	1545	5703	43%
USA	525	1032	1137	7306	27%
Switzerland	99	332	378	9191	8%
Slovakia	16	73	103	9808	2%
Norway	4	24	52	9920	1%
Austria	0	6	7	9987	0%
Slovenia	0	1	3	9996	0%
Latvia	0	1	1	9998	0%

Table 4.4
2014 Prediction Results - All % Chance to Medal



Team	ELO	Massey	TrueSkill
Sweden	43	53	62
Russia	41	50	62
Canada	37	61	51
Finland	36	26	43
Czech Rep	32	48	44
USA	31	23	27
Switzerland	25	22	8
Slovakia	16	8	2
Norway	14	6	1
Austria	10	0	0
Latvia	7	1	0
Slovenia	7	0	0

5.CONCLUSIONS AND FUTURE WORK

Our analysis of Olympic sports used ranking methods and Monte Carlo simulation to assess America's competitiveness. Such analysis allows U.S.O.C officials to make statements such as the following regarding men's ice hockey.

1. The United States was the sixth best team in the world based on recent results at the beginning of 2014 Olympics;
2. The United States Men's Ice Hockey Team had roughly a 27% chance to medal in the 2014 Olympics;
3. While the United States did not medal, they did finish 4th, exceeding any metric of expectation for the Olympic Games;
4. Since the 2010 World Championships, the United States has been improving at ice hockey.

However, these statements only deal with one sport. It is possible that for other head-to-head sports we cannot arrive at similar conclusions. For instance, ice hockey is a team head-to-head sport, and teams can replace athletes. Contrast this with individual head-to-head sports, such as fencing or taekwondo, where it might not be as intuitive to use the same type of tests done in section 2.4. It is possible that for individual head-to-head sports that predicting the top four athletes in the current year based on the top six or top eight from the previous year would result in much lower percentages. Also, for team head-to-head sports, there could be a more or less balance of play than that of ice hockey. Thus, more research needs to be conducted on other types of head-to-head sports to see if the three ranking methodologies will arrive at similar results. We have not yet run Monte Carlo simulations from the rankings generated for multi-competitor sports but intend to undertake this as future work along with implementing the ranking methods for other head-to-head sports.

REFERENCES

[1] Langville, A. N., Meyer, C.D. (2012) *Who's #1?: The Science of Rating and Ranking*, Princeton University Press, Princeton, NJ.

[2] Massey, Kenneth. Statistical Models Applied to the Rating of Sports Teams. Thesis. Bluefield College, 1997. N.p.: n.p., n.d. Massey Ratings.com. Sponsoring Professor: Mr. Alden Starnes. Web. ¡http://www.masseyratings.com/theory/massey97.pdf¿.

[3] ”TrueSkill Ranking System.” Microsoft Research. Microsoft, n.d. Web. 25 Feb. 2014. ¡http://research.microsoft.com/en-us/projects/trueskill/¿.