

EXPLORING COMPETITION IN NASCAR

A THESIS

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE

MASTER OF SCIENCE

BY

CHRISTOPHER A HURLEY

BALL STATE UNIVERSITY

MUNCIE, INDIANA

MAY 2019

Contents

- 1 Acknowledgements** **4**

- 2 Introduction** **5**

- 3 Current Measures of Competition** **6**
 - 3.1 Normalization 6
 - 3.2 Winning 8
 - 3.3 Competitive Balance Ratio 8
 - 3.4 Concentration Ratios 9
 - 3.4.1 Normalization for Concentration Ratios 10
 - 3.4.2 Normalization for Top X Finishes for Concentration Ratios 10
 - 3.5 Herfindahl-Hirschman Index 11
 - 3.5.1 Normalization of HHI 12
 - 3.5.2 Normalization for Top X Finishes for HHI 12
 - 3.6 Ideal Standard Deviation Ratio 13
 - 3.6.1 Normalization for Idealized Standard Deviation Ratio 14
 - 3.6.2 Normalization for Top X Finishes for Idealized Standard Deviation Ratio 15
 - 3.7 Churn 16
 - 3.8 Example 17

- 4 Data and Calculations** **18**
 - 4.1 Data 18
 - 4.2 Calculated Metrics 19
 - 4.2.1 Adjusted Churn 19
 - 4.2.2 Wins 19
 - 4.2.3 Top 5 Finishes 21
 - 4.2.4 Top 10 Finishes 21

| | | |
|----------|-------------------------|-----------|
| 4.2.5 | NCR_5 | 23 |
| 4.2.6 | NCR_{10} | 23 |
| 4.2.7 | HHI^* | 24 |
| 4.2.8 | IRN | 25 |
| 5 | Conclusions | 28 |
| 6 | Further Research | 28 |
| 7 | References | 30 |

1 Acknowledgements

I would like to thank my entire thesis committee for their assistance with both the thesis itself and my thesis presentation. Without Rebecca Pierce there is no way I would have been able to submit either on time or with this quality of work. I would like to especially thank both Drew Lazar and Todd Nesbit for their contribution; for the majority of my research it felt like I had two thesis advisors to help me with the process. Additionally, I want to thank both James McClure and Erik Nesson for their suggestions during my presentation, as well as Guy David for his assistance with determining upper and lower bounds for some of the measures.

2 Introduction

Measures of competitive balance are used to test the uncertainty of outcomes hypothesis. As is discussed by Knowles et al. (1992), the uncertainty of outcomes hypothesis is the idea that fans are more interested in watching games or races when it is more difficult to predict the outcome. In turn, fan interest translates to a higher revenue for sports leagues. In this thesis, because of a lack of data and difficulty of measurement of fan interest, I do not test the uncertainty of outcomes hypothesis. Instead, I deal with exploring season-long measures that are less affected by the unique issues that racing sports, specifically the National Association for Stock Car Auto Racing (NASCAR), have. Then the season-long measures can be used with some measure of “popularity” of the sport to test the uncertainty of outcomes hypothesis.

In any given race in NASCAR there are approximately 40 drivers, denoted d . These d drivers are not the same from race to race, or even year to year, which makes measuring competitive balance more complicated. For each race in the seasons considered in this thesis, drivers had to qualify by driving around the track by themselves. The top d fastest qualifying times gave the order which the drivers start in for the race that week. It is common, especially in the past, for the total number of drivers over a season, D , to be significantly higher than d . D is often significantly higher than the number of races in a year, N . Recently seasons have been $N = 36$ races long, but they have been as short as $N = 28$ races.

In head-to-head sports like soccer or basketball, in each competition there are two competitors with a single winner and a single loser (ignoring ties). This makes it very clear what winning means: if the opposing competitor or team lost then you won. Unlike head-to-head sports, while there is a single winner in a NASCAR race, there is more than one loser. It is reasonably clear that just because a driver does not win a particular race that does not mean they are a loser because in addition to wins, NASCAR records both top 5 and top 10 finishes. This makes it difficult to apply to NASCAR the measures of competition that are used in head-to-head sports. At least partially because of this fact, research involving

competition in NASCAR often deals with individual races, by Berkowitz et al. (2011) for instance, rather than a single season-long metric. The multiple winners and non-winners in NASCAR also make it difficult to compare the level of competition to other sports.

A possible way to solve this problem is to consider more than just the winner of each race in measuring competition. In this way, generalizing commonly used measures of concentration and competition to NASCAR seasons seems appropriate. These generalized measures then allow comparisons between head-to-head sport seasons and NASCAR seasons.

3 Current Measures of Competition

3.1 Normalization

Many of the measures that I examine have upper and lower bounds that depend on the structure of the league or the season being measured. In general, however, these upper and lower bounds are achieved when wins are most or least concentrated, respectively. Normalization recasts these measures as the proportion of the linear difference between least and most competitive and thus provides comparable measures between seasons and sports leagues with different structures. After “normalization” the value of the measure will be between 0 and 1, where 0 means roughly that there is either “perfect” competition or parity in the league, and 1 means the league is not competitive at all. Formally we have

$$M_N = \frac{M_C - M_{lb}}{M_{ub} - M_{lb}} \tag{1}$$

where M_C is the calculated value for the metric, M_{lb} and M_{ub} are the theoretical lower and upper bounds, respectively, for the measure, and M_N is the normalized value for the metric. This properly scales the value of the measure into the interval $[0,1]$.

The distribution of wins that gives M_{ub} or M_{lb} is the same for each of the measures discussed below. When all the races are won by a single driver for example, then the variance

in wins is the highest, and the values of the measures of competition are maximized. This can be seen as follows, where W has the empirical distribution of wins in the league among the D drivers over N races:

$$Var(W) = \frac{\sum_{i=1}^D (W_i - \frac{N}{D})^2}{D} = \frac{\sum_{i=1}^D (W_i^2 - 2 \cdot W_i \cdot \frac{N}{D} + \frac{N^2}{D^2})}{D} = \frac{\sum_{i=1}^D W_i^2 - \frac{N^2}{D}}{D} \quad (2)$$

where W_i is the number of races won by driver i for $i = 1, \dots, D$. $Var(W)$ is maximized when $\sum_{i=1}^D W_i^2$ is maximized, which is when $W_i = N$ for some i . Similarly, when the number of wins are spread out as evenly as possible, the variance is minimized, and the value of the measures of competition will be minimized as well.

A difference between normalizing the measures being looked at when considering top X finishes with $X > 1$ instead of wins ($X = 1$) is that there is often more total top X finishes, $N \cdot X$ in fact, than there are drivers. To determine the lower bound for these measures I will assign wins to each driver as evenly as possible, since unless D divides $N \cdot X$ not every driver will have the same number of wins. Since drivers can only earn whole wins and not fractions, set

$$q = \left\lfloor \frac{X \cdot N}{D} \right\rfloor \quad (3)$$

$$r = \text{mod}(X \cdot N, D) \quad (4)$$

This will then indicate that a league that has $D - r$ drivers with q wins and r drivers with $q + 1$ wins will have the lowest variance and concentration of wins.

One downside of normalization is a potential loss of too much information. If each of the terms in (1) have a factor in common, then that factor is lost after normalization. For many of the measures, D or N is a factor that gets lost, which is part of the goal of normalization since it removes some of the yearly characteristics of the league. If, on the other hand, the terms are a ratio of some calculated values to a fixed value as in section 3.6, the fixed value is lost. I later propose a solution for this case.

3.2 Winning

In all traditional measures of competition, either number of wins or winning percentage is used. This leads to an important distinction between NASCAR and head-to-head sports in which these measures are used: a race in NASCAR has more than one non-winner. In this thesis I calculate these measures not just in the traditional sense of winning, but if a top five or top ten finish is considered winning as well. While seemingly arbitrary, these definitions of “winning” have a basis in the sport since the top five and top ten finishes are recorded in the yearly data and in the NASCAR Hall of Fame data, NASCAR (2019). For instance, in the NASCAR Hall of Fame profile for Alan Kulwicki, the discussion of his 1992 season includes an emphasis on his “11 top fives and 17 top 10s” during the season as being instrumental in his championship victory.

It should be noted that this is not the only, or perhaps not even the most intuitive, way to generalize winning. In the structure that I propose, the top X finishing drivers in each race effectively receives 1 point while every other driver receives no points. There is no reason, other than simplicity, that each finish could receive some non-zero amount of points instead. I want to emphasize, however, that using this method makes a stronger assumption than the one I chose to explore. For these points, a weighting scheme must assign a point value for each position. Without data to test the uncertainty of outcomes hypothesis, any weighting system will be arbitrary with no way to determine whether a particular choice for weights is better or worse than any other.

3.3 Competitive Balance Ratio

The competitive balance ratio (CBR) defined by Humphreys (2002) is a common way to measure competition in a sports league, and it is calculated as

$$CBR = \frac{\bar{\sigma}_T}{\bar{\sigma}_N} \tag{5}$$

where $\bar{\sigma}_T$ is the mean of the standard deviations of wins for the teams/drivers in a league over time and is calculated as the mean over the i years of the following:

$$\sigma_{T,i} = \sqrt{\frac{\sum_{t=1}^T (wpct_{i,t} - \overline{wpct}_i)^2}{T}} \quad (6)$$

where T is the number of years in the league being used and $wpct_{i,t}$ is the percentage of races driver i won in year t . $\bar{\sigma}_N$ is the mean of the standard deviations of winning percentages for the league in each year, and is traditionally calculated as

$$\sigma_{D,i} = \sqrt{\frac{\sum_{i=1}^D (wpct_{i,t} - .500)^2}{D}} \quad (7)$$

where D is the number of competitors in year t .

This measure, as is, is not appropriate to measure competition in individual NASCAR seasons. First of all, it measures competition over time instead of a single season. Secondly, since the same drivers are not driving each year it becomes much more difficult to normalize. Finally, the .500 in the calculation for $\sigma_{D,i}$ would change to be the average winning percentage for drivers in year t , but that will change from year to year with the number of drivers. Thus, while *CBR* is an important and common measure for competition over time, I will not be calculating *CBR* in this thesis.

3.4 Concentration Ratios

A common, and easy to understand, measure of competition is the concentration ratio. The concentration ratio is simply the percentage of races won by the drivers who won the most races. For a league of size D , where the drivers are ranked from 1 to D in descending order of wins, the concentration ratio for the top $k < D$ drivers, CR_K , is

$$CR_k = \frac{\sum_{i=1}^k wins_i}{\sum_{i=1}^D wins_i} \quad (8)$$

3.4.1 Normalization for Concentration Ratios

As previously discussed in section 3.1, it is important to normalize this measure, and it was shown by Manasis et al. (2011) that a normalized concentration ratio for the top k drivers, NCR_k , can be defined as follows:

$$NCR_k = \frac{CR_k - CR_{k_{lb}}}{CR_{k_{ub}} - CR_{k_{lb}}} \quad (9)$$

where $CR_{k_{ub}}$ and $CR_{k_{lb}}$ are the upper and lower bounds for CR_k , respectively. We then have $0 \leq NCR_k \leq 1$.

The upper bound occurs when the top k drivers combine to win every race. The lower bound occurs when each of the top k drivers each has exactly 1 win. Thus,

$$NCR_k = \frac{CR_k - \frac{k}{N}}{1 - \frac{k}{N}} \quad (10)$$

where N is the number of races in a particular season.

3.4.2 Normalization for Top X Finishes for Concentration Ratios

The upper bound for CR_K is as follows:

$$CR_{k_{ub}} = \min\left\{1, \frac{k}{X}\right\} \quad (11)$$

and this upper bound occurs when all of the top X finishes belong to k drivers. As can be seen above in (10) and (11), if $X > k$, then the maximum proportion of top X finishes from the k drivers is $\frac{N \cdot k}{N \cdot X} = \frac{k}{X}$ where N is the number of races in the season.

The lower bound for CR_k is only slightly more difficult to determine. Using the definitions of q and r from (3) and (4), we have:

$$k > r \Rightarrow CR_{k_{lb}} = \frac{1}{N \cdot X} \cdot [(k - r) \cdot q + r \cdot (q + 1)] = \frac{k \cdot q + r}{N \cdot X} \quad (12)$$

$$k \leq r \Rightarrow CR_{k_{lb}} = \frac{k \cdot (q + 1)}{N \cdot X} \quad (13)$$

CR_k does, however, have downsides. First of all, the choice of k is completely arbitrary. Additionally, CR_k and NCR_k only directly take into account the actual performance of k drivers in the league and not the distribution of wins among the remaining drivers. The next measure of competition does a better job of taking the performance of all the drivers in a season.

3.5 Herfindahl-Hirschman Index

The Herfindahl-Hirschman Index (HHI) is a common way to measure concentration of revenue in an industry or wins in a league. HHI is the sum of squared winning percentages for every driver in the league and thus, unlike the concentration ratio, directly takes into account every driver's performance. For a league with D drivers, let $wpct_i$ be the winning percentage (races won divided by total races in a season) for each driver $i = 1, \dots, D$. HHI is then defined as

$$HHI = \sum_{i=1}^D wpct_i^2. \quad (14)$$

Squaring the winning percentages causes HHI to be bounded between 0 and 1, but there are also upper and lower bounds which are based on the highest winning percentage. These bounds are

$$0 < wpct_{max}^2 \leq HHI \leq wpct_{max} \leq 1$$

Since at least one competitor has a non-zero winning percentage we have $0 < HHI$, as above. It is noted by Depken (1999) that this strict inequality can be removed by subtracting the minimum possible value of HHI , which we will call HHI_{lb} . HHI adjusted in this way is called $dHHI$ and is defined as

$$dHHI = HHI - HHI_{lb} = \sum_{i=1}^D wpct_i^2 - HHI_{lb}. \quad (15)$$

3.5.1 Normalization of HHI

For HHI , we have the straightforward upper and lower bounds as

$$HHI_{lb} = \frac{1}{D} \quad (16)$$

$$HHI_{ub} = 1 \quad (17)$$

where the lower bound occurs if each of the N races has a unique winner. The upper bound represents each race being won by the same driver. HHI normalized, denoted as HHI^* , is then

$$HHI^* = \frac{HHI - \frac{1}{D}}{1 - \frac{1}{D}}. \quad (18)$$

3.5.2 Normalization for Top X Finishes for HHI

When we change the definition of winning for a driver to finishing in the top X positions, HHI then no longer measures sum of squared winning percentages. It is instead the sum of squared percentages of each driver's share of top X finishes. That is,

$$HHI = HHI = \sum_{i=1}^D wpct_i^2$$

with $wpct_i = \frac{f_i}{X \cdot N}$ where f_i is the number of top X finishes for driver i . If a driver i had a top X finish in each race then $wpct_i = \frac{1}{X}$. Thus, we have the straightforward upper bound as

$$HHI_{ub} = \sum_{i=1}^X \left(\frac{1}{X} \right)^2 = \frac{1}{X} \quad (19)$$

which occurs when the top X drivers are the same for each race. The lower bound is more involved, but similar to the derivation for CR_k . Using the definitions of q and r from (3)

and (4), the lower bound for HHI is

$$HHI_{lb} = \sum_{i=1}^{D-r} \left(\frac{q}{N \cdot X} \right)^2 + \sum_{i=1}^r \left(\frac{q+1}{N \cdot X} \right)^2. \quad (20)$$

Unfortunately, $dHHI$ then has the opposite issue of HHI : instead of $0 < HHI \leq 1$ we have $0 \leq dHHI < 1$. This is corrected by Owen et al. (2007) by simply dividing $dHHI$ by its range, $HHI_{ub} - HHI_{lb}$, and the new measure is referred to as HHI^* , which is defined as

$$HHI^* = \frac{HHI - HHI_{lb}}{HHI_{ub} - HHI_{lb}} = \frac{\sum_{i=1}^D wpct_i^2 - HHI_{lb}}{HHI_{ub} - HHI_{lb}} \quad (21)$$

where HHI_{lb} depends on the size and structure of the league, and HHI_{ub} is always 1 for $D \geq X$. While HHI (and by extension $dHHI$ and HHI^*) is easy to both calculate and understand, it only considers competitors in the league with a non-zero winning percentage. This means that HHI does not capture the fact that in NASCAR there are a lot of drivers each season that do not win a single race. The next measure captures this information, but since it can be shown to be a function of HHI , it possibly does not provide any extra useful information.

3.6 Ideal Standard Deviation Ratio

A third commonly used measure of competition, defined in Fort and Quirk (1995), for a single season is the ratio of actual standard deviation of win percentages, σ_A , to the standard deviation of win rates in an “ideal” league, σ_I . This ratio, denoted as IR for idealized standard deviation ratio, is calculated as

$$IR = \frac{\sigma_A}{\sigma_I} = \sqrt{\frac{\sum_{i=1}^D (wpct_i - \mu_{wpct})^2}{D}} \cdot \frac{1}{\sigma_I} \quad (22)$$

where $wpct_i$ is the win percentage for driver $i = 1, \dots, D$ and μ_{wpct} is the mean winning percentage in the league.

3.6.1 Normalization for Idealized Standard Deviation Ratio

It is a little more involved to normalize IR than the other ratios. The ratio could equal zero (when the teams are at parity in terms of wins), but that would imply that the league is more competitive than in the “ideal” case. While empirically unlikely to occur, it would be preferable for all values of IR with $\sigma_A < \sigma_I$ to be mapped to 0, so I normalize IR in the following way

$$IR_N = \frac{\max\{IR, 1\} - 1}{IR_{ub} - 1} \quad (23)$$

where IR_N is the normalized value, IR is the calculated value, and IR_{ub} is the upper bound for IR . An additional benefit of normalizing in this way is that σ_I does not get normalized out in the calculation. As discussed in section 3.1, some characteristics of the league we would like normalized out, like N or D . If σ_I is normalized out then the only item actually being measured is the standard deviation of wins in the league.

For any particular league σ_I is a constant, and in other sports it is much easier to derive. The assumption that drives the derivation is that in an “ideal” league each competitor has an equal chance to win each competition. Thus, the number of wins for a particular driver, W , is a binomial random variable with number of trials being the number of races that driver participated in and the probability of success being $\frac{1}{d}$ where d is the number of drivers per race. If we let $WP = \frac{1}{N} \cdot W$ where N is the number of races that the driver participated in, then WP is the distribution of win percentage for a driver in an “ideal” league. The calculation for σ_I is then

$$\sigma_I = \sqrt{Var(WP)} = \sqrt{\frac{1}{N^2} Var(W)} = \sqrt{\frac{1}{N^2} \cdot N \cdot \frac{1}{d} \cdot \frac{d-1}{d}} = \sqrt{\frac{d-1}{N \cdot d^2}} \quad (24)$$

Since σ_I does not depend on the data, to normalize IR we only need to consider the upper and lower bounds for σ_A . As discussed above, while $\sigma_A < \sigma_I$ is possible, I set the minimum for IR to be 1. Clearly variance is maximized if a single driver won every race, so thus we

have:

$$IR_{ub} = \sqrt{\frac{(D-1)\bar{\mu}_{wpct}^2 + (1-\bar{\mu}_{wpct})^2}{D}} \cdot \frac{1}{\sigma_I} \quad (25)$$

The normalization of IR is then

$$IR_N = \frac{\max\{IR, 1\} - 1}{\sqrt{\frac{(D-1)\bar{\mu}_{wpct}^2 + (1-\bar{\mu}_{wpct})^2}{D}} \cdot \frac{1}{\sigma_I} - 1} \quad (26)$$

The upper bound is obtained when the variance of win percentages is maximized, which can be seen from (2) replacing W , the distribution of wins, with WP , the distribution of win percentages.

3.6.2 Normalization for Top X Finishes for Idealized Standard Deviation Ratio

Again, the lower bound is being set to 1 for the purposes of this thesis. The upper bound, which occurs when the top X drivers are the same for each race, is calculated as follows:

$$IR_{N_{ub}} = \sqrt{\frac{\sum_{i=1}^X (1 - \overline{wpct})^2 + \sum_{i=1}^{D-X} (0 - \overline{wpct})^2}{D}} = \sqrt{\frac{X \cdot (1 - \overline{wpct})^2 + (D - X) \cdot \overline{wpct}^2}{D}} \quad (27)$$

It has been shown by Depken (1999) that the standard deviation of winning percentage within a head-to-head sports league is a function of the positive square root of HHI , which also holds for NASCAR. This should be obvious; HHI is the sum of squared winning percentages and IR involves the standard deviation of winning percentages. Thus, while IR is influenced by a large number of competitors with win percentages of zero (an issue with HHI and CR_K for NASCAR as seen above), it is a more difficult measure to interpret while being a function of an easier to interpret measure.

3.7 Churn

The final measure, churn, is calculated as

$$CHURN_t = \frac{\sum_{i=1}^d |f_{i,t} - s_{i,t}|}{d} \quad (28)$$

where d is the number of drivers in race i , $f_{i,t}$ is the finishing position of driver i in race t , and $s_{i,t}$ is the starting position of driver i in race t . As with the other measures, a normalized, or “adjusted”, churn has been developed by Mizak et al. (2007). To be consistent with the other measures, take the difference between one and $CHURN_t$ so that zero represents a perfectly competitive race and one represents a race that is not competitive at all. $ADJCHURN_t$ is then

$$ADJCHURN_t = 1 - \frac{CHURN_t}{MAXCHURN_t} = 1 - \frac{\sum_{i=1}^d |f_{i,t} - s_{i,t}|}{n} \cdot \frac{1}{MAXCHURN_t} \quad (29)$$

where $MAXCHURN_t$ is the upper bound for churn in race t . It was shown by Mizak et al. (2007) that for a race with an even number of drivers

$$MAXCHURN_t = \frac{d}{2} \quad (30)$$

and for a race with an odd number of drivers

$$MAXCHURN_t = \frac{d^2 - 1}{2 \cdot d}. \quad (31)$$

Berkowitz et al. (2011) stated that the previous measures discussed are not appropriate when viewing a single NASCAR race. However in this analysis competitiveness of the season as a whole is of concern; thus, the mean for each season is represented by $ADJCHURN$ and calculated as

$$ADJCHURN = \frac{\sum_{i=1}^N ADJCHURN_t}{N}. \quad (32)$$

3.8 Example

For the non-churn measures of competition discussed above, it is possibly not immediately clear how different the values of the measures could be, before and after normalization. As an illustration, consider a league where the number of races, N , the number of drivers, D , and the the number of drivers per race, d , are all equal to five. As discussed in 3.4, the value of k in CR_k is arbitrary, so consider CR_2 . This choice is reasonable because 3 would be over half the total number of drivers and 1 is just the concentration of wins for the single best driver. Table 1 illustrates a possible way the wins could be distributed in the 5 races. Here

| Driver | R ₁ | R ₂ | R ₃ | R ₄ | R ₅ |
|----------------|----------------|----------------|----------------|----------------|----------------|
| d ₁ | 1 | 2 | 2 | 2 | 2 |
| d ₂ | 2 | 1 | 3 | 3 | 3 |
| d ₃ | 3 | 3 | 1 | 4 | 4 |
| d ₄ | 4 | 4 | 4 | 1 | 5 |
| d ₅ | 5 | 5 | 5 | 5 | 1 |

Table 1: Example Race Results

each driver won one race each. The non-churn measures for this hypothetical league are

$$HHI = \frac{1}{25} + \frac{1}{25} + \frac{1}{25} + \frac{1}{25} + \frac{1}{25} = \frac{1}{5} (0), \quad CR_2 = \frac{2}{5} (0), \quad IR = 0 (0)$$

where the values in red are the normalized values for each measure. Since the numerator of IR is the standard deviation of wins for the drivers in the league, which is 0 as there is no difference in wins among the drivers. In the example it should be noted that each of the non-churn measures is normalized to 0 since these measures achieve their theoretical minimum value, which corresponds to the lowest level of concentration. An issue is apparent if anything other than first place finishes are considered, as d_1 clearly did much better during these five races than d_5 did. With this in mind, Table 2 shows the results when, for example,

finishing in either first or second is considered winning, and below the table are the calculated values of the non-churn measures with their normalized values in red.

| Driver | R ₁ | R ₂ | R ₃ | R ₄ | R ₅ |
|----------------|----------------|----------------|----------------|----------------|----------------|
| d ₁ | 1 | 2 | 2 | 2 | 2 |
| d ₂ | 2 | 1 | 3 | 3 | 3 |
| d ₃ | 3 | 3 | 1 | 4 | 4 |
| d ₄ | 4 | 4 | 4 | 1 | 5 |
| d ₅ | 5 | 5 | 5 | 5 | 1 |

Table 2: Example Race Results Redefining Winning

$$HHI = \frac{25}{100} + \frac{4}{100} + \frac{1}{100} + \frac{1}{100} + \frac{1}{100} = \frac{32}{100} (.4), \quad CR_2 = \frac{7}{10} (.5), \quad IR = 1.73 (.374)$$

Not only is it clear that this league is not perfectly competitive when viewing finishing positions other than first favorably, but it is a good example of the benefits of normalization. Here, IR is calculated to be 1.73, and without some sort of context this value is meaningless. The normalized value of .374 suggests that the results are neither perfectly competitive nor perfectly monopolistic and is closer to the perfectly competitive lower bound.

4 Data and Calculations

4.1 Data

The data employed to calculate the above-discussed measures was obtained from NASCAR (2017), which is freely available and is in a format that allows a straightforward computation of the above metrics. This study uses the yearly summary by driver to compute HHI^* , NCR_K , and IR_N . The final measure, $ADJCHURN$, is calculated using race-level data from 1972 to 2015. Each year drivers that participated in less than half of the races are removed because these drivers were not competing for the championship and would have unduly skewed the measures of competition that are calculated below.

4.2 Calculated Metrics

4.2.1 Adjusted Churn

The summary statistics and correlation matrix for the calculated metrics can be seen in Tables 3 and 4. Compared to the other measures, *ADJCHURN* is fairly stable through the time period, but when considering top 10 finishes to be wins it is negatively correlated with the metrics other than *HHI** and *IRN*. The relative stability makes it less likely that *ADJCHURN* is capturing any of the changes in the overall competition in the league. On the other hand, competition in the league could have been stable during the time period, which contradicts the variability in the other measures, so is unlikely. This stability can be seen graphically in Figures 1, 2, and 3. With this in mind, most of the analysis will be focused on the other metrics.

4.2.2 Wins

In Figure 1, the definition of winning used is a first place finish. Both *NCR₅* and *NCR₁₀* show a close to perfect lack of competition in the distribution of wins, with a noticeable downward trend over time, i.e. they suggest NASCAR has become more competitive over time. Additionally, they follow or match each other very closely, which can also be seen in Table 3. In general, both of these measures are above *ADJCHURN*. The remaining two measures, *HHI** and *IRN*, both remain below *ADJCHURN* for the entire time period. Like the concentration ratios above, these two measures are very highly correlated with each other, which is to be expected from the discussion in section 3.6 about the relationship between these measures. It is also clear to see that all four of these measures reach high and low points in the same years, so all of the measures are capturing the same or similar information about competitiveness.

Table 3: Correlation Matrix for Selected Metrics

| | hhi | hhi_5 | hhi_10 | ncr5 | ncr5_5 | ncr5_10 | ncr10 | ncr10_5 | ncr10_10 | irn | irn_5 | irn_10 | adjc |
|----------|--------|--------|--------|--------|--------|---------|---------|---------|----------|--------|--------|---------|--------|
| hhi | 1 | 0.427 | -0.377 | 0.946 | 0.582 | 0.525 | 0.776 | 0.513 | 0.215 | 0.982 | 0.605 | -0.294 | -0.275 |
| hhi_5 | 0.427 | 1 | 0.436 | 0.529 | 0.940 | 0.939 | 0.588 | 0.879 | 0.829 | 0.468 | 0.927 | 0.496 | -0.154 |
| hhi_10 | -0.377 | 0.436 | 1 | -0.268 | 0.229 | 0.267 | -0.067 | 0.301 | 0.492 | -0.309 | 0.256 | 0.973 | 0.267 |
| ncr5 | 0.946 | 0.529 | -0.268 | 1 | 0.672 | 0.622 | 0.869 | 0.624 | 0.346 | 0.934 | 0.704 | -0.163 | -0.189 |
| ncr5_5 | 0.582 | 0.940 | 0.229 | 0.672 | 1 | 0.986 | 0.688 | 0.902 | 0.760 | 0.588 | 0.972 | 0.322 | -0.183 |
| ncr5_10 | 0.525 | 0.939 | 0.229 | 0.622 | 0.986 | 1 | 0.625 | 0.873 | 0.829 | 0.545 | 0.957 | 0.368 | -0.168 |
| ncr10 | 0.776 | 0.588 | -0.067 | 0.869 | 0.688 | 0.625 | 1 | 0.692 | 0.402 | 0.777 | 0.719 | -0.0002 | -0.127 |
| ncr10_5 | 0.513 | 0.879 | 0.301 | 0.624 | 0.902 | 0.873 | 0.692 | 1 | 0.789 | 0.489 | 0.945 | 0.398 | -0.084 |
| ncr10_10 | 0.215 | 0.829 | 0.492 | 0.346 | 0.760 | 0.829 | 0.402 | 0.789 | 1 | 0.270 | 0.775 | 0.577 | -0.086 |
| irn | 0.982 | 0.468 | -0.309 | 0.934 | 0.588 | 0.545 | 0.777 | 0.489 | 0.270 | 1 | 0.596 | -0.241 | -0.327 |
| irn_5 | 0.605 | 0.927 | 0.256 | 0.704 | 0.972 | 0.957 | 0.719 | 0.945 | 0.775 | 0.596 | 1 | 0.371 | -0.094 |
| irn_10 | -0.294 | 0.496 | 0.973 | -0.163 | 0.322 | 0.368 | -0.0002 | 0.398 | 0.577 | -0.241 | 0.371 | 1 | 0.290 |
| adjc | -0.275 | -0.154 | 0.267 | -0.189 | -0.183 | -0.168 | -0.127 | -0.084 | -0.086 | -0.327 | -0.094 | 0.290 | 1 |

Table 4: Summary Statistics for Calculated Metrics

| Statistic | N | Mean | St. Dev. |
|-----------|----|-------|----------|
| hhi | 44 | 0.122 | 0.054 |
| hhi_5 | 44 | 0.257 | 0.061 |
| hhi_10 | 44 | 0.245 | 0.043 |
| ncr5 | 44 | 0.674 | 0.141 |
| ncr5_5 | 44 | 0.384 | 0.083 |
| ncr5_10 | 44 | 0.274 | 0.094 |
| ncr10 | 44 | 0.897 | 0.102 |
| ncr10_5 | 44 | 0.633 | 0.099 |
| ncr10_10 | 44 | 0.126 | 0.045 |
| irn | 44 | 0.207 | 0.079 |
| irn_5 | 44 | 0.380 | 0.080 |
| irn_10 | 44 | 0.386 | 0.051 |
| adjc | 44 | 0.499 | 0.038 |

4.2.3 Top 5 Finishes

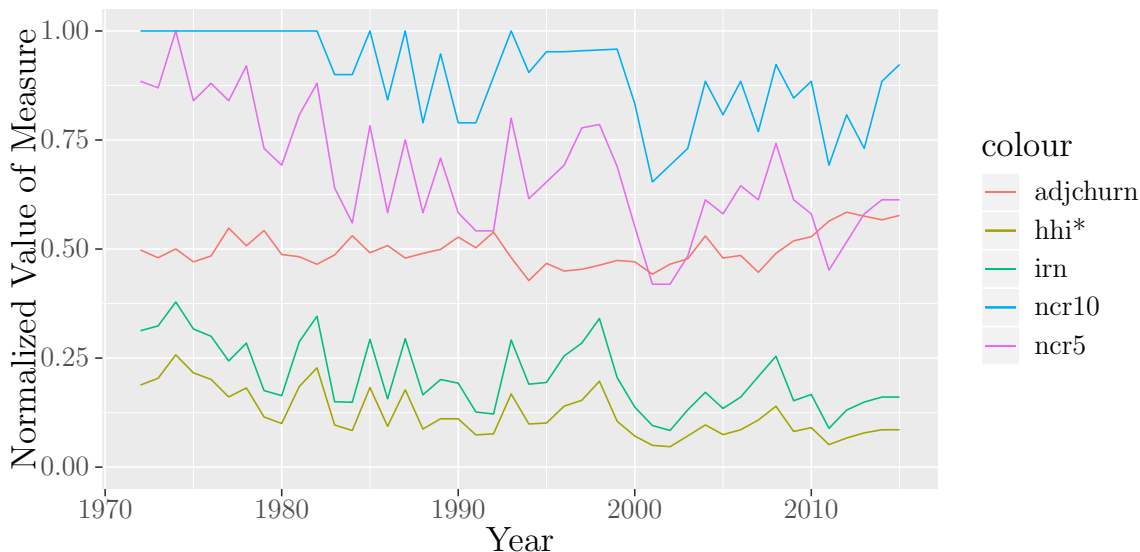
In Figure 2, the definition of winning is a top 5 finish. Here, as in the case where winning was just a first place finish, NCR_5 and NCR_{10} are highly correlated. For most of the time period, however, NCR_{10} is above $ADJCHURN$ while NCR_5 is below it. For both of these measures, the mean decreased going from counting first place finishes to counting top 5 finishes. Additionally, the standard deviation for NCR_5 for top five finishes is almost half what it is for NCR_5 for wins while the standard deviation for NCR_{10} is nearly the same for both wins and top five finishes.

With HHI^* and IRN computed for top 5 finishes, the means for both almost doubled in the opposite direction of the concentration ratios. On the other hand, the standard deviations made almost no change at all. Interestingly, while HHI^* and IRN are still very highly correlated, the correlation is higher between IRN and the concentration ratios.

4.2.4 Top 10 Finishes

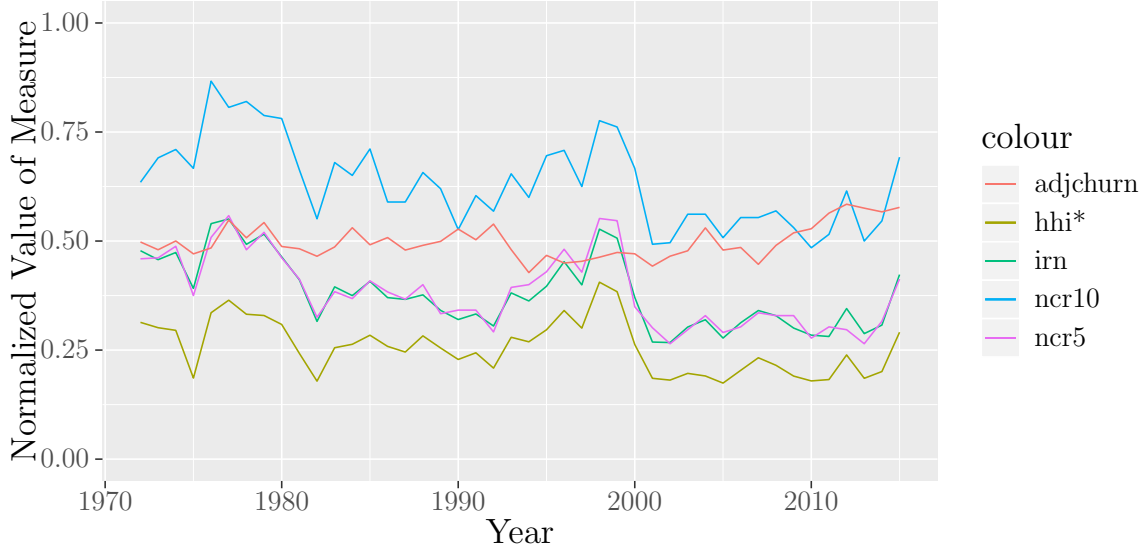
When going from defining winning by top five finishes to top ten finishes, major differences start to appear in Figure 3. For NCR_5 and NCR_{10} the correlation is noticeably lower than

Figure 1: Measures of Competition Using Wins, 1972-2015



it was for either top 5 finishes or first places finishes. The mean of NCR_{10} is approximately 20% of the value for top 5 finishes, while the mean of NCR_5 is about 70% of the value for top 5 finishes. The standard deviation of NCR_5 increased by almost 70% in comparison to the values for top 5 finishes, but the standard deviation of NCR_{10} is less than half of what it was for top 5 finishes. Additionally, the concentration ratios are now the lowest measures over the time period, with NCR_{10} being lower than NCR_5 every year. This result is due to the interaction between the choice of both k and W . Since $W = 10$, for NCR_{10} the denominator is $1 - CR_{10_{lb}}$ while the denominator for NCR_5 is $.5 - CR_{5_{lb}}$, so NCR_5 is inflated in comparison since not every top 10 finish can be earned by the same group of 5 drivers. This is a good example of one of the key features of normalization: it gives context as to the meaning of the value of the measure in terms of the theoretical maximum and minimum value of the measure. The other two measures, HHI^* and IRN , have almost no change in their means between top 5 finishes and top 10 finishes. The standard deviations, however, are about 2/3 as high when calculated with top 10 finishes as with top 5 finishes. These measures are still highly correlated with each other, but are no longer nearly as correlated with the concentration ratios as they are when calculated with first place finishes or top 5

Figure 2: Measures of Competition Using Top 5 Finishes, 1972-2015



finishes. Additionally, this is the only instance where any of the four measures that change based on the definition of winning are positively correlated with *ADJCHURN*.

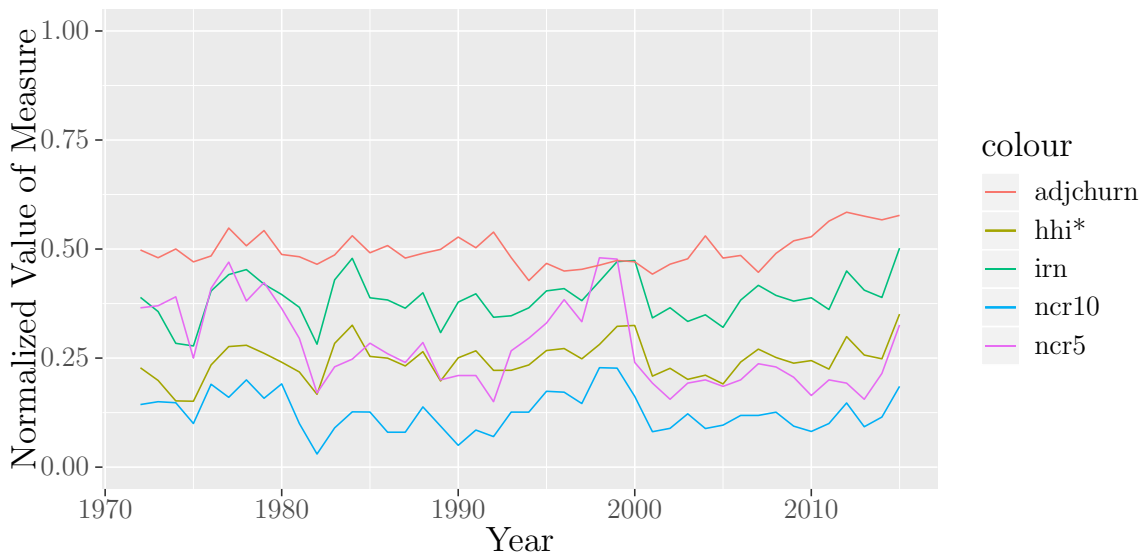
4.2.5 NCR_5

When looking at just NCR_5 , in Figure 4, for different definitions of winning, the trend over time is clearly from less competition to more competition. Interestingly, NCR_5 for top five finishes and NCR_5 for top ten finishes are almost perfectly correlated, much more so than the correlation between NCR_5 and NCR_{10} for top five finishes or top ten finishes. While NCR_5 has a lower mean when winning is defined by top 10 finishes rather than top 5 finishes, the standard deviation is almost 15% higher. In both cases, however, the standard deviation is significantly lower than when winning is defined to be only first place.

4.2.6 NCR_{10}

The results for NCR_{10} are similar to those for NCR_5 , but more pronounced in some ways while being less pronounced in others. The mean for NCR_{10} with different definitions of winning decreases as more finishing positions are considered “winning,” but the drop from

Figure 3: Measures of Competition Using Top 10 Finishes, 1972-2015

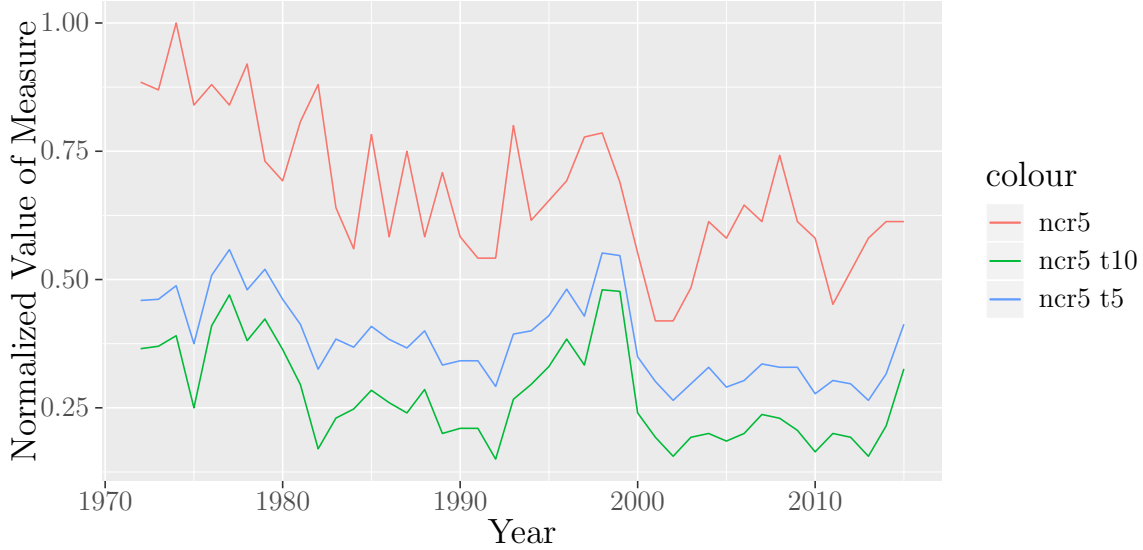


wins to top 5's is not nearly as pronounced as it was with NCR_5 . On the other hand, going from top 5 finishes to top 10 finishes, NCR_{10} drops to less than a fifth of the value with a significant decrease in standard deviation as well. In fact, there is no discernable overall trend for NCR_{10} with winning being a top 10 finish, while there is a noticeable overall slight downward trend when considering just first place finishes or top 5's. In any case, the concentration ratios show a movement closer to the theoretical perfect competition when the definition of winning is changed to include more of the field.

4.2.7 HHI^*

The story for HHI^* is much less clear. To start, the opposite result regarding competition is true compared to the concentration ratios: as the definition of winning is changed to including more finishing positions, HHI^* moves further from the theoretical perfect competition. Also, the means for top 5 and top 10 finishes are very close, but the standard deviation for HHI^* with top 5 finishes is almost 50% higher than it is for top 10 finishes. It can also be seen that there is a noticeable trend towards more competitive for HHI^* with just first place finishes, but when considering top 5 and top 10 finishes there is no clear trend at all. Looking at the

Figure 4: NCR5 for Wins, T5, and T10, 1972-2015



three definitions together, there is the same positive correlation between HHI^* for wins and top 5 finishes as there is between top 5 finishes and top 10 finishes, however the relationship between HHI^* for wins and top 10 finishes is slightly weaker but in the negative direction instead.

4.2.8 IRN

IRN is almost the same case as HHI^* when looking at the three definitions of winning, which should come as no surprise. The mean almost doubles going from first place finishes to top 5 finishes, whereas in HHI^* it was slightly more than double. The standard deviation stayed the same for that change, where for HHI^* the standard deviation increased by over 10%. When going from calculating IRN from top 5 finishes to top 10 finishes, the mean stayed nearly the same and the standard deviation dropped to approximately 2/3 of what it was. The same result was true for HHI^* . Finally, the correlation between IRN for first place finishes and top 5 finishes was stronger than it was for HHI^* , while the other two correlations were weaker. The correlations were in the same direction as they were for HHI^* .

Figure 5: NCR10 for Wins, T5, and T10, 1972-2015

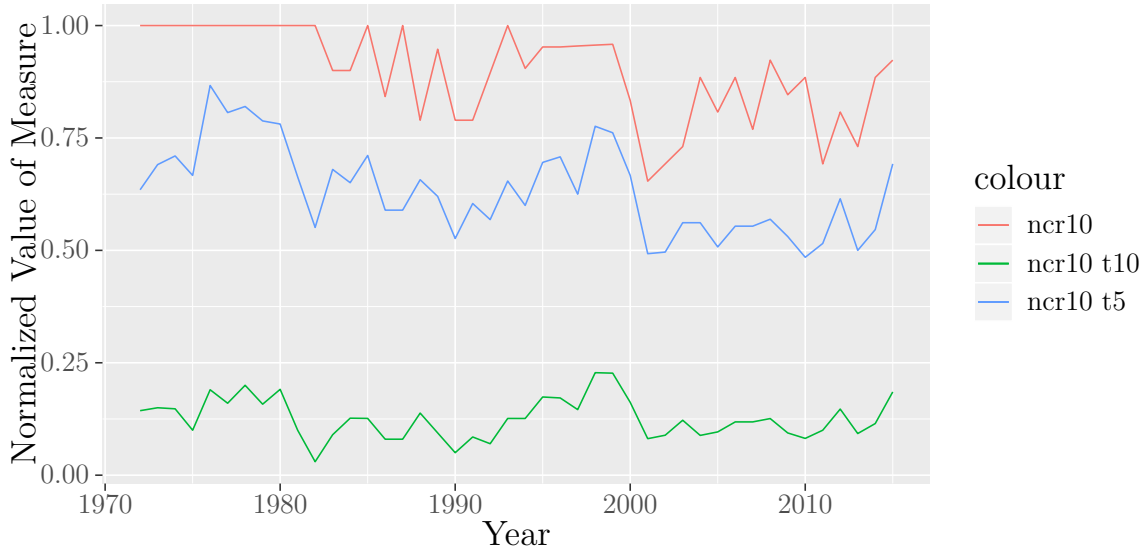
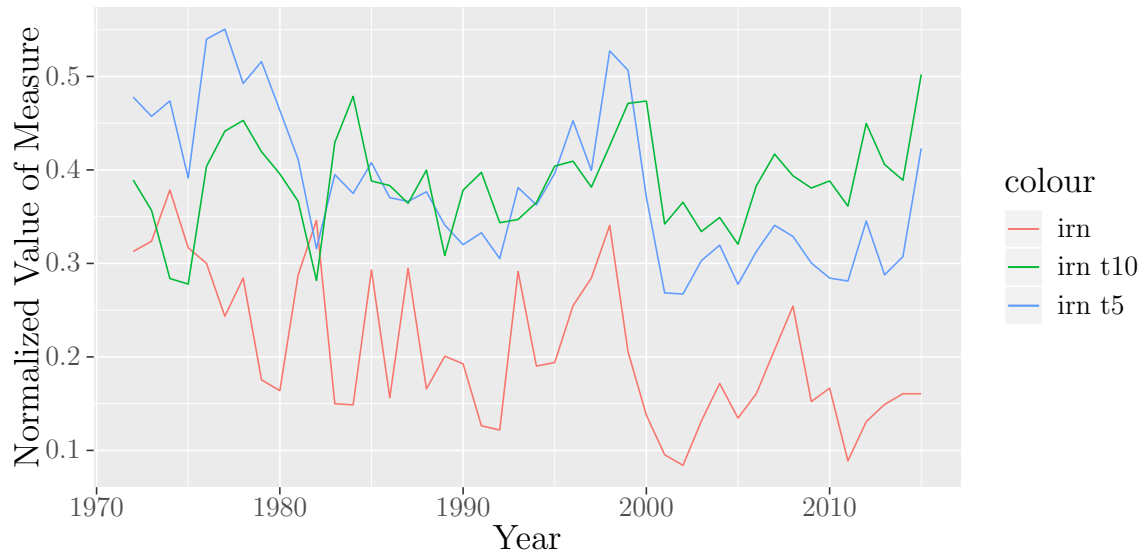


Figure 6: HHI for Wins, T5, and T10, 1972-2015



Figure 7: IRN for Wins, T5, and T10, 1972-2015



5 Conclusions

When looking at NASCAR at the season level, it is clear that the measures of competition explored in this thesis not only do not agree with each other when looking at a generalized definition of winning, but they are not consistent with themselves when the definition of winning is changed. Some of the more extreme examples of the differences can be explained by looking at the raw data. For example, looking at NASCAR (2017) in 1982 a single driver, Darrell Waltrip, won 12 of the 30 races, and combined with Bobby Allison won 20 of the 30 races. There is a clear spike counter to the trend in this year when looking at the measures with first place being the definition for winning, but when the definition is changed to include top 5 or top 10 finishes there is almost as large of a negative spike from the trend in that same year. When looking at the driver data for that year, it is seen that drivers all the way down to 26th in terms of points (as determined by NASCAR) have double digit top ten finishes.

Additionally, the value of normalization is clear, especially graphically. In the case of first place finishes, IRN is only slightly higher (less competitive) than HHI^* in general, but without normalization that relationship is not as clear. Similarly, the concentration ratios for the top 5 and top 10 drivers are in the opposite order in terms of magnitude than would be expected when considering top 10 finishes, and this relationship would not be noticed without normalization.

6 Further Research

As discussed under generalizing winning, rather than simply choosing a set of finishes to have weight of 1 and the rest 0, a more nuanced system of weighting finishes could be used. For instance, a first place finish could be considered 1 win, a top 5 finish could be considered $1/3$ of a win, a top 10 finish could be considered $1/8$ of a win, and so on. Alternatively, as used by Berkowitz et al. (2011), NASCAR's points system could be used for the weights.

The amount of prize money could be as well. In any case, these systems to determine weights can be analyzed in a similar manner as measures of competition were in this analysis.

The uncertainty of outcomes hypothesis also needs to be tested at the season level. If NASCAR's yearly revenue were available that would be the best way to measure overall fan interest, but there are other options. Just as was employed by Berkowitz et al. (2011), both in person and broadcast viewing numbers could be used. For years after 2000, internet search data could be used. In any case, the measure of fan interest that is used to test the uncertainty of outcomes hypothesis will also shed light on the nature of the competition that is preferred by fans. NASCAR can then change their set of rules to try to move the level of competition to a level preferred by fans rather than to try to make the league more competitive based on a measure that may not match fan preferences and thus limit profitability of the sport.

With the focus on optimal competition as viewed by fans, the discussion of how the measures and definitions from this thesis can be used shifts also to the ability of NASCAR to change the set of rules in a manner to optimize a chosen measure of competitive balance. For instance, it is possible that a given rule change, such as the adoption of the "Car of Tomorrow" or the conversion to the current playoff structure, could advantage select teams or change driver incentives within a season and thus qualitatively impact the various measures in predictable ways. As such, one of the next steps is to empirically measure the quantitative impact of such rules changes on these measures.

7 References

References

- Berkowitz, J. P., Depken, C. A., and Wilson, D. P. (2011). When going in circles is going backward: Outcome uncertainty in NASCAR. *Journal of Sports Economics*, 12(3):253–283.
- Depken, C. A. (1999). Free-agency and the competitiveness of Major League Baseball. *Review of Industrial Organization*, 14(3):205–217.
- Fort, R. and Quirk, J. (1995). Cross-subsidization, incentives, and outcomes in professional team sports leagues. *Journal of Economic literature*, 33(3):1265–1299.
- Humphreys, B. R. (2002). Alternative measures of competitive balance in sports leagues. *Journal of Sports Economics*, 3(2):133–148.
- Knowles, G., Sherony, K., and Hauptert, M. (1992). The demand for major league baseball: A test of the uncertainty of outcome hypothesis. *The American Economist*, 36(2):72–80.
- Manasis, V., Avgerinou, V., Ntzoufras, I., Reade, J. J., et al. (2011). Measurement of competitive balance in professional team sports using the Normalized Concentration Ratio. *Economics Bulletin*, 31(3):2529–2540.
- Mizak, D., Neral, J., and Stair, A. (2007). The adjusted churn: an index of competitive balance for sports leagues based on changes in team standings over time. *Economics Bulletin*, 26(3):1–7.
- NASCAR (2017). Racing-reference.info.
- NASCAR (2019). Nascar hall of fame.

Owen, P. D., Ryan, M., and Weatherston, C. R. (2007). Measuring competitive balance in professional team sports using the Herfindahl-Hirschman index. *Review of Industrial Organization*, 31(4):289–302.