Generating and Analyzing Rankings of Cross Country Skiers

Karl Holub

Contents

Introduction

World Cup Cross Country Skiing

World cup cross country ski racing is an endurance athletic where skiers traverse uphills, downhills, and flats as rapidly as possible. Depending on the type of event, skiers may be skiing together or alone (mass start or individual start). Regardless, each world cup ski race features around 50 of the world's best skiers competing against one another. Some skiers specialize in short races, others long distances. Some skiers perform better in mass start races compared to individual start. It is somewhat uncommon for a skier to dominate all facets of racing; even if they are, all skiers experience peaks and valleys in performance throughout a season. All these factors make generating a generic ranking system difficult, and to date, no overall ranking system is in place for the over 100 active skiers in the FIS World Cup.

Applications of a Ranking System

Luckily, the FIS maintains a full record of all World Cup ski results over the last 15 years. So, building such a ranking system is within reach. A ranking system would be useful to race directors in being able to order skiers' seeding at races, advertisers and sponsors to determine who the most consistent and talented racers are, and ski coachs and analysts who may wish to understand different skill levels across countries or other skier characteristics.

Furthermore, with a rich history of results available, a timeline of rankings can be used to generate a skier's career trajectory. It is widely held that skiers "peak" in terms of athletic performance for a given number of years before their performance begins to decline. A ranking system might be used to generate a database of ranking timelines that would be useful in classification and prediction of current skiers' career trajectories.

Harkness Ranking System

The Harkness Rating System (Harkness, 1956), introduced in the 1950's for use in ranking chess players, is one of the earliest and simplest ranking systems. A player enters a tournament with a score of R_{pre}. In the tournament, the player plays a number of matches with k wins and m losses. Let $\bar{R}$ be the average rating of all players in the tournament. Then the player's score after the tournament is:

$$R_{new} = \bar{R} + 1000 * (\frac{k}{k+m} - .5)$$

In plain language, the updated score is the average of the tournament, with a bonus/loss of 10 points for every percentage point above/below average expected performance. In the case that a player enters the tournament without a score, the average may not include that player, or a value may be somehow imputed.

One weakness of the Harkness Ranking system is that it reacts to potential skill changes very

rapidly. If a player has an off day and losses a greater than expected number of matches for R_{pre}, the score immediately dips to reflect the losses. Another weakness is that a player may play a non-representative sampling of other players in the tournament (the average of $|R_{pre} - \bar{R}|$ is large). Although in the long term this effect is washed out and negligible, this does mean that short term windows in a player's ranking may be unreliable estimates of skill.

Harkness Computational Considerations

The above score update must be individually applied to each tournament participant. However, this system is computationally efficient, requiring $O(n)$ arithmetic operations where n is the number of participants. So, applying to a timeline of p fixed size tournaments, the entire algorithm costs $O(pn)$ operations to compute Harkness timelines.

Elo Ranking System

The Elo Ranking System (Elo, 1978) introduced in the 1970's was widely considered an improvement to the Harkness and other early ranking systems. A player enters a tournament with score R_{pre}. In the tournament, the player wins and losses some number of matches, represented as sets of players. The resulting Elo score is computed as:

$$R_{new} = R_{pre} + K * (\sum_{p \in wins} (1 - \frac{10^{R_{pre}/400}}{10^{R_p/400} + 10^{R_{pre}/400}}) + \sum_{p \in losses} (0 - \frac{10^{R_{pre}/400}}{10^{R_p/400} + 10^{R_{pre}/400}}))$$

where K is some positive value. The inner terms of the summations can be thought of as the difference between the expected outcomes (ratings on a log-transformed scale) and the actual outcome (1 for a win, 0 for a loss) (Glickman & Jones, 3). K is interpretable as a parameter controlling the rate at which the score changes. If a player is underrated with R_{pre}, a large value of K will rapidly bring the rating close to the true value (Glickman & Jones, 4) . The other constant appearing in the update is 400. This value is of little consequence to the

meaning of the equation, but rather sets the scale on which the ratings exist. It designates that a hypothetical rating increase of 400 makes an order of magnitude difference in the expected outcome of a match:

$$10^{(R+400)/400} = 10 * 10^{R/400}$$

In this sense, any set value will suffice as it only influences the scale on which the rating is interpretted. 400 is used as convention.

If a player enters a tournament with only a few matches or less, the player's rating is considered unreliable. Several imputations exist (Glickman & Jones, 5), the simplest of which is to set some starting default value. Typically this value is chosen to be less than the average Elo score, as a new player presumably is not very skilled.

Elo Computational Considerations

The above update equation must be applied to each player in the tournament. While memoization or dynamic programming may be used to avoid some repeated computations, asymptotically the update requires $O(n^2)$ arithmetic computations for a tournament, where n is the number of participants in the tournament. So, across p tournaments of fixed size, on the order of $O(n^2p)$ arithmetic operations will be required to compute Elo timelines.

Proposed Approach and Analyses

Modified versions of both the Harkness and Elo ranking systems will be used to generate score timelines for World Cup Skiers. The systems will be compared on an applied level. The systems will also be subject to a heuristic validation using news articles reporting on World Cup skiing.

With a series of rankings of active skiers, the most recent rankings will be used to evaluate

who the best skier as of 2016.03.12 is, if there is any effect on country of origin on score, and to test if there are any significant differences in score between specific countries. Specifically, any difference in Elo between Canada and France will be tested, since these two nations often train with one another.

Methods and Implementations

Target Data

To limit the volume of data handled, only Male skiers active in the period of 2000.01.01 to 2016.03.12 with 10 or more recorded races will be used in computing scores. The 10 race minimum is imposed to prevent racers with insufficient data from entering the scoring pool; as explained in later sections, racers enter the scoring pool with an average score. Given that it takes multiple races for the ranking algorithms to converge on a racer's real score, a minimum of 10 races, or roughly half a season, is required.

Any non-individual race (e.g. relays) was ignored, since this data is not within the scope of individual rankings. Additionally, no results of qualifying races were used, since these do not typically reflect the full effort of the skiers.

Web Scraping Race Results

All FIS World Cup Cross Country Ski Results can be found on the FIS website (see Appendix 1). The FIS offers no API to download this data en masse, so a scraper was implemented using python to obtain the data (see Appendix 2).

Over 100,000 race results are stored on the FIS website- many of these races fall outside the target specifications and were ignored. However, still several thousand races needed to be processed and stored efficiently. Since querying a webserver is a slow, I/O bound task, a

multithreaded program was used to read results. Each set of race results is presented on the FIS website inside an HTML table, so the webpage content was scraped by running it through an XML parser. Once race data was read using the parser, it was stored on the local filesystem as a tab separated value file. By storing results, expensive repeated queries to the FIS servers are avoided.

Although most data were consistently formatted, some exceptions did exist in the FIS database. Since the data set was so large, individual verification and correction of these exceptional data entries was not possible; Any unexpected data was simply ignored. In practice, this approach caused the loss of less than 10 races.

Implementation of the Modified Elo Ranking Algorithm

As noted in the introduction, the Elo system is intended for handling 1-1 matches in a tournament. To extend to a setting with multiple players, the total results are broken down into individual results that are as if the race was 1-1. So, a first place finisher has expanded results that are all victories. The second place finisher has expanded results of 1 loss to the winner and wins against the remainder of players. Using these expanded results, the Elo as explained in the introduction was used with k=2 (typical values are 8,16,32). A smaller value was used since the shear number of expanded results greatly bolsters the effect of performing well, even at a small K.

The actual Elo implementation in python is straightforward (see Appendix 3); Elo scores are stored in a matrix. Support structures like hash tables are used to quickly find the row or column entry given a date and an FIS id or date. All the results are sorted in ascending order of occurrence, all pairs of (winner,loser) are generated for a given race, and the update function is applied for each competitor in the race.

Implementation of the Harkness Ranking Algorithm

A modified version of the Harkness Ranking Algorithm was not required, as long as individual races are treated as total tournaments.

Implementation of the Harkness (see Appendix 4) was also straightforward with a 2 dimensional matrix for scores and hash tables for quick row and column indices. All race results are sorted in ascending order of occurrence, the mean of the competitors for a race is computed, and then the update function is computed for each competitor in the race.

Time Series Smoothing

Once time series data is obtained, it will become necessary to visually inspect it with plots. The rankings will wildly vary within small windows, making visual interpretation difficult. So Cubic Smoothing Splines will be used to generate globally interpretable lines from the data scatter without losing all of the local variation.

ANOVA and t-tests for Rating Differences

Using the most recent (last recorded race was 03/12/2016) ranking, questions about the state of international competition may be answered. The specific question posed in this project is: does the country of origin impact the skill level of the skier? An appropriate analysis for the question is a one-way ANOVA, given some constraints to maintain observation independence (see Results section for more detail).

Another question of interest is to determine if two specific countries produce skiers of different caliber. Two countries of interest are Canada and France; they are notable because they frequently train and compete with one another throughout the off-season. One might reasonably expect these two countries to have comparable ratings given their similar training regimen.

Table 1: Algorithm Runtimes

System	Avg Runtime
Elo	28s
Harkness	1.3s

Results

Overview

Given the target data restrictions, 555 skiers, potentially competing in 470 races over the 15 year window, were obtained. Running the algorithms described in the Methods section, the runtimes shown in Table 1 where achieved. As expected from the asymptotics, the Harkness system is faster to compute- an important consideration if the racer space were to be expanded to the larger world of skiers outside just the World Cup.

To establish a baseline for the Elo rating system, it is necessary to consider how average Elo scores change over the years.

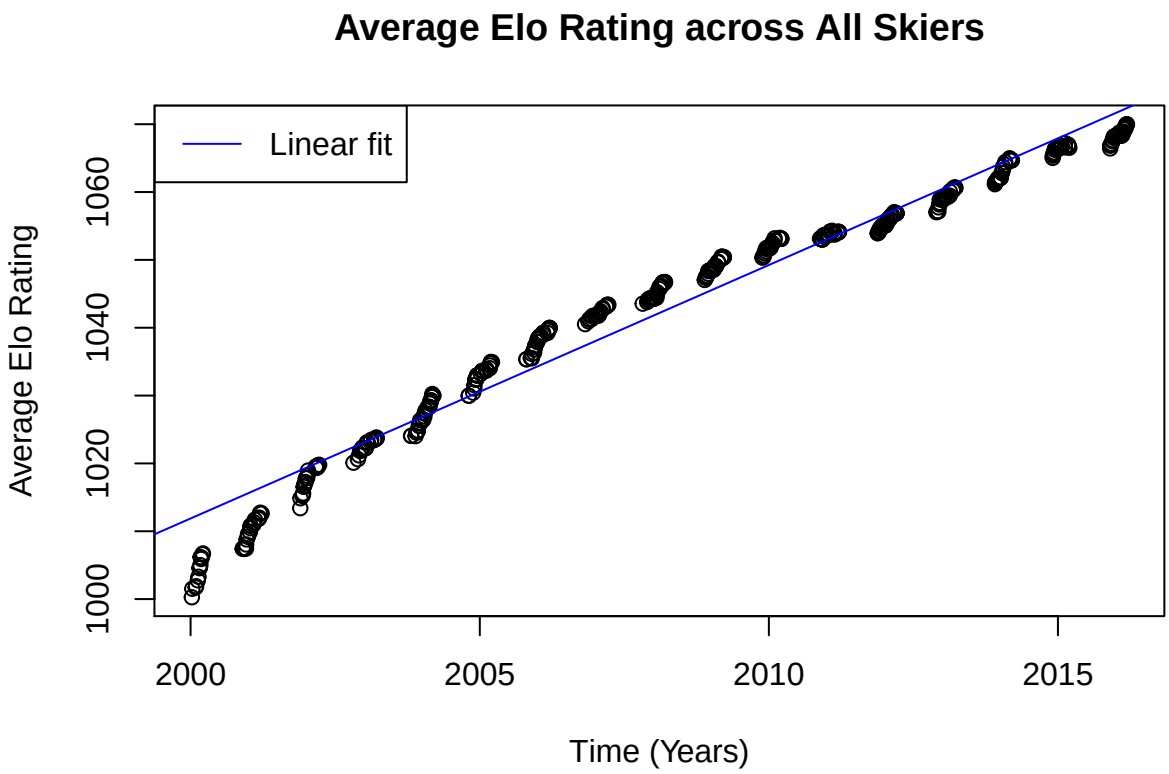

Figure 1

Average scores increase with time, but the trend is beginning to plateau. A likely explanation for this trend is that lower caliber skiers tend to be dropped from the World Cup level of competition very rapidly if they are not as successful as their counterparts.

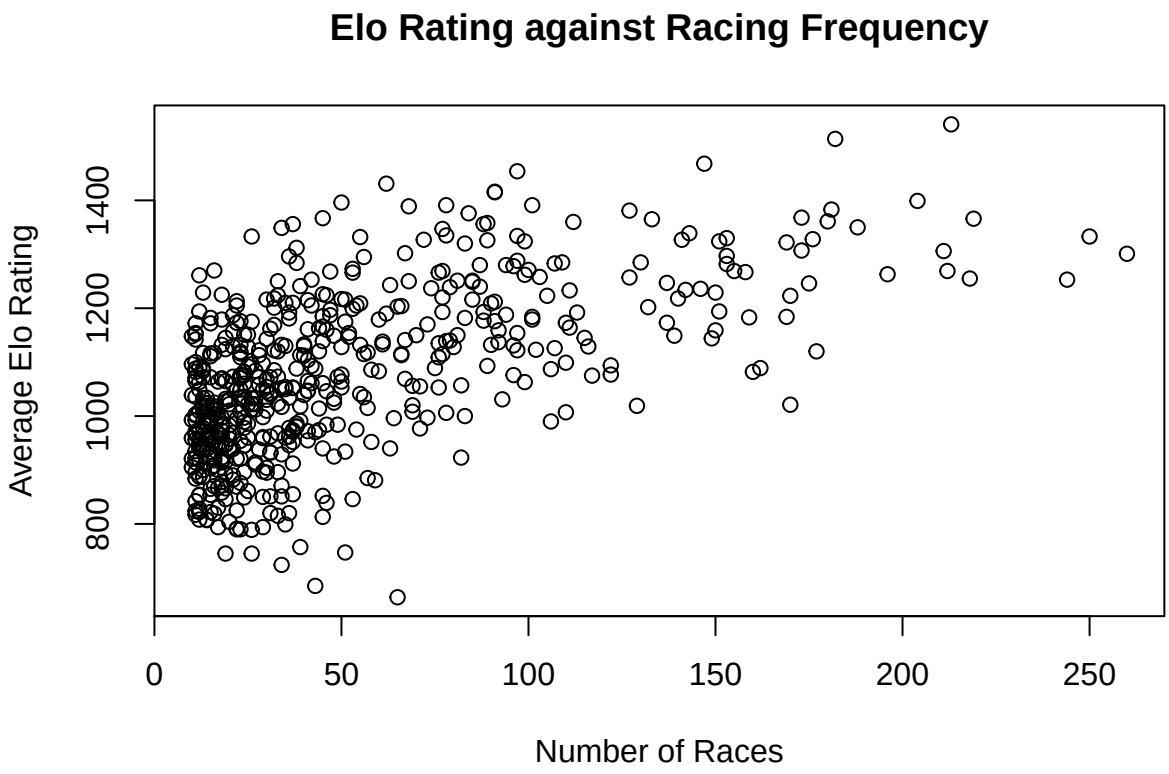

Figure 2

This figure indicates that almost all skiers with an ELO below 1000 drop off the World Cup race circut with fewer than 50 races. So, the average ELO rankings over time are less likely to include the individuals with these lower scores, explaining the increase in average score over time.

Heuristic Validity of Ranking Methods

It is difficult to empirically assess the validity of a ranking system, since the measure of quality is subjective and the true ranking of skiers is unknowable. Instead, we rely on subjective judgement of the system (by finding skier placements in the all results over the years), media sources, or using metrics giving some indication of skier quality. Several skiers have had remarkable career trajectories that are very unique and identifiable, so this analysis proceeds

with those skiers. Below is a plot of their smoothed Elo ratings.

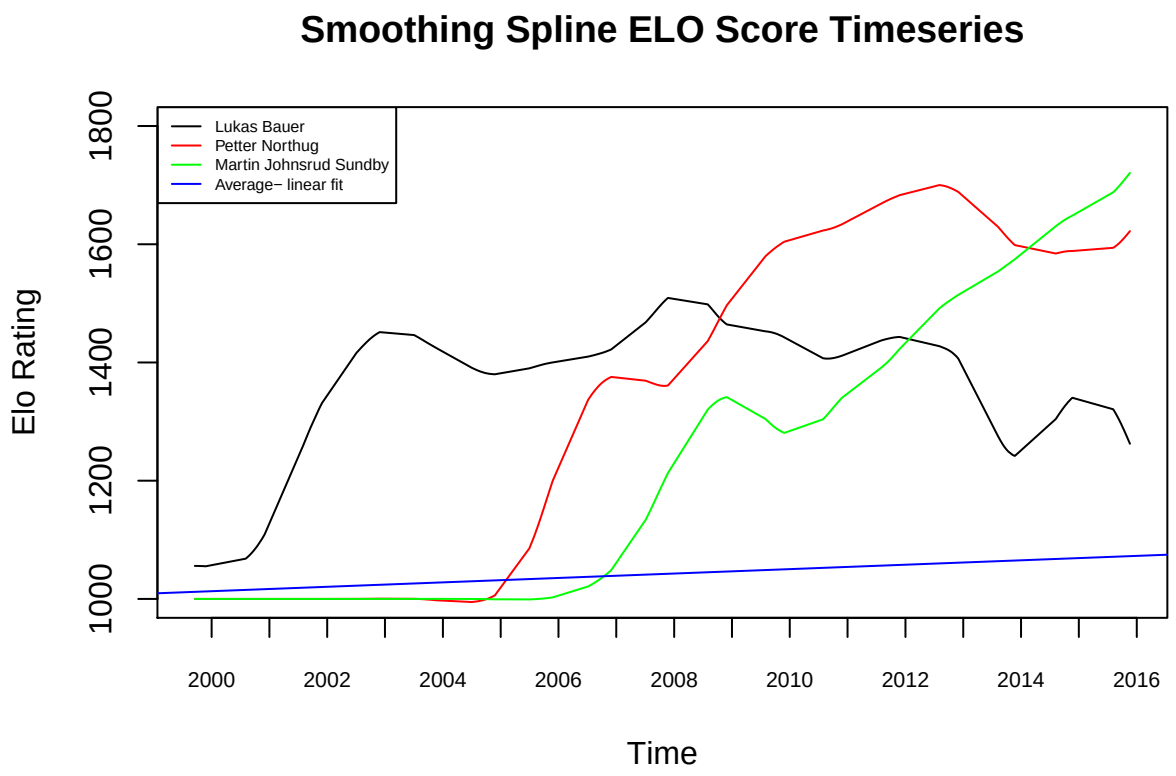

Figure 3

Petter Northug was widely held to be the top skier in the world for the late 00's and early 10's. His results have been strong, but not as dominant since the 2014 Olympic games. Again, the Elo timeline matches well with this narrative. The plot does not show it, but Northug indeed had the highest Elo rating of the field from 2010 to 2013. For a more quantitative approach, we consider the number of podiums (top 3 finishes in races) Bauer achieved in comparison to his Elo rating. The below plot shows that these metrics align well.

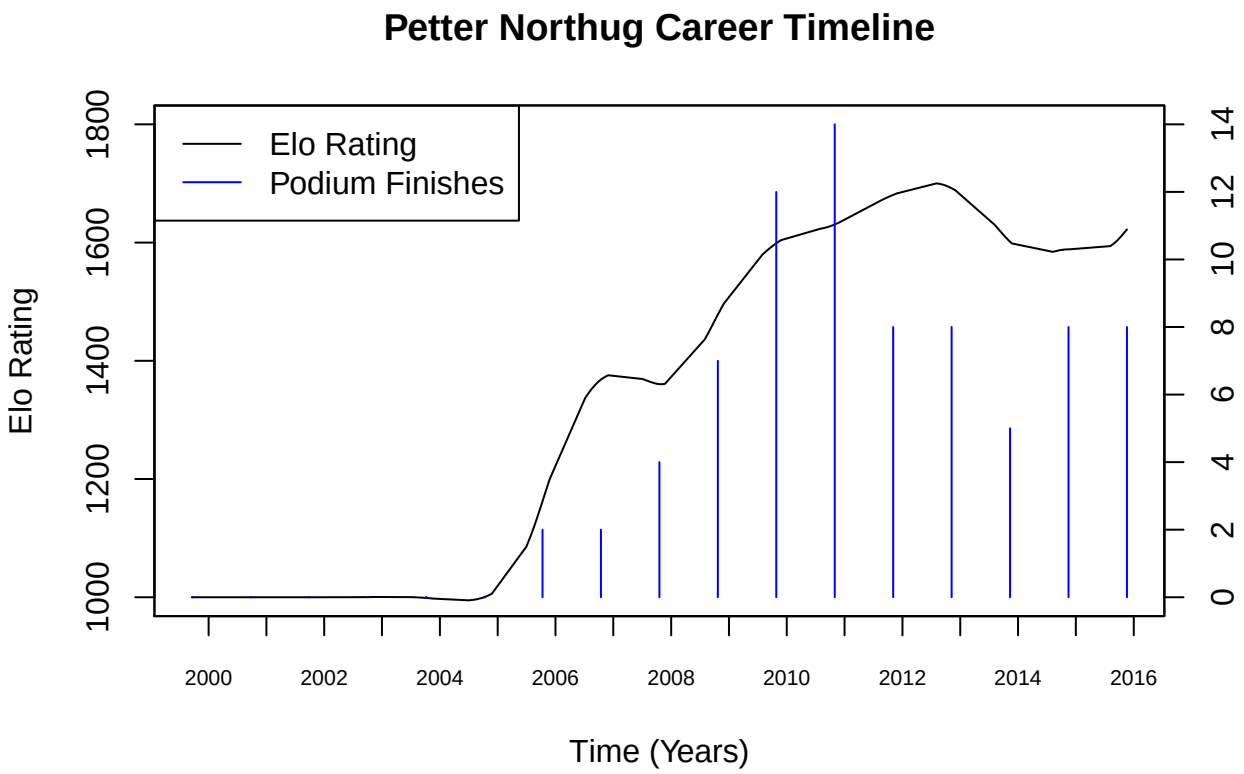

Figure 4

Lukas Bauer (black line) is known for his consistent top results and longevity- a career that has spanned over a decade. In recent years, he has had success, but less consistently. His Elo timeline very much aligns with this assessment- a marked increase in 2003, consistent, near top results through 2013, and then a drop off after. For a more quantitative approach, we consider the number of podiums (top 3 finishes in races) Bauer achieved in comparison to his Elo rating. The below plot shows that these metrics align well.

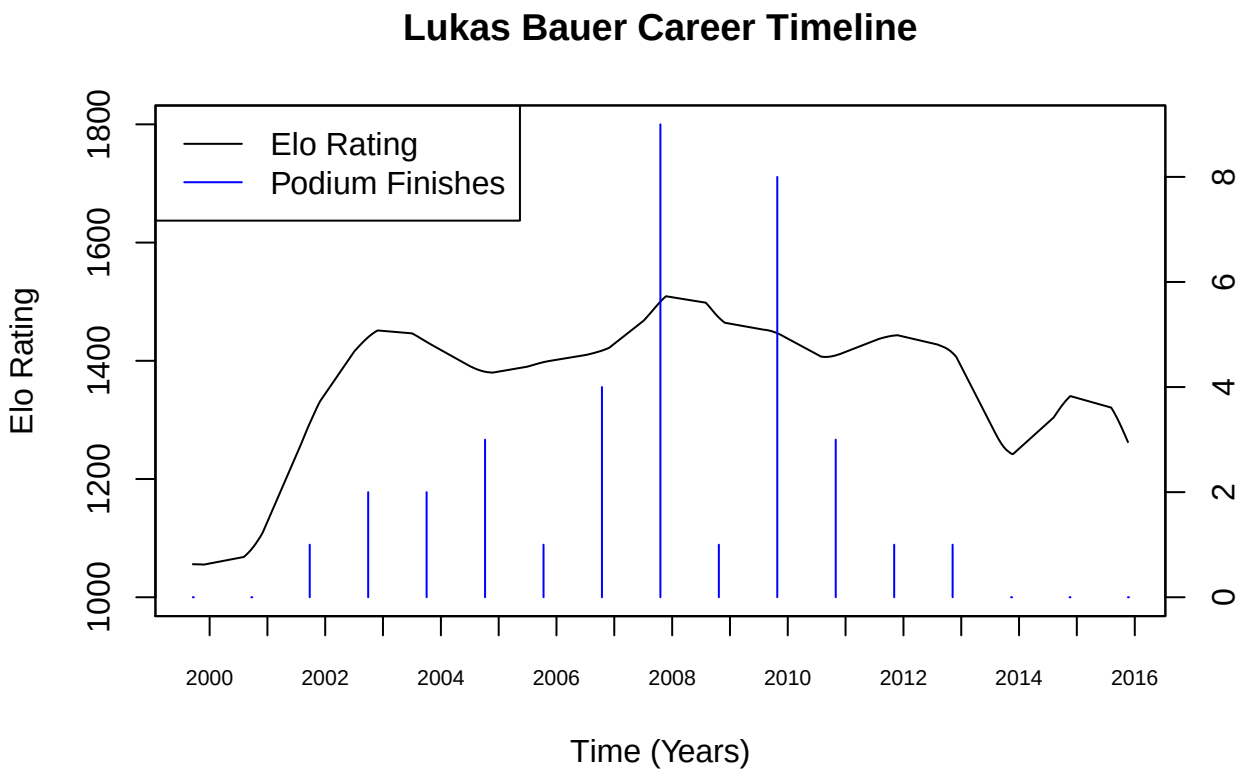

Figure 5

Finally, Martin Johnsrud Sundby is considered the top skier in the world, as of 2016.03.12. He is renowned for his seemingly constant improvement, and after the winning the recent 2016 Tour of Canada, has left analysts questioning if he is the most dominant skier of all time. The Elo timeline again aligns well with Sundby's career. Steady improvement from 2007 onward, surpassing Northug in the 2014 season, and a spike in late 2016 representing his win in the Tour of Canada all align well with the rating.

It would require a large time investment to repeat this analysis for all 500 skiers; these 3 examples give a very compelling case that the Elo ranking system reliably indicates skier rank.

Comparison with Harkness

The Harkness scores for the same 3 skiers:

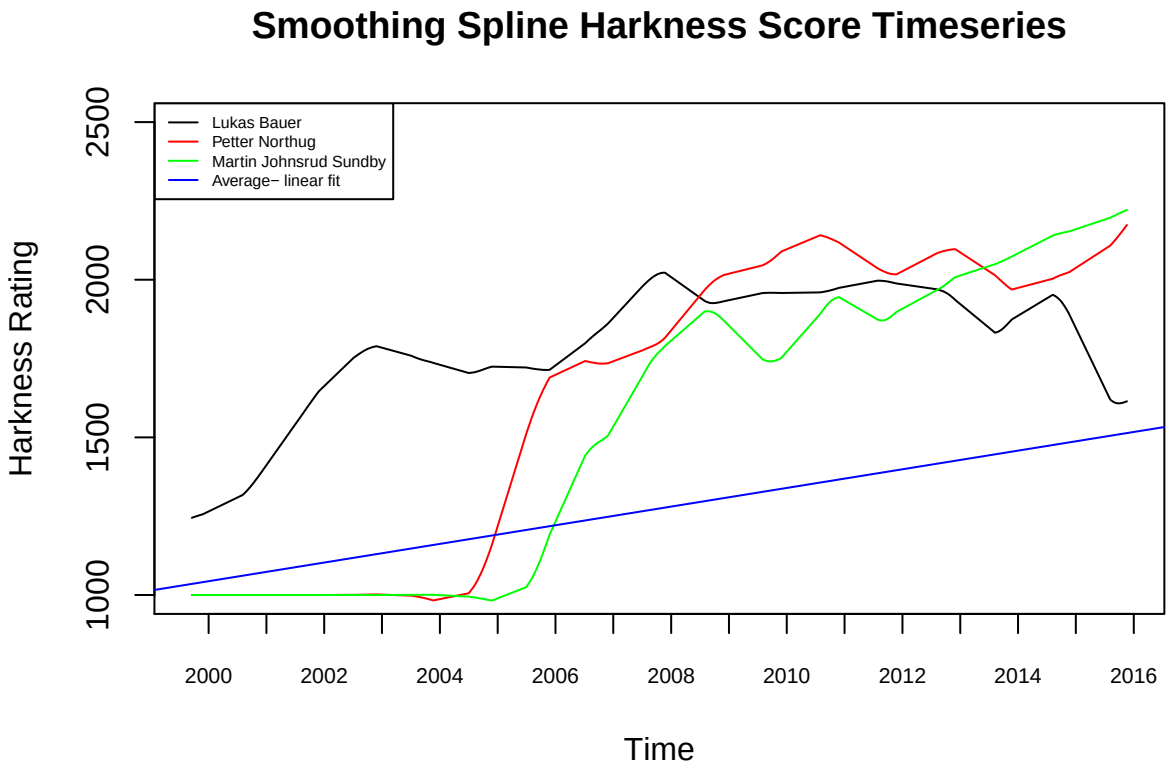

Figure 6

Note that the Harkness ratings are on a different scale. Still, the timelines contain much of the same information, but with one telling, significant difference. Northug (blue line) and Sundby (green line) are almost of the same rank at the end of 2016. The Elo score does not have them so closely matched at that time, and rightfully so; Northug had a difficult 2016 except for an excellent series of races in the Tour of Canada that he was 3rd overall in. Recall that Harkness ranking reacts quickly to perceived changes in skill level. So a small series of placements close to Sundby in the Tour of Canada launched Northug close to Sundby's level, despite being far off Sundby's level for much of the season.

Another interesting point of comparison for the methods is the similarity of the intersection

Table 2: Time of rating intersection

	Harkness	Elo
Northug >Bauer	early 08	late 08
Sundby >Bauer	early 12	early 12
Sundby >Northug	late 13	early 14

points of the rating timelines.

We see that the intersections are all close, with the Harkness intersection points coming a bit sooner. This is again because upswings and downfalls are quickly reflected in the rating.

Analysis of Skier Caliber Across Nations

Note that for the following analysis, all tests performed rely on the assumption that observations are independent. The observations in this case are certainly dependent to some extent, given that after a race one skier's score increase is another's score decrease. However, we are performing an analysis on a small subset of all skiers (around 50) within a larger field of nearly 500. So, while the skiers in this subset have some dependence with one another, a lot of the dependence is washed out in the larger field of 500 that is not being analyzed. In other words, the magnitude of dependence may not be large enough to disrupt the below analysis.

We wish to determine whether country of origin, within the group of France, USA, Canada, and Switzerland, has a significant effect on Elo rating. This grouping of countries was selected since they have all been involved in World Cup racing for comparable amounts of time. Inspecting the data, it certainly appears that there are centrality differences in the distribution of the data across countries.

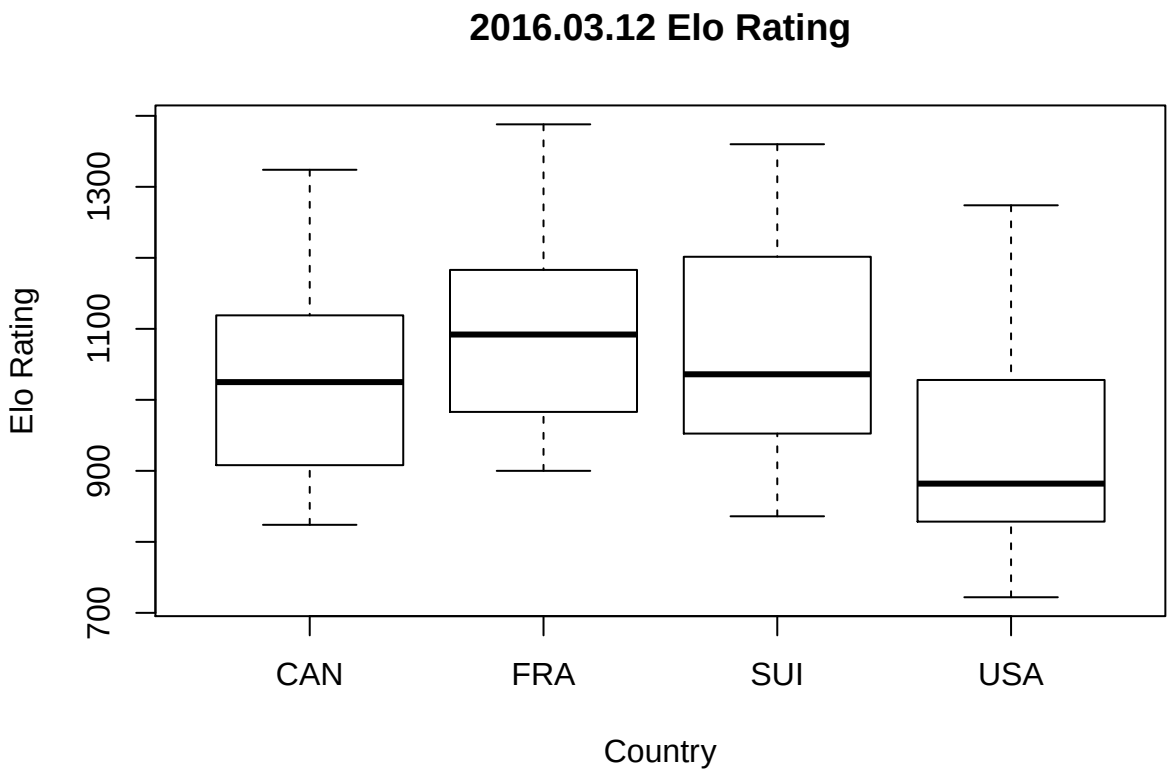

Figure 7

To determine an appropriate test, first normality must be assessed. Histograms reveal that many of the countries have unacceptable departures from normality, typically excessively long tails.

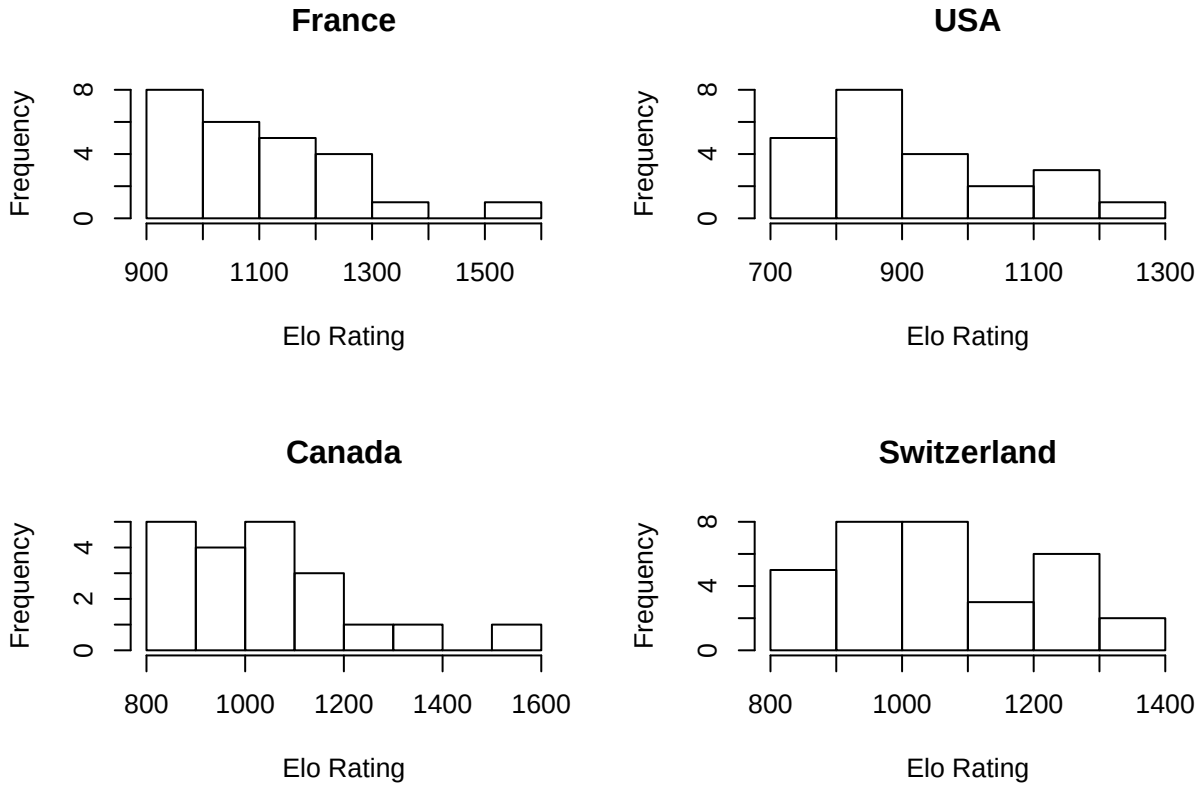

Figure 8

Since there is a departure from normality, an ANOVA would be inappropriate for analyzing mean differences in the data. Instead, the non-parametric Kruskal-Wallis ANOVA will be used to test the following hypothesis:

$$H_0 : The\, data\, derive\, from\, the\, same\, distribution$$

$$H_a : The\, data\, derive\, from\, different\, distributions$$

The test statistic of 266.2 (associated p-value $< 2.2^{-16}$) leads to a rejection of H_0, and we may conclude that certain countries have different distributions, and that some countries produce more successful skiers than others.

Next, we investigate a pair of countries of specific, predetermined interest. France and Canada

regularly train and compete with one another, moreso than other nations. Their mean Elo ratings appear similar at 1043 and 1114. From their histograms, the distributions of France and Canada both appear significantly non-normal, indicating that a non-parametric test must be used. A two-sided Wilcoxon Rank-Sum test will be used to test the hypotheses:

$$H_0 : \ French\, and\, Canadian\, Elo\, ratings\, are\, drawn\, from\, same\, distribution$$

$$H_a : \ French\, and\, Canadian\, Elo\, ratings\, are\, drawn\, from\, different\, distributions$$

The resulting test statistic was 332 with an associated p-value of .063. Given this non-extreme p-value, we fail to reject H_0 that the distributions are different. This finding fits the narrative that the French and Canadian skiers often train together and have comparable skill levels.

Discussion & Conclusions

Two ranking systems, Harkness and Elo, were applied to 15 years of cross country skiing data scraped from an online repository. Both ranking systems proved reasonably successful. It is difficult to define accuracy of a ranking in the setting of cross country skiing, where there is no identifiable ground truth, nor even a universal agreement what the criteria for being successful are. Instead, the ranking systems are subjectively judged on their ability to match general perception of the trajectory of a skier, as well as providing accurate discrimination between skiers of obviously different calibers. Both ranking systems performed well in matching the career trajectories of 3 of the world's top skiers over the period of 2000-2016.

In fact, the FIS offers a points competition throughout the season, where racers can accumulate points to compete for a season long victory. This competition should not be mistaken for an overall ranking mechanism (it fails to consider the quality of the opponents at races and does not generate fair rankings for lower quality skiers). However, the top 10 opponents in this FIS classification were all present in the top 15 of the ranking generated by the Elo ranking.

On the issue of Elo vs. Harkness, the two measures provide very similar information. However, the algorithm for computing Harkness ratings is much more efficient and leads to faster runtimes in practice. In the face of a larger pool of skiers, the Harkness ranking system would be preferable to Elo for this reason. One weakness of the Harkness was identified in the over-rating of Petter Northug at the end of the 2016 season. Based on his strong performance in a few races at the end of the season his ranking was propelled close to that of best skier (Sundby), despite a season of much poorer results. Northug's Elo ranking was more conservative and still reflected a large, although reduced, gap between Northug and Sundby.

The Elo ranking also appeared to be useful tool for verifying existing knowledge- specifically that athletes from certain countries perform better than those from other countries consistently, and also that French and Canadian athletes have similar Elo rankings.

References

Elo, A. E. (1978). The rating of chessplayers, past and present. Arco Pub.

Glickman, M. E., & Jones, A. C. (1999). Rating the chess rating system. Chance-Berlin Then New York-, 12, 21-28.

Harkness, K. (1956). Official chess handbook. D. McKay Co.

Appendix

1- http://data.fis-ski.com/global-links/all-fis-results.html

2- https://github.com/holub008/skilo/blob/master/scraper.py

3- https://github.com/holub008/skilo/blob/master/elo_run.py

4- https://github.com/holub008/skilo/blob/master/harkness_run.py