Journal of Data Science 15(2017), 373-390

A Data Analytics Approach to Evaluation of Competition in the 2012 Summer Olympics

John L. Salmon[1] and Willie. K. Harrison[2]
[1]Department of Mechanical Engineering, Brigham Young University,
[2]Department of Electrical and Computer Engineering,
University of Colorado at Colorado Springs,

Abstract: This paper analyzes the competitive nature of a subset of Summer Olympic sporting events by presenting a new data metric that effectively measures the gap between the athletic performance of Olympians in a variety of data visualizations. This metric can be compared across events and across Olympiads so as to determine an event's relative competitiveness as well as the history of the event's competition level. We address our efforts here in identifying several Olympic events that can be classified as *less competitive*, in an effort to aid countries in selecting events in which to compete that could provide an increased probability for medaling. These *less competitive* events can be viewed as opportunities for smaller countries to finally achieve an Olympic medal, or as opportunities for historically successful countries to increase their dominance. We further show how analysis of our metrics specific to any one country can provide an ordered list of Olympic events where the country is nearing the Olympic podium. These results may suggest an event priority ranking for a specific country for investment in improved technology and training techniques.

Key words: data analytics,competition, Olympic sports, performance metrics, data visual-ization

1. Introduction

Increasingly more and more money has been directed towards both the Summer and Winter Olympic Games every two years. China reportedly spent more than $40 billion at the XXIX Olympiad in 2008, while London invested more than $14 billion at the 2012 Summer Olympics. These values were trumped by Russia at the most recent Winter Olympics with an estimated $50 billion in total expenditures for hosting the games in 2014 as reported in Müller (2014).

Although the exact allocation of these funds may be unknown, this growth in overall in-vestment likely increases the funding toward the training and preparing of athletes, especially for the host nation. Recently, even smaller countries are participating more and investing with the intent to produce Olympic medalists for increased national pride and recognition in the international community. The number of participating National Olympic Committees (NOC) from 1996 to 2012 increased from 197 to 204, the respective number of NOCs which won a medal was 79 vs. 85, and all NOCs sent women in 2012 for the first time (see International Olympics Commitee (2015)).

With more countries now competing in each Olympiad, each with improved access to re-cent training techniques and research, updated equipment, and improved coaching (due to the equalizing effects of the Internet and globalization trends), the level of individualperformance is expected to increase over time and concurrently the difference between the top athletes is expected to diminish. In other words, one would expect a continuation of broken world records into the future, while the sporting events themselves have closer finishes and tighter races over

time as each of the events at the Olympic Games becomes more competitive. The former is, of course, evident by the continual breaking of world records in a number of sporting events at every Olympiad. On the other hand, the increase in competition from year to year and within each event is less apparent (or not observable).

The difference between the gold and silver medalists' performances may be ever decreas-ing but the medal podium's discretization of that data masks the true measure of an event's competition level. For example, a second place finisher jumping only half as far as the best performance still receives a silver medal. On the other hand, even if the three medals represent outlier performances from the pool of athlete candidates on any particular day, but the per-formance difference between bronze and fourth place, fourth and fifth place, etc. is narrowing, it suggests that the global pool of athletes is improving and *catching up* to the gold medal performance.

This paper proposes a simple metric to analyze the competition within an Olympic event and discusses some of the findings when this metric is used to compare Olympic events applied in various visualization techniques. Furthermore, based on these analyses, it discusses which sports may be currently *less competitive* and thus a potential area in which those countries without any medals could invest and increase the probability for reaching the medal podium or for relatively successful countries to increase their dominance. Somewhat recently, others have also attempted to identify Olympic events where smaller countries may have more of a chance for medaling, by investigating the history of Olympic events and their winners (as given in Silver (2012)). In this paper, we provide a more mathematically rigorous approach by presenting and analyzing metrics that can quantitatively measure competitiveness, and rank a large subset of Olympic events using the same metric. In other words, our approach can answer the question of which gap is wider between athletes: 10 seconds in a race, or 10 kilograms in weightlifting?

2. Background

2.1 Dataset Description

The 2012 Summer Olympics in London awarded 962 medals in 302 events. Interestingly, those medals are not distributed equally across gold, silver and bronze. As shown in Figure 1, the expected 302 gold medals were awarded but 304 silvers and 356 bronzes (17.8% more than gold) were presented to successful athletes.

The two additional silvers were both awarded in swimming events. Yang Sun and Taehwan Park both finished the mens' 200m freestyle behind Yannick Agnel with the identical time of 1:44.93 min:s and Evgeny Korotyshkin and Chad Le Clos tied with a time of 51.44 s behind

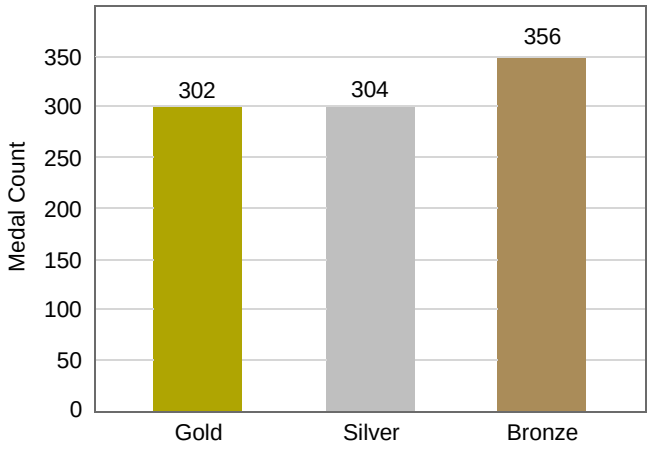

Figure 1: Olympic medals awarded in the 2012 Summer Olympics

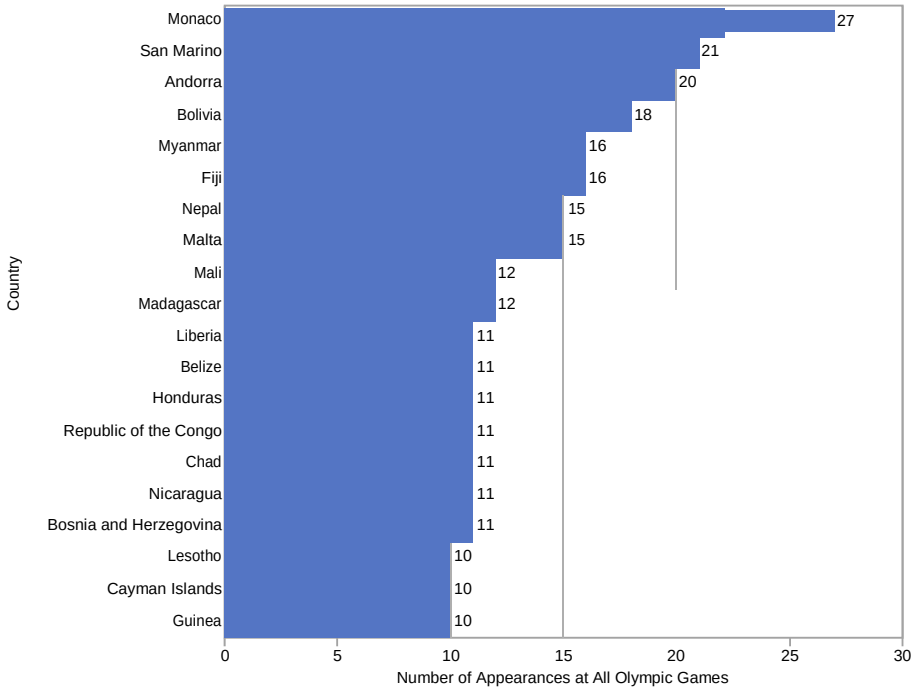

Figure 2: Top 20 number of Olympic appearances for countries with no medals (pre-Sochi 2014)

Micheal Phelp's gold medal in the 100m butterfly. Among the 54 additional bronze medals, most come from combat events where the defeated athletes in the semi-final round both received a medal, 18 are from Wrestling, 8 from Taekwondo, 14 from Judo and 13 from Boxing. The last medal was from the new Keirin event in cycling track where both Teun Mulder and Simon Van Velthooven were awarded bronze. Finally, with the lack of bronze medals in the two swimming events aforementioned, the total of 356 is reached by the addition of three bronze medals awarded in High Jump where Robert Grabarz, Derek Drouin, and Mutaz Essa Barshim all jumped 2.29 m in the competition.

Although, the swimming and high jump results mentioned are unique, the combat sports are consistently awarding 33% more medals per event; first place through *fourth place* go home with a medal. Assuming many factors are similar across events and athletes (i.e. the number of competitors per event is the same, access to training is equal for each athlete, etc.) the chances to go home with a medal are best in these combat events and countries should take note.

In particular, this could be useful information for those countries without any medals in any Olympics but who have multiple Olympic appearances. In Figure 2, the top 20 countries with the most number of Olympic appearances, but with no medals, is presented.

Even with 27 appearances (pre-Sochi 2014), Monaco has yet to find one of their own on top of the medal podium. The other 19 countries, despite each competing in 10 or more Olympiads, also have no medals. The population size, Gross Domestic Product (GDP), and other political, social and economic factors clearly play a role in these results (see Morton (2002); Bernard and Busse (2004); Lozano et al. (2002)). Neural network models have likewise been employed to predict the success of nations (Condon et al. (1999)), but is there something that these and other countries can do differently to improve their chances of becoming a country with an Olympic medalist other than simply improving their economy and increasing their population size? Although appearing at the Olympics is already a prestigious accomplishment for any individual and their country, could an event be identified as *less competitive* and thus more likely for one of these countries to win (or at least medal) with focused investment? Should these countries attempt to compete more in events with two bronzes? Should they compete in

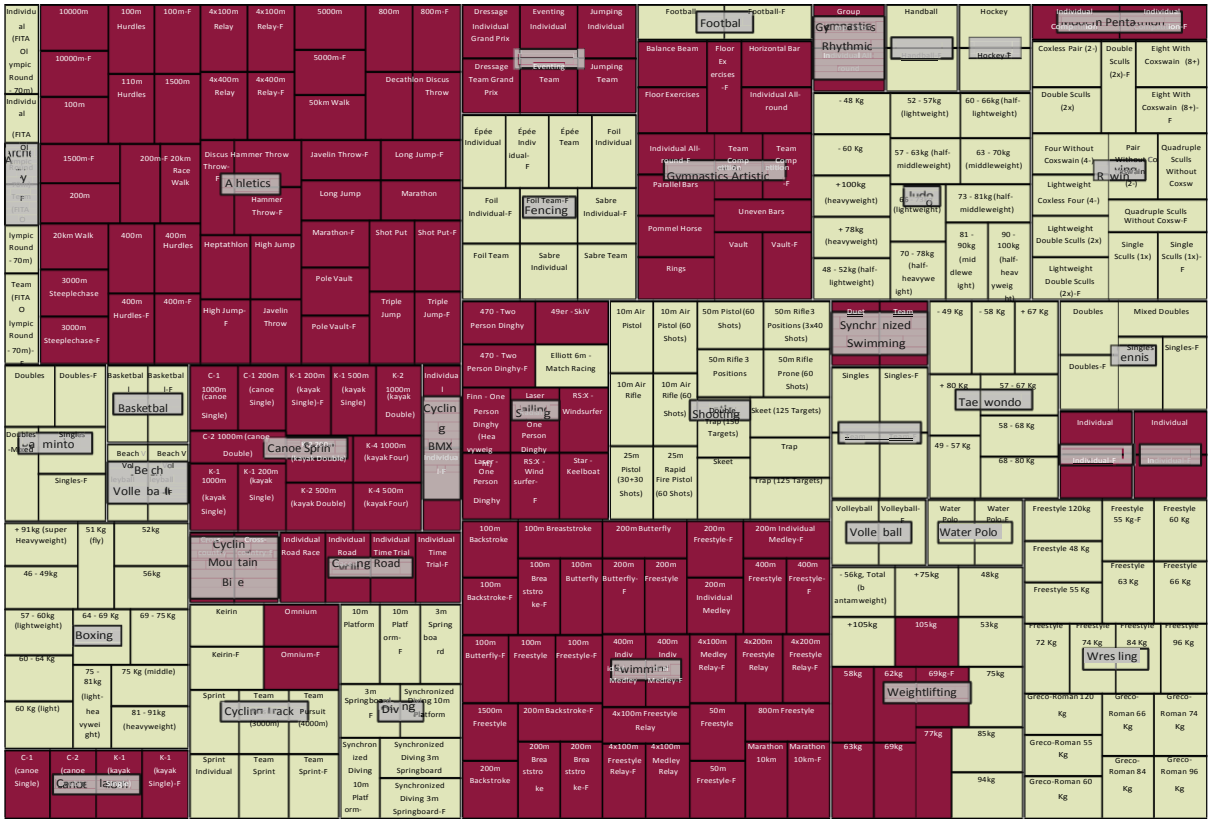

Figure 3: Olympic events with (dark red) and without (beige) quantitative values

an event that is more insensitive to GDP or population size?

From the above discussion, various strategies are clearly available but selecting which one is feasible or appropriate requires some new ways of analyzing, aggregating, and visualizing the numerical data.

2.2 Dataset Preparation

To identify the events that could be rationally compared to each other in some method, the data was first sorted based on available quantitative results. Of the 302 events, only 154 list quantitative values (e.g. time, distance, score) on the summary database available on the official Olympics Games website [3]. The 148 non-quantitative events are typically based on the binary win-loss of matches between competitors in a single elimination or round-robin tournament, or position finish in a heat. When these events are grouped into sport categories, as shown in the tree map of Figure 3, a few useful observations can be used to clean the data further and perform a preliminary down selection of all events for processing (see Johnson and Shneiderman (1991)). The two-level hierarchy presented in Figure 3 demonstrates the classification of each "Event" (the lower level) within each "Sport" (the higher level).

First of all, most Sport categories (e.g. Swimming, Athletics, etc.) are either all quantitative or all non-quantitative with the exception of Match Racing in Sailing and Omnium in Cycling Track. The former event is clearly a tournament based competition and the latter is based on the points accumulated for five different cycling sub-events similar to a pentathlon. The other sport category with a mixture of quantitative and non-quantitative results is Weightlifting. This may simply be an indication that the database is not complete since values for *snatch* and *clean and jerk* were likely recorded elsewhere but not entered int the current state of the database.

[3] Official website of the Olympic Movement, http://www.olympic.org

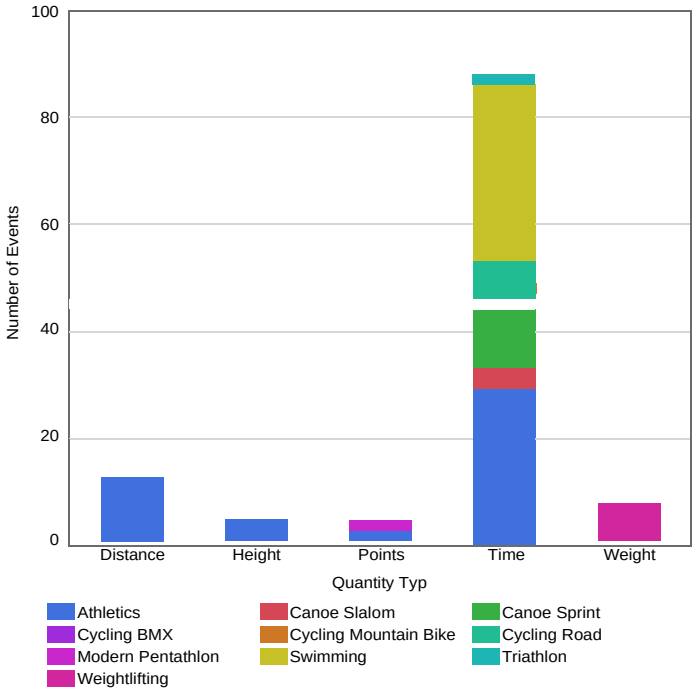

Figure 4: Metric type for 116 Olympic events

Also, groups of events or sports can be collectively defined as *judged events* including Eques-trian, Gymnastics Rhythmic, Synchronized Swimming, Gymnastics Artistic, Cycling BMX, and Trampoline. Due to the subjective nature of these events and the scores, they will be initially removed from evaluation in this study but will potentially be reintroduced in future research directions. This is a similar situation for those 148 non-quantitative events (e.g. Rowing) which naturally have times, or some quantitative values associated with the performance, but only the placement, position, or order of the event is essential to define the medalists.

Another aberration to the common metrics is in the Sailing events where an athlete receives one point for finishing first, two points for finishing second and so on. The athlete with the fewest points after all rounds and heats in these events is awarded the gold. In Tennis, although a combat-type of event, the available quantitative value is the number of matches won in the tournament. Thus, the men's single gold medalist will have a value of 6, followed by silver with 5 matches won, etc. This scale does not lend itself well to comparison in a similar manner (as does time or distance) and has therefore also been removed from the quantitative events analyzed here.

Finally, the Modern Pentathlon and Omnium (along with the Heptathlon and Decathlon) are also points-based taken from a transformation method from *distance to points* or from *time to points*. Although this presents a challenge in providing an absolute comparative technique across all events, this type of scoring will not necessarily cause an event to be removed from our analysis. More will be spoken of this in later sections.

With the removal of the non-quantitative events, judged events, and other events with quantities that are not feasible to use as comparisons (e.g. tennis matches won), the list of events is trimmed down to 116. The types of quantities that measure Olympian performance across all considered events is displayed in stacked bar chart (see Streit and Gehlenborg (2014)) of Figure4. As indicated in this figure, most events (89 events or 79%) use time to measure performance and award medals, with the metrics of height, weight, distance, and points comprising the remaining 21%.The Athletics sport category includes four different types of metrics and makes up 47 of the

116 (41%) events while Swimming events, which only use time as a metric, comprise 34 (29%) of the events.

The distinction is made here since the five data types used in this study do not share the same objective direction (e.g. minimize vs. maximize) and are not normalized in the same way. While many of the respective events want to minimize the performance time, the other four categories seek to maximize the points, height jumped, weight lifted, or distance (thrown or jumped), which serves as a crucial observation in developing the competition metric in the following section.

3. Metric Deftnition

The mechanism we use to measure the competitive nature of these Olympic events is to discern, based on final results, the percentage of the gold-level that was not completed by the silver medalist, and the percentage of the silver-level that was not completed by the bronze medalist. Intuitively, we can then make statements such as *the silver medalist trailed the gold medalist by 2%,* and similarly for the bronze medalist with respect to the silver. The measures provide information as to the competitive nature of the battle for gold (or silver), by effectively measuring the normalized gap between competitors. This analysis can clearly be extended to the *i*th rank in an event so as to evaluate the competitive nature of the battle for *i*th place.

3.1 Timed Events

To be specific, in a timed event, let t_1 be the gold medalist's finishing time, and similarly t_2 and t_3 are the respective silver and bronze medalists' finishing times. Then, we measure the normalized gap between the gold and silver medalists as

$$C_1 = \frac{t_2 - t_1}{t_2}, \tag{1}$$

and the normalized gap between the silver and bronze medalists as

$$C_2 = \frac{t_3 - t_2}{t_3}, \tag{2}$$

These metrics are designed so that C_1 conveys the relative distance of the race that the silver medalist has yet to cover when the gold medalist crosses the finish line, and similarly for C_2. An estimate of the location of the silver medalist at the time the gold medalist crosses the finish line can be given by assuming a constant velocity for the silver medalist throughout the race. While this is not realistic, the model still provides a reasonable estimate for our purposes. Let v_2 be the constant velocity required by the silver medalist to complete the race in time t_2. Let the total distance of the race be simply d. Then, the distance covered by the silver medalist in time t_1 can be estimated by

$$d_2 = v_2 t_1 = \frac{d}{t_2} t_1. \tag{3}$$

Now, the relative fraction of the race yet to be covered by the silver medalist at time t_1 can be estimated as

$$C_1 = 1 - \frac{d_2}{d} = 1 - \frac{\frac{d}{t_2} t_1}{d} = 1 - \frac{t_1}{t_2} = \frac{t_2 - t_1}{t_2}. \tag{4}$$

This completes our derivation of C_1, and the derivation for C_2, or any C_i, follows an identical line of reasoning. More generally, we can measure the normalized difference between the *i*th and *j*th ranked Olympians in timed events by calculating

$$C_{i,j} = \frac{t_j - t_i}{t_j} \tag{5}$$

where it is assumed that $j>i$, but j need not be equal to (i+1). In this paper, we will always use consecutive finishing ranks i and j=(i+1) for analysis of competitions in the Olympics. Thus, we need only reference the best finishing rank i when labeling the value as Ci.

Example 1 The men's marathon race in the London 2012 Summer Games resulted in Stephen Kiprotich of Uganda winning gold with a time of 2:08:01 hr:min:s. The silver and bronze medalists, Abel Kirui and Wilson Kipsang Kiprotich, were both from Kenya and finished with respective times of 2:08:27 hr:min:s and 2:09:37 hr:min:s.

To measure the competitive nature of the competition for gold, we first convert these times to a common unit, e.g., seconds, and calculate

$$C_1 = \frac{t_2 - t_1}{t_2} = \frac{7707 - 7681}{7707} = 0.00337\,. \tag{6}$$

This fractional gap corresponds to a gap of 0.337% that the silver medalist lacked compared to the gold, which seems highly competitive. A similar analysis is completed to determine the competitiveness of the competition for silver, which results in

$$C_2 = \frac{t_3 - t_2}{t_3} = \frac{7777 - 7707}{7777} = 0.009\,, \tag{7}$$

or, in other words, the bronze medalist lacked 0.9% of the silver medalist's capability in that race, also seemingly quite competitive.

3.2 Scored Events

The derivation is simpler for events where competitors either accrue points, or progress throughout the event to a maximum ability, e.g., weightlifting, high jump, etc. The four data types in Figure 4 besides time all benefit from the following definitions of C_1, C_2, etc., and may collectively be referred to as *scored events*. Here, we need only consider the normalized fraction of the gap between medalists so that if the final score (or amount lifted or height jumped, etc.) by the gold medalist is S_1, and the final score for the silver medalist is S_2, then the normalized gap between the two medalists is

$$C_1 = 1 - \frac{S_2}{S_1} = \frac{S_1 - S_2}{S_1}\,. \tag{8}$$

Notice the difference between timed events and scored events is effectively the normalizing divisor after the difference in final scores is obtained. For timed events, we use the larger of the two times to normalize, e.g., the silver medalist's finishing time t_2 for calculating C_1, while in scored events (events where a competitor wants to achieve as high a final score as possible) the divisor is S_1, the score of the gold medalist. Thus for scored events, the relative gap between the silver and bronze medalists can be measured by

$$C_2 = \frac{S_2 - S_3}{S_2}\,, \tag{9}$$

where S_3 is the final score of the bronze medalist. Similarly as was shown for timed events, in the most general sense we can calculate the normalized difference between the ith and jth finishing ranks in scored events by the expression

$$C_{i,j} = \frac{S_i - S_j}{S_i}\,, \tag{10}$$

where $j>i$. Such a metric could be used to identify how close a low finisher is to competing for an Olympic medal. For example, C_3,10 compares the bronze medalist to the 10th place finisher, and may indicate in some cases that the 10th place finisher is closer to mounting an Olympic podium than a fourth place finisher is in a different event.

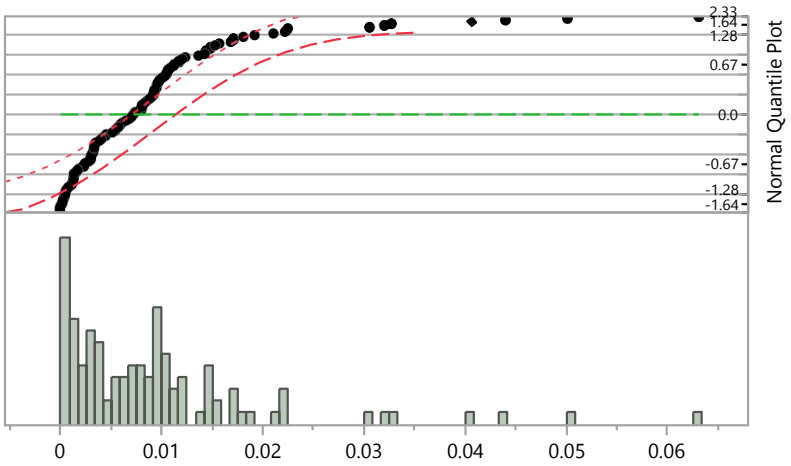

Figure 5: One-dimensional C_i value analysis: distribution of C_1

Example 2 In the women's high jump competition at the London 2012 Summer Olympic Games, Anna Chicherova of Russia won gold clearing 2.05 meters on her second try. Brigetta Barrett of the United States and Svetlana Shkolina of Russia both cleared 2.03 meters, but Bar- rett cleared the height on her second try, while Shkolina took three tries to clear the height. Thus, Barrett won silver and Shkolina won bronze. Observing only the final heights cleared, we calculate

$$C_1 = \frac{S_1 - S_2}{S_1} = \frac{2.05 - 2.03}{2.03} = 0.00976\,, \tag{11}$$

or 0.976%, while

$$C_2 = \frac{S_2 - S_3}{S_2} = \frac{2.03 - 2.03}{2.03} = 0. \tag{12}$$

While we may conclude that the competition for gold was competitive, our metrics indicate a dead heat between silver and bronze medalists. There are other ways to factor in the number of misses into the final score of the Olympians, but based only on the final performance results, the event would be classified as highly competitive.

4. One-dimensional C_1 Value Visualizations

Applying the above defined metric $C1$ to each of the 116 events in the feasible subset for the 2012 Summer Games, and sorting with respect to this metric identifies that Weightlifting events comprise five of the top seven *least competitive* events (those with the largest $C1$ values). Furthermore, the *least competitive* event was Javelin Throw for women, where the silver medalist score (distance thrown) was more than 6% below the gold medalist. On the other end, four events are identified with a $C1$ of zero, suggesting a *photo finish* type of event in three of these, namely the women's triathlon, and the individual cycling road race for both genders. The other events with high competitive scores (and thus low $C1$ values) include a variety of longer races such as the 10,000m run where only 0.029% separated the gold from the silver medalist, an expected physical difference of $\sim$ 2.9 m, or roughly two stride lengths.

Of course, all other events fall between these two extremes, with 79 of 116 events having a C_1 less than 1%. Only 11 have a C_1 greater than 2%. Figure 5 aggregates these values showing the distribution of the 116 C_1 calculations. As evidenced from the normal quantile plot above the histogram, the performance between the gold and silver medalists across Olympic events in not distributed normally (see Ott and Longnecker (2008)). The shape of this distribution shows

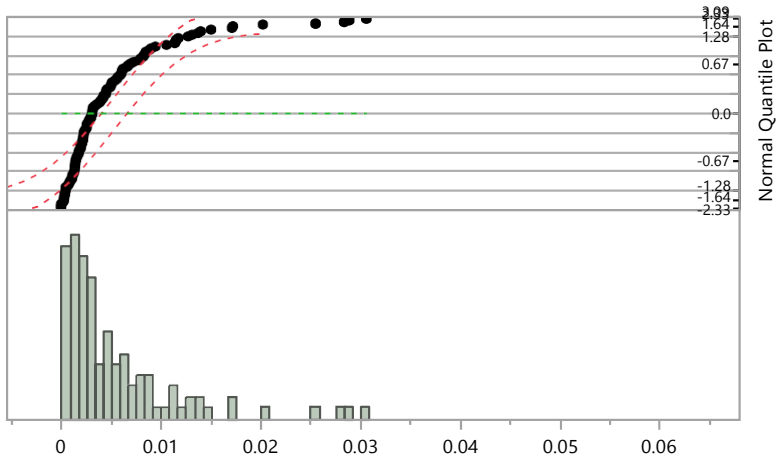

Figure 6: One-dimensional C_i value analysis: distribution of C_2

the piling-up against the natural constraint at zero. Thus, at a *more competitive* Olympiad, the distribution would be expected to shift closer to zero with both a lower mean and variance.

We obtain Figure 6 for the C_2 values using the same analysis as in Figure 5. With the distribution plotted over identical x-axis ranges, it can be observed that the mean and variance of the C_2 distribution is smaller than that compared to the C_1 distribution. This was confirmed with the numerical values for mean μ and standard deviation σ for C_1 and C_2 respectively, where $\mu_1 = 0.0092$, $\sigma_1 = 0.0104$, $\mu_2 = 0.0053$, and $\sigma_2 = 0.0061$.

Interestingly, therefore, the overall competition for the silver medals at the 2012 Olympics Games were more competitive than that for gold medals based on analyses of the C_1 and C_2 metrics. In some sense then, we can consider gold medalists as outliers who find ways to achieve larger relative gaps in performance than the other medalists. One can hypothesize that a percentage of the outlier performance by the gold medalist (e.g., 95% of $s1$) would be approached by a greater proportion of athletes over time. However, since more athletes could approach the silver medalist performance, the C_2 value in general should be lower and the competition for the silver medal should then be *more competitive*, especially as a larger number of athletes approach the human physical limit of performance (assuming it exists). This theory and its applicability to lower degree Ci values (i.e. C_3, C_4, C_5, etc.) will be tested and explored in a later section.

To further investigate the trend that the mean for C_2 is generally smaller than that for C_1 a sample of 500 hypothetical performances are taken from a normal distribution representing a generalized Olympic sport from a larger population. The C_1 and C_2 values are then calculated from the three extreme outliers. This process is repeated 1000 times to acquire a distribution of C_1 and C_2. As shown in Figure 7 the distribution for C_2 does in fact indicate a smaller mean and variance than those for C_1.

5. Two-dimensional C_i Value Visualizations

The previous conclusion is supported by plotting $C1$ versus $C2$ metrics on a single graph as shown in Figure 8 with the bivariate scatterplot and univarite histograms on the sides (see Scott (2009)). Let us now refer to a C value pair, where

$$C = (C1, C2). \tag{13}$$

The data is projected into the same histograms on the top and right, shown previously in Figures 5 and 6 respectively. Assigning identical ranges to the x and y axes clearly shows that

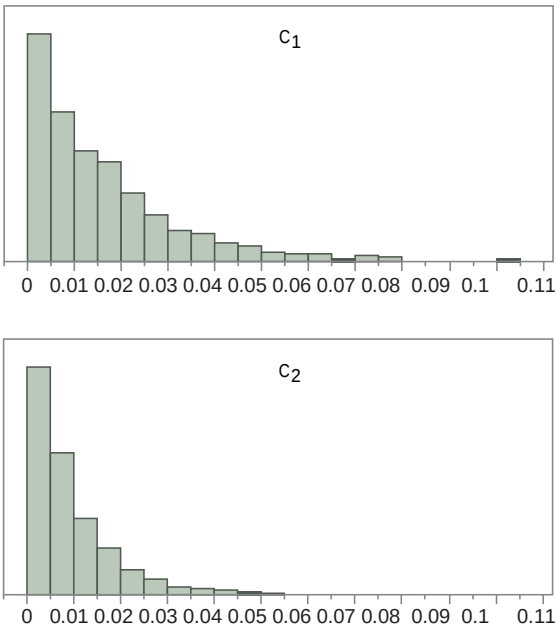

Figure 7: One-dimensional C_i value analysis: comparing expected C_1 and C_2 distributions

the points are located in the bottom half of the graph or at lower C_2 values. This is further established by the trend line of the data from a linear regression approach suggesting that the competition for the silver medal is *more competitive* in comparison to the gold.

In Figure 8, the events far removed from the origin are of particular interest and warrant additional investigation since one or both of these Ci values can be characterized as less com- petitive. Thus, the same data is plotted on Figure 9 with co-centric circles overlaid (with radii increments of $0.01n$ for $n = 1, 2, \ldots, 5$), so as to characterize and group events in terms of their C value pairs. As pairs move away from the origin, the event is deemed less and less competitive. Furthermore, the individual events are encoded with colors and markers to improve identifying trends in the C value space. Events with C value pairs outside the circle with radius 0.02 are labeled.

One apparent trend which is readily identified is that many of those events outside the 2% circle (i.e. outside circle with radius 0.02) are from Athletics and Weightlifting. Further consideration suggests that some of these events have *low resolution* in the discretization of scoring methods such as weightlifting or high jump where each successive lift or jump respec-tively is augmented by a set value for future attempts. Decathlon and Heptathlon also include a non-continuous scoring method (points) while the time-based events approach more closely a continuous scale (often down to a resolution of hundredths of a second).

Even with the limitations of these *non-continuous* scoring scales in these events, a couple of propositions can still be made which are in opposition to each other and encourage additional inquiries: 1) the traditional or older events (such as those in Athletics) have been around for so many decades that sufficient time and experience has allowed athletes to *fill in the distribution* and reach *outlier status,* leaving a larger gap between the gold medalists and the next best athlete, or 2) the event is *younger* or less popular and has not been saturated sufficiently with a long history of athletes and so large differences in athletic performance are observed.

For example, the C_1 value derived from the men's Shot Put comes from a differential distance of just 3 cm, while the C_2 value comes from a differential of 63 cm (a differential 21 times larger than from the C_1 value). The questions can be asked, are there fewer athletes to *fill in* the performance betwee these two extremes or was the tight competitive performance between the gold and silver medalist a random occurrence? If the answer is yes to the former question,

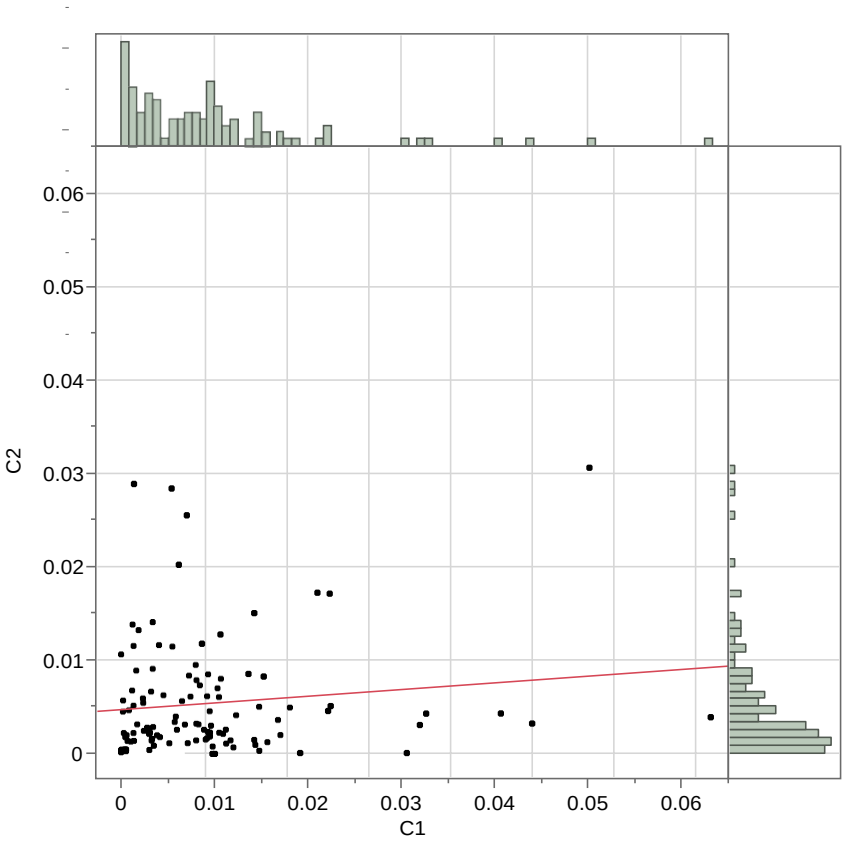

Figure 8: Two-dimensional *Ci* value analysis

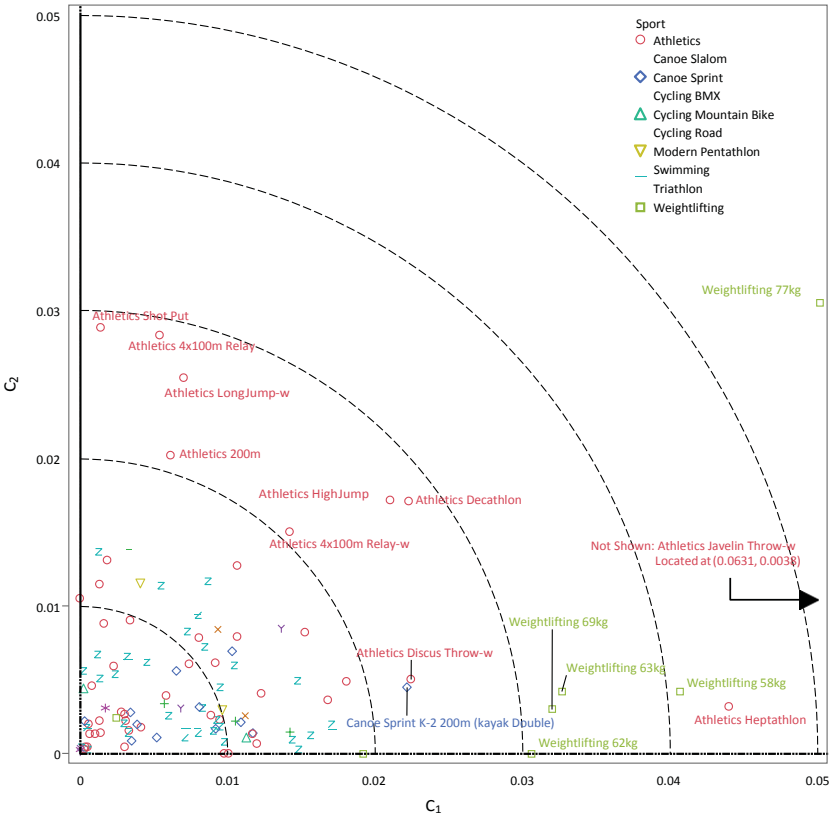

Figure 9: Two-dimensional *Ci* value analysis with labeled events beyond radius 0.02

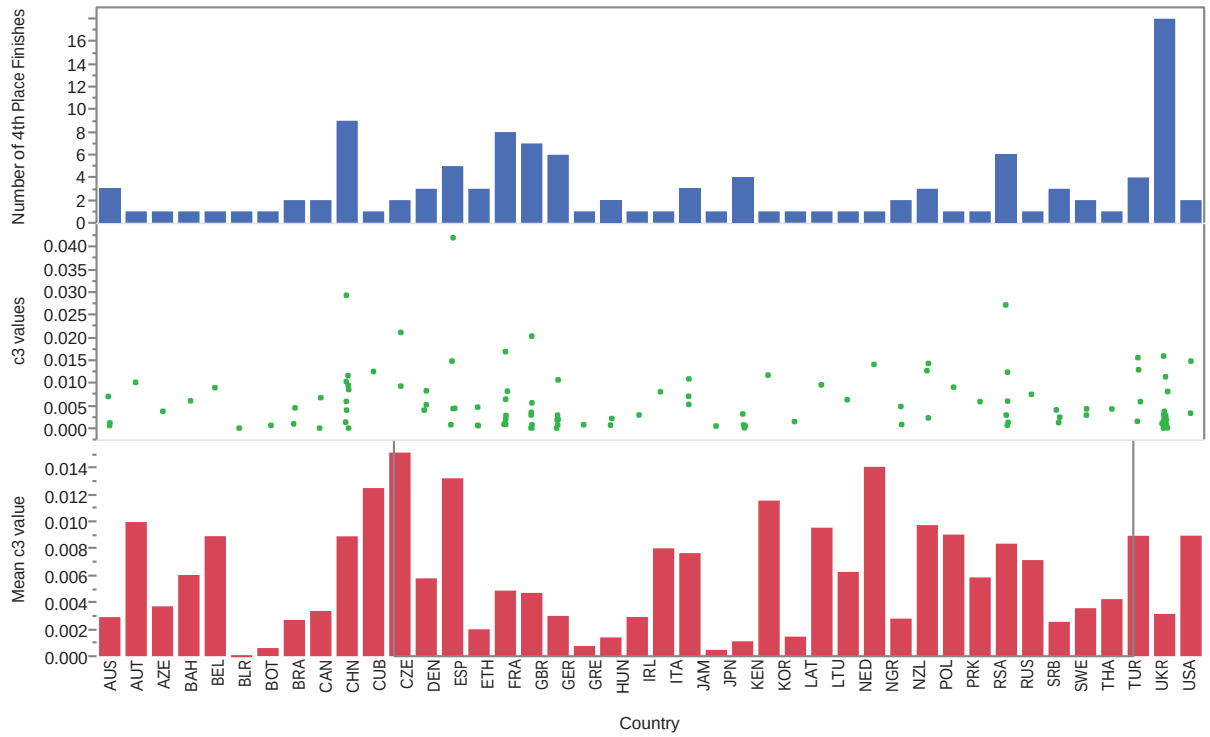

Figure 10: Number of fourth place finishes by country with associated $C3$ values and means

then this event could be identified as one which should be pursued by a country seeking for representation on the podium, assuming all else is equal.

Similarly, the $C1$ value for women's Weightlifting in the 58kg class is greater than 0.04. This was derived from a 10 kg difference between the total weight lifted between the gold and silver medalist. However, the true difference could have been even greater if Li Xueying had been successful in her attempt of breaking the world record with a 144 kg clean and jerk lift. She still won the gold medal with her successful lift of 138 kg but the $C1$ could have been higher if she chose to lift a weight between these two amounts. Therefore, the rules of the event and the strategy taken by the weightlifting athletes themselves may be a factor in causing the large $C1$ values, but this still leaves open the possibility of approaching the gold medalist by additional training, technique, or strategy improvements. Furthermore, weightlifting is an event with two *subevents* (i.e. *snatch* and *clean and jerk*) and thus the deserving gold medalist has *twice as many chances* to expand the distance between first and second place.

6. Application of the *Ci* Metric

6.1 A Case Study: USA

In order to further down select the potential Olympic events in which a country could invest to maximize the probability of reaching the podium, the *Ci* metric can be combined with other metadata available about each event. For example, Figure 10 shows the number of fourth place finishers for all 116 events distributed across the respective countries. The USA, for example, finished 18 times in fourth place (more than 15% of the 116 events considered). If one assumes that the effort in finishing one place higher (i.e. finishing third) to obtain a medal is easier than multiple place increases (e.g. sixth to third), a country should consider focusing, at least ini-tially, on those events where they are currently *just off* the podium for prioritized investments. Furthermore, sorting those events with the $C3$ values would provide additional event prioritiza-tion or at least indicate which event's bronze medal is most competitive (i.e. lowest $C3$ value)

Figure 11: Number of US team finishes in the top four when a US athlete finished fourth

and thus a higher probability for one's athlete to overcome the smaller performance differential. In the multi-coordinated view (see Baldonado et al. (2000)) in Figure 10, the C_3 values are shown in clusters for each country. The cluster for USA shows 15 of the 18 C_3 values less than 0.005. The mean value for those 18 C_3 values is approximately 0.003, one of the lowest mean values for the 42 countries that finished fourth in at least one event. This suggests a number of potential opportunities for the USA to increase their medal count. Taking the ratio between the number of fourth place finishes over the mean C_3 values presents a weighted average that indicates potential for increase in medals by a country. Using this technique, the USA has the highest ranking, followed by Kenya and then Japan, indicating that the USA has perhaps the highest potential of any country to increase overall medal count in the next Summer Olympiad.

If we investigate further, however, this potential is perhaps reduced for the USA because there are already American athletes on the podium in several of the events where the US finishes in fourth place. Figure 11 lists these 18 events by the sorted number of US athletes who finished in the top four. Any event with more than one finisher suggests a US medalist whereas the 10 events with only one (fourth place) finisher is indicative of a podium void of any US athlete. The eight events with a US medalist are events where sufficient training, resources, efforts, etc. are already bearing fruits in the form of medals, while the other 10 events show potential for more medals with perhaps relatively small additional investments. These 10 events can be investigated with respect to their own individual C_3 values as shown in Figure 12. The closest fourth place finish for the USA in London 2012 was in the Cycling Road men individual road race. In this case, Taylor Phinney of the United States finished a more than 5 hour long race within the same second as the bronze medalist, coming across the finish line second in a pack of 23 riders, all of whom were racing for a bronze medal (the gold and silver medalists were eight seconds faster than this group of 23 riders). Expanding this study to several of the latest Olympiads may show trends in events where the US routinely secures fourth-place finishes. A more comprehensive study of, say fourth through eighth place finishes may also be useful in identifying those events where a slight amount of improvement results in a more competitive USA Olympic Team.

6.2 Applications for Smaller Countries

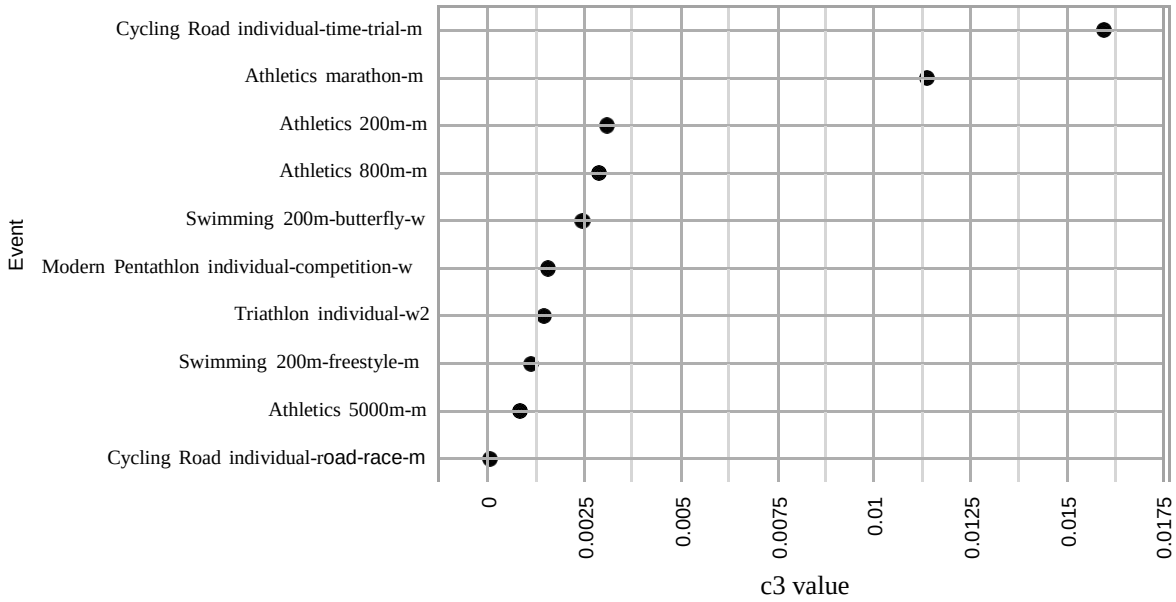

Figure 12: Events with US Team fourth place finisher with no medal

Now what about the countries from Figure 2, mostly smaller countries who have been competing in the Olympic Games for between 10 and 27 Olympiads with no medals earned to date? Our analysis can help these countries in choosing less competitive events in which to compete, and in monitoring their progress towards the podium over Olympiads. Certainly the events previously described as less competitive (i.e. many of the Weightlifting events) may be reasonable candidates for countries to attempt to win medals. A more in-depth study of the three Ci values that are calculated from medal-winning performances may also shine a light on events that can be leveraged in an attempt to win a medal. Figure 9 shows some of these events to be Weightlifting events, Athletics events, and one Canoeing event.

6.3 Comparison to the 2016 Olympics

In both the previous case studies, a country would want to verify these trends over Olympiads, rather than hastily investing in Olympian development in events that are less competitive in 2012 alone. This final analysis compares the preceding events between the 2012 and 2016 Summer Olympics.

Performing the same procedure to the 116 events (with quantitative measurements) from the 2016 Olympics to calculate the C_1 and C_2 metrics enables a comparison of the statistics from the two Olympiads under analysis. Table 1 presents the minimum, maximum, mean, standard deviation, and median statistics for both C_1 and C_2 values of all 116 events for the 2012 and 2016 Olympics. Interestingly, a minimum Ci value of zero is observed for all four categories suggesting that at least one event results in an effectual tie for both gold and silver medals. Additional analysis shows that at least 3 events resulted in ties with as many as 7 ties in 2016 (for C_1 values). However, not all "ties" result in awarded medals. For example, in the men's 10K swimming marathon, after a recorded time of exactly 1:52:59.8 for both the first two finishers, photo analysis revealed the gold medal would eventually go to Ferry Weertman from the Netherlands. Other events such as high jump may have identical Ci values but additional rules will be applied for awarding medals in these situations.

Some statistics in Table 1 suggest a reduction in competitiveness from 2012 to 2016. The maximum C_1 increases from 0.063 to 0.074 suggesting at least one sport became less competitive with a higher C_1 value in 2016 compared to 2012. The same phenomenon is observed for the

Table 1:Statistics Comparison of C_1 and C_1. from 2012 and 2016 Olympiads

	C_1 - 2012	C_1 - 2016	C_2 - 2012	C_2 - 2016
Minimum	0	0	0	0
Maximum	0.063	0.074	0.031	0.047
Mean	0.009	0.010	0.005	0.007
Std. Dev.	0.010	0.012	0.006	0.009
Median	0.007	0.006	0.003	0.003

silver medals (i.e. a higher C_2 maximum value in 2016). Furthermore, mean values increase from 2012 to 2016 while the median values decrease. This is indicative of a slight distribution skew towards lower Ci values suggesting that some events became more competitive while others less competitive, resulting in a necessary event by event comparison between the years. This is visualized in a parallel plot of Figure 13.

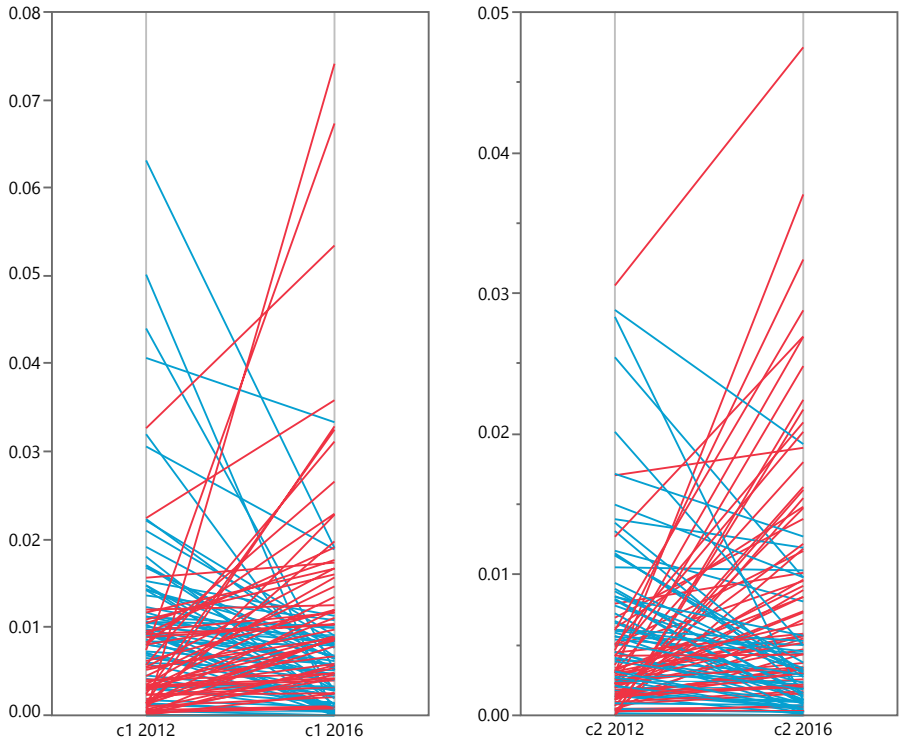

Figure 13:Comparison of C_1 (left) and C_2 (right) values between the 2012 and 2016 Olympics

All the events which became less competitive in 2016 for the gold medal are colored in red on the left hand side of Figure 13 while the events that became more competitive are in blue. This color scheme is repeated for the C_2 values on the right hand side of Figure 13. The number of events that became less competitive for C_1 values is 53 and for C_2 is 59. Therefore, more than half of the events under consideration became less competitive for the silver medal in 2016 and almost half for the gold medal. Graphing the 2016 Ci scores onto the results presented previously in Figure 9 results in Figure 14. Each line segment represents the movement of the point C_1, C_2 in 2012 to 2016 for each event. The open and closed circles or points subtending each line segment represent 2012 and 2016 respectively.

Of most interest are the Olympic events (i.e. the line segments) that are less competitive for both years and thus located relatively far from the origin. This would suggest a large gap in performance for both the silver and gold medals for both years. Furthermore, any line segment with a positive slope with the closed circle further away from the origin represents events that

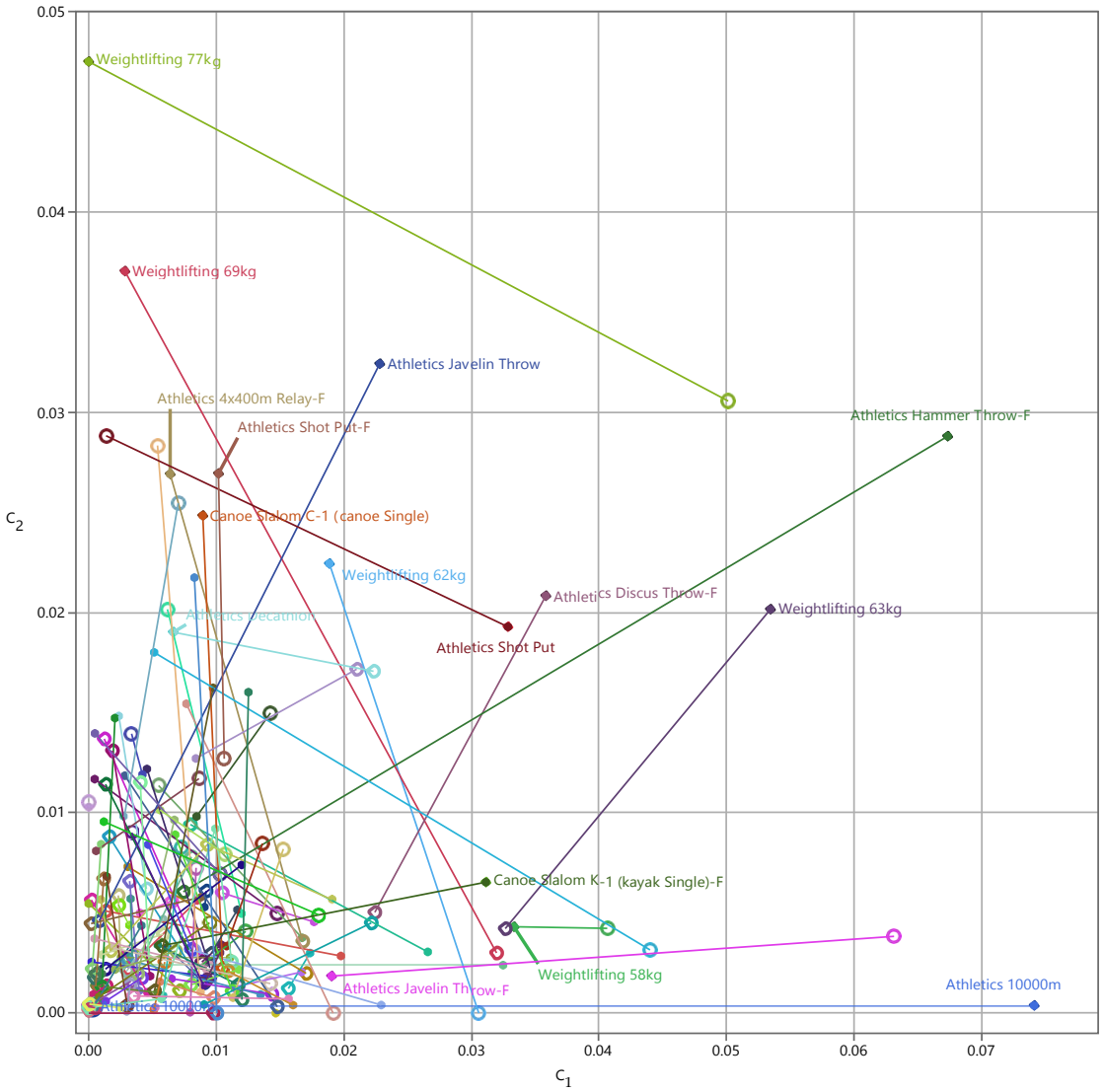

Figure 14:Two-dimensional *Ci* value analysis for 2012 (open circles) to 2016 (closed circles)

became less competitive for both the gold and silver medals. A few of these events are labeled in Figure 14 including the Women's Hammer Throw, Men's Javelin Throw, Women's Discus Throw, and Weighting Lifting (63kg). Other events, positioned at a large relative distance from the origin, have either the gold or silver medal competitions becoming less competitive such as Weightlifting (77kg) and Men's Shot Put. These are also labeled in Figure 14 as well as other events that suggest the competitiveness of these events are potentially not as high. Overall, this figure suggests that of all the sports some events maintain a low level of competitiveness in recent history and over time which could be exploited by an individual athlete or country seeking additional medals.

7 Conclusion and Future Work

One limitation of the aforementioned findings is that we only consider those events that are readily applicable to *Ci* calculations (116 of the 302 total events), but there could exist several other events that are less competitive than those we have outlined. Expanding our results to consider all Olympic events is a clear open area for expansion of this research. In particular, we can extend our results to the Winter Olympic Games. At the Winter Olympics, many events are

already quantitative using time measures, etc. and would yield themselves to the simple direct application of the analysis in this paper, perhaps highlighting more events that could prove wise investments for Olympic teams around the globe. In terms of judged events, tournament style events, and scored events that we have yet to consider, we will need to adopt normalizing strategies in all cases that can result in *Ci* calculations across the board such as intermediate points awarded in combat events or tournaments.

Although mentioned in this paper, the more specific metric $C_{i,j}$ allows us to make compar-isons of an *i*th place finisher and a *j*th place finisher for any i and j. It may prove useful to analyze all $C_{3,j}$ metrics for j 3 across events so as to more appropriately order events in which a country exhibits potential for winning medals. In some events, finishing 10th may indicate a wide chasm between the medal podium and the athlete, whereas in others, the fractional differ-ence could be incredibly small. Although we may be quick to point out that a 10th place athlete was outperformed by nine other athletes, and therefore, is very far from an Olympic podium, an extremely small $C_{3,j}$ value may actually indicate that the third through 10th place athletes are more or less equal in capability, and any of them are potential Olympic medal winners. In other words, small environmental factors may actually determine the medalist on any given day, rather than an actual superiority in skill level. This more detailed metric also gives us a mechanism for comparing two athletes who finish well off the podium in their respective events. We may wish to ask, which of them is closest to winning a medal, and their $C3_{,j}$ values can provide this information even if the athletes compete in different events. Additionally, athletes that compete in more than one event may want to know how close they are to medaling in each of their events so as to more properly distribute their training time and effort.

Finally, after having characterized all Olympic events as *competitive* or *less competitive*, it will also be useful to apply additional methods to the available data sets that can determine how to improve athletes in specific events. Can we essentially follow the sabermetricians of baseball like Baumer and Zimbalist (2014) and break down Olympic events so as to understand what percentage of a medal is won by following different strategies, attaining certain goals, etc? The Olympic Games provide a fascinating stage for statistical and visual analysis for athletic performance for several reasons: (1) the athletes in question are, by definition, world class, (2) the world as a whole is deeply interested in the outcomes of the Olympic games, (3) the variability in sports and events covered by the Olympics is sufficiently large that there will always be something new to discover, and (4) properly visualizing data can assist in better decisions by both athletes and countries.

References

[1] Baldonado, M. Q. W., Woodruff, A., and Kuchinsky, A. (2000). Guidelines for using multiple views in information visualization. In *Advanced Visual Interfaces*, pages 110–119.

[2] Baumer, B. and Zimbalist, A. (2014). *The Sabermetric Revolution: Assessing the Growth of Analytics in Baseball*. UPCC book collections on Project MUSE. University of Pennsyl-vania Press, Incorporated.

[3] Bernard, A. B. and Busse, M. R. (2004). Who wins the Olympic Games: Economic resources and medal totals. *The Review of Economics and Statistics*, 86(1):pp. 413–417.

[4] Condon, E. M., Golden, B. L., and Wasil, E. A. (1999). Predicting the success of nations at the summer olympics using neural networks. *Computers & Operations Research*, 26(13):1243-1265.

[5] International Olympics Commitee (2015). Women and Sport: Leaping forward .

[6] Johnson, B. and Shneiderman, B. (1991). Tree-maps: A space-filling approach to the visual-ization of hierarchical information structures. In *Visualization, 1991. Visualization'91, Proceedings., IEEE Conference on*, pages 284–291. IEEE.

[7] Lozano, S., Villa, G., Guerrero, F., and Corts, P. (2002). Measuring the performance of nations at the Summer Olympics using data envelopment analysis. *The Journal of the Operational Research Society*, 53(5):pp. 501–511.

[8] Morton, R. H. (2002). Who won the Sydney 2000 Olympics?: An allometric approach. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 51(2):147–155.

[9] Müller, M. (2014). Introduction: Winter Olympics Sochi 2014: What is at stake? *East European Politics*, 30(2):153–157.

[10] Ott, R. and Longnecker, M. (2008). *An Introduction to Statistical Methods and Data Analysis*. Available 2010 Titles Enhanced Web Assign Series. Cengage Learning.

[11] Scott, D. (2009). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. Wiley.

[12] Silver, N. (2012). Let's play medalball! *The New York Times*, page MM40.

[13] Streit, M. and Gehlenborg, N. (2014). Points of view: Bar charts and box plots. *Nature methods*, 11(2):117–117.

John L. Salmon
Department of Mechanical Engineering Brigham
Young University
Provo, UT 84604

Willie. K. Harrison
Department of Electrical and Computer Engineering University of
Colorado at Colorado Springs
1420 Austin Bluff Parkway, Colorado Springs, CO 80918