Streaky Good Models

Incorporating Short-Run Trends into NBA Predictions

Harry Jain

CPSC 490: Senior Project [1]
May 2023
Advisors James Glenn && Kevin O’Neill

Abstract

One of the most prevalent discussion points in sports is the idea of a ”hot hand” or ”streakiness” that reflects a short-run change in a player or team’s performance. However, there is no consensus on whether this phenomenon exists, and if it does, how it can be measured statistically. Moreover, there are very few attempts to directly apply this phenomenon to game predictions. In this vein, Streaky Good Models is a project that aims to quantify the effect of short-run trends on NBA team performance and apply it to create more streak-sensitive models for predicting game results.

In order to rigorously quantify the effect of short-run trends, this project is comprised of several key components, each of which includes mathematical reasoning for the chosen approach paired with statistical analysis of real NBA data. First, the existence of the hot hand is validated by proving a bias in previous research that deemed it to be a fallacy and demonstrating appreciable hot hand effects in NBA data once this bias is corrected. Next, more sophisticated methods of quantifying the hot hand are explored, focused on the development of new metrics that capture short-run trends in a player’s performance; these metrics are then verified using mock ”streaky” data and applied to real NBA data via significance tests using Monte Carlo simulations. Finally, these metrics are tested as possibilities for improving Elo model predictions for NBA games. The results of this project are summarized in the documents below, along with the interactive webpage which includes several options for exploring these effects in real-time NBA data.

Contents

1 Background and Motivation

"The streak has become my identity; it's who I've become."

– Cal Ripken, Jr. [2]

The above quote from famed baseball "Iron Man" Cal Ripken, Jr. serves as an apt introduction to the core topic of this project: streaks in sports. From Ripken's legendary 2,632 consecutive games played to Caltech basketball's 207 consecutive losses to Jahangir Kahn 555 straight squash match victories, nothing captures the imagination of fans and athletes alike more than the idea of a streak [3].

While often viewed the perspective of superstitions and luck, streaks can also reveal changes in form that may have useful predictive value. As Ripken explains so eloquently, streaks are not just an aspect of sports but an encapsulation of their identity, and thus they serve as an obvious avenue for analytical analysis.

1.1 NBA Game Data

While there are certainly streaks of various kinds in each and every sport, one sport whose streaks are particularly explanatory of a team's form is basketball, in particular the NBA. As explained by FiveThirtyEight, "NBA data is subject to relatively little randomness," as there are less than 10 players in a typical rotation, all of whom play a fairly regular distribution of minutes when uninjured [5]. Moreover, there are no major differences in positions from game to game, unlike baseball with starting pitchers, and the large amount of games (82 in the regular season, as much as 28 in the postseason) provides ample time for variable short-run trends to emerge across a season.

As a result of these characteristics, an extended streak of good or bad performance in the NBA is more likely to represent a true change in the performance of a team, rather than some fluky trend due to randomness. This is often due to changes like a new scheme or lineup, poor or good health of important players, or even just an establishment of better chemistry within a team (as was the case with the 2021-22 Boston Celtics, who started the season 18-21 and finished 51-31). Thus, recent performance has strong predictive power with regards to NBA games, something acknowledged implicitly by fans, journalists, and coaches alike when assessing the probability of success for a given team.

1.2 Typical Predictive Models

In terms of mathematical models predicting NBA games, there has been consistent and continual development, with several methods emerging to prominence in recent years. Most of them build on a staple of more casual game predictions: power ratings.

Whereas most traditional power ratings merely rank teams from 1 to 30, modern models are typically structured to allow precise predictions for a theoretical game between any two teams. In order to do so, there are two main strategies: a mathematical metric that can be used to calculate win probabilities, or a direct measure of how many points a team would be expected to score "relative to average."

1.2.1 Elo Ratings

The first type of model is typified by the now ubiquitous Elo rating system, a method for calculating the relative skill levels and win probabilities for games with two opponents (most often thought of with regards to chess). These types of ratings assume that each participant's performance in a random variable that is normally distributed, which represents the uncertainty natural in any competitive game while also providing a predictor of the outcome of the match based on the difference in Elo scores. Then, when an actual game is played, the scores the teams or participants update, with larger adjustments in the case of an "upset" [4].

The aforementioned FiveThirtyEight provides the most popular rendition of Elo-based ratings for NBA teams, expanding on the typical Elo model with NBA-specific adjustments. In particular, given teams A and B with ratings R_A and R_B (which average about 1500), the probability that team A wins is

$$E_A = \frac{1}{10^{\frac{-(R_A - R_B + Adj)}{400}} + 1} \qquad [6]$$

where Adj represents adjustments from aspects of the game like home-court advantage. Then, after a game in which team A has win probability E_A and has actual result S_A (1 for a win and 0 or a loss), their score is updated to R'_A according to

$$R'_A = R_A + K(S_A - E_A) \qquad [4]$$

where the K-factor determines how responsive the ratings are to each result (20 for FiveThirtyEight's model). Overall, this strategy has proven to be highly effective with regards to win probabilities, essentially providing a dynamic rating for each team that consistently updates with every result.

1.2.2 Adjusted Point Differentials

On the other hand, there are a variety of popular models that instead provide something more in line with a betting spread, i.e. by how many points would a team be expected to win on average. One prominent ranking of this sort is ESPN's NBA Basketball Power Index (BPI) [7], with Basketball Reference's Simple Rating System utilizing a similar technique [8].

While the precise methods are far more opaque than the well-known Elo ratings, they usually involve some Bayesian framework with priors according to statistics like offensive and defensive efficiency, schedule strength, days of rest, and preseason expectations [9]. They also update in real time with new game results, though these equations are also typically proprietary.

1.3 Limitations of Current Models

While the various predictive models described in the previous section naturally update according to new results, they are designed under the assumption that the expected performances of a team changes *slowly* over time [4]. However, this assumption naturally breaks down in the wake of a significant winning or losing streak.

As mentioned in the background on streaks, the somewhat deterministic behavior of NBA games means that significant streaks often reflect equally significant changes in performance. Thus, a model that properly incorporates streaks should be very responsive to these changes.

While Elo ratings do update every game and theoretically compound during a streak, they only adjust a set amount based on the result of a single game and the K-factor. They will thus take a while to properly adjust to a team's short-run performance, potentially only truly reacting once a streak is over. Likewise, the predicted point differential models typically depend strongly on prior assumptions about a team from previous years, roster construction, etc. So while these assumptions can be updated based on team changes like injuries, they are also slow to react to recent form.

From the "sticky" title favorite whose odds never seem to fall enough with their poor performances to the hot young team that just can't move up the BPI rankings, the explosion of mathematical models of the NBA has been coupled with an understatement of the importance of streaking teams. This is the core discrepancy this project will seek to address.

2 Project Outline and Goals

Given this disconnect between the predictive potential of streaks and the basic assumption of consistent performance in the most popular models, there is significant room for development of more streak-sensitive models. This project aims to fill this void by creating such models for NBA game predictions. In order to do so, it follows the basic outline below:

- **Goal:** Increase the sensitivity of NBA models so that they better incorporate short-run performance, aka "streakiness"
- **Justification:** Prove that the "Hot Hand" does exist and significantly impacts "success" probability for NBA teams
  - Disproves research concluding that the hot hand does not exist by demonstrating a selection bias
  - Determine more suitable measures for streakiness
  - Perform significance tests using these measures to demonstrate the streakiness of NBA teams
- **Method:** Utilize the aforementioned streakiness measures to update the Elo ratings of NBA teams more dynamically

Ultimately, this process allows us to rigorously construct new models that more strongly incorporate the current form of teams while still maintaining reasonable overall accuracy.

3 Measuring Streaks

In order to justify the creation of streak-dependent models, we will first validate the existence of "hot hand" effects that would necessitate their existence. To do so, we will disprove canonical results that concluded the hot hand to be a fallacy, consider desired characteristics for a measure of streakiness, and apply these measures to NBA data via Monte Carlo simulations.

3.1 The Hot Hand

The "Hot Hand" refers to the idea that a player is more likely to make a shot if they have made their previous shots. This is a common belief among basketball fans, yet beginning in the 1980s the dominant mathematical view was that the hot hand was a fallacy. However, this conclusion failed to properly account for a significant sampling bias, which we outline below.

3.1.1 The Hot Hand Fallacy

The idea of the "Hot Hand Fallacy" was popularized by Gilovich, Vallone, and Tversky (GVT) in their 1985 paper "The Hot Hand in Basketball: On the Misperception of Random Sequences." We can summarize their study as follows:

- **Goal:** Determine whether empirical and experimental data on shooting streaks differ from binomial expectations.
- **Method:** Calculate the percentage of makes after streaks of three makes and misses, and complete a paired t-test
- **Results:** 49% shooting percentage on "Hot Streaks" and 45% on "Cold Streaks"
- **Conclusion:** Using a paired t-test, this difference is not significant and the hot hand is "a powerful and widely shared cognitive illusion"

3.1.2 Binomial Representation of NBA Games

In order to verify or disprove GVT's conclusion, we need to model the game results (or equivalently shot results) as random variables. The most natural way to do this (and the way that GVT did it) is to model a set of n games (or shots) as a sequence of independent Bernoulli trials for n random variables, each with equal probability p, i.e.

$$\mathbf{W} = \{W_1, \dots W_n\} = \{W_i\}_{i=1}^{n} \text{ s.t. } P(W_i = 1) = p$$

or equivalently a Binomial distribution $B(n, p)$.

Crucially, this model assumes that each trial (shot/game) has the same probability of success, and that the trials are independent. If the hot hand exists, then these assumptions are violated. In particular, there are two possibilities for how the hot hand could affect the game results:

1. **Autocorrelation:** the current trial depends on the outcome of the previous trials
   - i.e. a hot streak leads to increased or decreased future performance
2. **Non-stationarity:** the probability of success changes over time
   - i.e. a team (or player) improves or regresses based on trades, injuries, etc.

The GVT technique for measuring streakiness is really meant to identify autocorrelation, which would intuitively result in higher success probability after a streak of successes; in reality, however, their sampling method overlooked a substantial bias which we will detail in the following section.

3.1.3 Selection Bias

While GVT was a generally accepted and unchallenged result for many years, later research by Miller and Sanjurjo (2015) found that GVT's method of measuring the hot hand was biased. In particular, they concluded that due to a combination of a form of sampling without replacement and the overlapping nature of the selection proceedure (e.g. the two possibilities for sampling from a win streak of length 5), the expected sample shooting percentage after a streak of 3 makes is less than the player's overall shooting percentage. Therefore, the supposed 4% increase above the expected probability is more accurately a 13% increase above the expected probability, a significant figure. Mathematically, this result is outlined in the theorem below.

Theorem: Streak Selection Bias

Let $\mathbf{W} = \{W_i\}_{i=1}^n, n \geq 3$ be a sequence of independent Bernoulli trials, each with equal probability of success $0 < p < 1$.

Then, say $S_k(\mathbf{W})$ represents the subset of trials that immediately follow k consecutive successes, i.e. a "run" of length k, i.e.

$$S_k(\mathbf{W}) \coloneqq \{i : \prod_{j=i-k}^{i-1} X_j = 1\} \subseteq \{k+1, \ldots, n\}$$

Furthermore, define $\hat{P}_k(\mathbf{B})$ to be the proportion of successes of trials in $S_k(\mathbf{W})$, i.e.

$$\hat{P}_k(\mathbf{W}) \coloneqq \frac{\sum_{i \in S_k(\mathbf{W})} W_i}{|S_k(\mathbf{W})|}$$

In this setting, $\hat{P}_k(\mathbf{W})$ is a biased estimator of the probability that an independent trial $W_t = 1$ given that the preceding k $(1 \leq k \leq n-2)$ trials were 1, i.e.

$$P(X_t = 1 \mid \prod_{j=t-k}^{t-1} W_j = 1) = p$$

In fact, we can say

$$E[\hat{P}_k(\mathbf{W}) \mid S_k(\mathbf{W}) \neq \emptyset] < p$$

Intuitively, this means that we expect a lower probability of success for this biased sampling technique.

For the complete proof of this theorem, see Miller and Sanjuro, but we will sketch it here.

Proof. First, we convert our expected value into a conditional probability where τ is a random trial drawn uniformly from all trials following a win streak of length k

$$E[\hat{P}_k(\mathbf{W}) \mid S_k(\mathbf{W}) \neq \emptyset] = P(W_\tau = 1 \mid S_k(\mathbf{W}) \neq \emptyset)$$

which we can then break down into a sum over all possible selections $\tau = t$

$$P(W_\tau = 1 \mid S_k(\mathbf{W}) \neq \emptyset) = \sum_{t=k+1}^{n} P(W_t = 1 \mid \tau = t, \prod_{i=t-k}^{t-1} W_i = 1) P(\tau = t \mid \prod_{i=t-k}^{t-1} W_i = 1)$$

Here, the rightmost probabilities sum to 1, so we just need to show the left is less than p in all iterations of the sum. Using Bayes's rule, we can get

$$\frac{P(W_t = 1 \mid \tau = t, \prod_{i=t-k}^{t-1} W_i = 1)}{1 - P(W_t = 1 \mid \tau = t, \prod_{i=t-k}^{t-1} W_i = 1)} = \frac{P(\tau = t \mid W_t = 1, \prod_{i=t-k}^{t-1} W_i = 1)}{P(\tau = t \mid W_t = 0, \prod_{i=t-k}^{t-1} W_i = 1)} \times \frac{p}{1-p}$$

so the left is less than p as long as

$$P(\tau = t \mid W_t = 1, \prod_{i=t-k}^{t-1} W_i = 1) < P(\tau = t \mid W_t = 0, \prod_{i=t-k}^{t-1} W_i = 1)$$

which is true because there are more possibilities for τ when $W_t = 1$. □

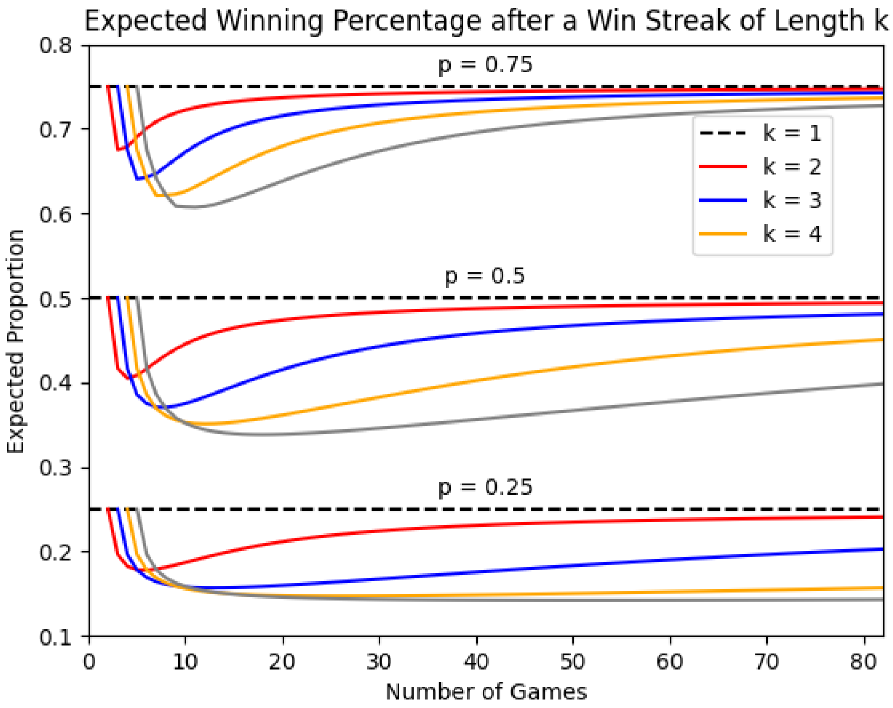
3.1.4 Quantifying the Bias

To see what an example of this selection bias looks like, consider taking $n = 3$ games, with a streak length $k = 1$ (i.e. sample after every win) and a win probability $p = 0.5$ (i.e. a coin flip). Below, we enumerate the possibilities and corresponding proportion of successes after each success, with the overall expectation of sampled trials being 5/12, which is less than the individual trial probability 0.5.

The Selection Bias in the Case of Three Games

Three-Game Sequence	Proportion of Ws
LLL	-
LLW	-
LWL	0
LWW	1
WLL	0
WLW	0
WWL	1/2
WWW	1
Expectation:	5/12

More generally, we can use the recursive algorithm in Miller and Sanjurjo. Below, we calculated the expected proportion of successes after streaks of various length k across $k \leq n \leq 82$ games. The bias clearly decreases with larger game samples, as both the sampling without replacement and overlap effects are decreased. Likewise, it has an inverse relationship with the independent probability p, as the lesser amount of overall wins increases the effect of sampling without replacement. Finally, it increases with k, for similar reasons of less sampled trials.



3.2 Streak Measures

Beyond the aforementioned bias, there are significant limitations in terms of the power of GVT's method of measuring the hot hand. This begs the question: how can we measure the hot hand in a more robust way? This section will consider using a variety of measures for streakiness, starting with the traditional Wald-Wolfowitz Runs Test and its limitations, continuing with a mathematically appealing class of inter-event time measures, and concluding with the empirically useful gap measure.

3.2.1 Wald-Wolfowitz Runs Test

This test counts the number of runs or streaks, i.e. subsequences made up of equal values, e.g.

$$\text{runs}(\{L, W, W, W, W, W, L, W, L, W\}) = 6$$

This test is appealing because its null hypothesis that the sequence is binomial allows for closed-form statistics, namely

- **Mean:** $\mu_r = \frac{2n_W n_L}{n_W + n_L} + 1$
- **Variance:** $\sigma_r^2 = \frac{2n_W n_L(2n_W n_L - (n_W + n_L))}{(n_W + n_L)^2(n_W + n_L - 1)} = \frac{(\mu - 1)(\mu - 2)}{n_W + n_L - 1}$
- **Standard score:** $z = \frac{r - \mu_r}{\sigma_r}$

- **p-value:** $p = \frac{2}{\sqrt{2\pi}} \int_{-\infty}^{-|z|} e^{-x^2/2}\mathrm{d}x$

where n_W represents the number of successes and n_L represents the number of failures.

However, the Runs Test does not address non-stationarity well given its exclusive focus on runs, resulting in low power to reject the null hypothesis. Thus, we want to consider other measures as well.

3.2.2 Measures of Inter-Event Times

One appealing option for measuring streakiness is a function of inter-event times, i.e. the number of trials between successes, as outlined in Zhang, Bradlow, and Small (2013). They specifically identified that measures of "clumpiness" should have the following four characteristics that allow for powerful determination of the hot hand:

1. **Minimum:** the measure should be minimized when events (wins/losses) are equally spaced
2. **Maximum:** the measure should be maximized if all the events are clumped together
3. **Continuity:** a small shift in events should only change the measure a small amount
4. **Convergence:** as events move closer, the measure should decrease

Theorem: Suitable Streak Measures

Any convex and symmetric function of inter-event times (IETs) satisfies the above four properties.

Moreover, they proved the above theorem, which gives us a wide range of potential measures to choose from, originating from a variety of different sources. This project in particular implements the following three measures:

1. **Second moment:** often utilized to describe and distinguish probability distributions

$$L_2 = \sum_{i=1}^{n+1} x_i^2$$

2. **Entropy:** a measure of uncertainty and disorder used in information theory

$$H_p = \sum_{i=1}^{n+1} x_i \log x_i$$

3. **Log utility:** used in economics, it “normalizes” the ranges being considered

$$M = -\sum_{i=1}^{n+1} \log(x_i)$$

3.2.3 Gap Measure

The final measure we will consider comes from Patrick R. McMullen and measures the “smoothness” of a team’s success throughout a set of games; in order to do so, it essentially measures how many games the team has won through the ith game compared to how many they would be expected to win based on overall winning percentage, summing up these differences. Mathematically, it is defined as

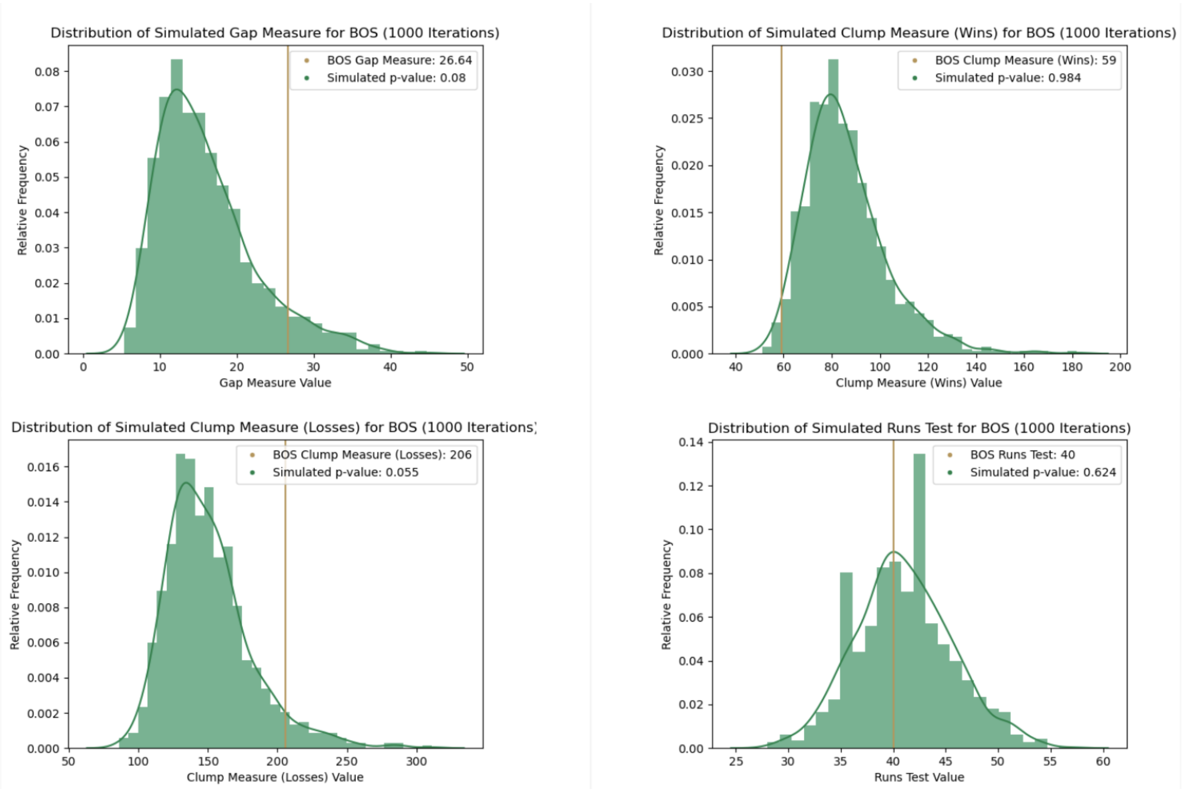
$$\text{gap} = \sqrt{\sum_{i=1}^{n} \left[\left(\sum_{j=1}^{i} w_j \right) - i * p \right]^2}$$

where w_i represents the number of successes through trial i and p represents the overall win percentage for the n trials.

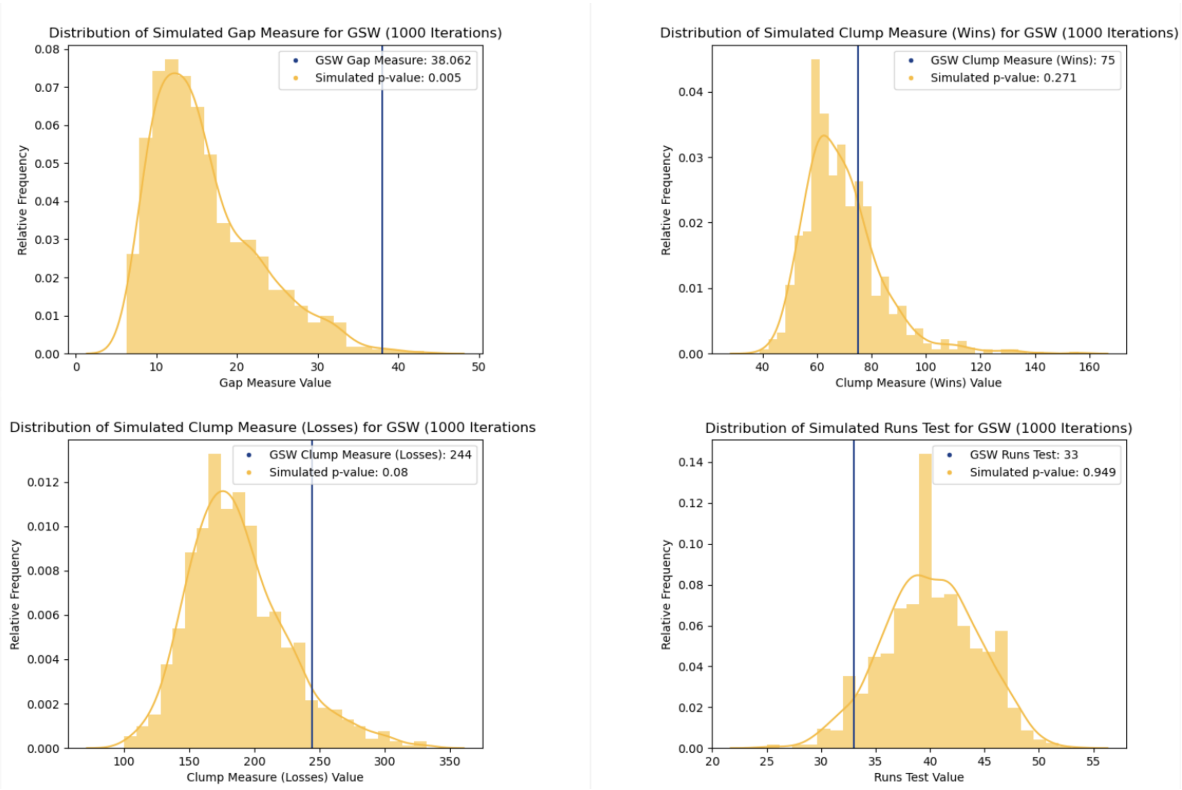
3.3 NBA Streakiness Significance Tests

Having established our desired set of streak measures, let’s apply them to actual NBA data. In particular, we can use Monte Carlo simulations to test whether a given measure for a given team implies high or low levels of streakiness. To do so, we randomly permute the game results for the given team some number of times. For each permutation, the measure is calculated and the number of times the measure is greater than or equal to the actual measure is counted. This count is then divided by the number of iterations to get the p-value of the actual measure. If the p-value is less than 0.05, the measure is considered significant at the 95% confidence level.

As illustrative examples, let’s consider the two 2021-22 finalists: the Boston Celtics and the Golden State Warriors. These two teams are interesting because they certainly had some element of streaky performance, with the Celtics beginning the season with a 18-21 record before a miraculous January turnaround resulting in a 51-31 final record. Likewise, the Warriors started off 41-13, then were plagued with injuries for a 7-16 stretch, before winning the final 5 games to finish with a 53-29 record. Thus, both of these teams should have interesting results in these simulations.



For the Celtics, we can see a relatively high gap measure, but it is not quite statistically significant at a 95% confidence level. Similarly, the clump measures for wins and losses (equivalent to the aforementioned Second Moment) are both nearly significant, demonstrating a fair amount of streakiness, both in terms of wins and losses. This effect is completely lost on the Runs Test, as the Celtics primarily experienced the type of non-stationarity barely considered by the structure of the test.



The Warriors have an extremely high gap measure, likely caused by the roller coaster nature of their season. At no point were they consistently performing at their final win percentage, with fluctuations from much above, to much below, to undefeated. This demonstrates the strength of the "two-sided" nature of the gap measure, which picks up on both hot and cold streaks. Moreover, the Runs Test is also quite significant, with an abnormally low amount of runs. Perhaps this illustrates the long hot and cold streaks of the Warriors, which still maintained long runs (and hence few runs overall) despite overall variability, or it may even hint at a traditional autocorrelative hot hand effect. Interestingly, neither of the clump/second measure statistics are significant at all, likely because of these extended periods of non-stationarity.

4 Elo Game Predictions

Because it is widely used, proven, and non-parametric, the Elo model serves as a strong starting point for our streak-sensitive models. Moreover, their basic structure allows for a convenient adjustment described at the end of this section.

4.1 Elo Model Details

At its core, the Elo model assumes that performance of each player in each game is a normally distributed random variable, essentially making it a normal or logistic regression model with the properties below.

- **Independent variables:** team ratings, represented by

$$\begin{aligned} R_A &= \text{rating for Team A} \\ R_B &= \text{rating for Team B} \end{aligned}$$

- **Dependent variables:** game results, represented as

$$S_A = \begin{cases} 1 & \text{Team A win} \\ 0 & \text{o.w.} \end{cases}$$

- **Probabilistic model:** the probability of Team A (with rating R_A) defeating Team B (with rating R_B) is

$$P(\text{Team A win}) = E_A = \sigma(r_{A,B}) \ \text{ with } \ \sigma(r) = \frac{1}{10^{-r/s} + 1} \ \text{ and } \ r_{A,B} = R_A - R_B$$

  Correspondingly, this also yields

$$P(\text{Team B win}) = \sigma(-r_{A,B}) = 1 - \sigma(r_{A,B})$$

Now, updating the model is equivalent to minimizing the log loss of the prediction, represented by:

$$\ell = \begin{cases} -\log \sigma(r_{A,B}) & S_A = 1 \\ -\log \sigma(-r_{A,B}) & S_A = 0 \end{cases}$$

which we can minimize with stochastic gradient descent

$$R'_A = R_A - \eta \frac{d\ell}{dR_A} = \begin{cases} R_A + \eta \frac{d}{dR_A} \log \sigma(r_{A,B}) & S_A = 1 \\ R_A - \eta \frac{d}{dR_A} \log \sigma(-r_{A,B}) & S_A = 0 \end{cases}$$

resulting in our update rule

$$R'_A = \begin{cases} R_A + K\sigma(-r_{A,B}) & S_A = 1 \\ R_A - K\sigma(r_{A,B}) & S_A = 0 \end{cases} \text{ with } K = \frac{\eta \log 10}{s}$$

or, more compactly

$$R'_A = R_A + K(S_A - E_A)$$

4.2 Gap Measure Adjustment

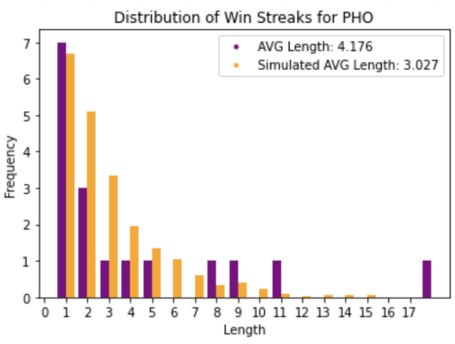
The key desired characteristic for our models is a sensitivity to recent performance, i.e. a method of incorporating the streakiness of the team. Thus, a natural modification of the Elo model is scaling the K-factor, intuitively equivalent to the "update velocity" or "learning rate." This is a common adjustment, e.g. the official chess ratings updating more slowly (lower K-factor) for more experienced players, soccer ratings updating more significantly for bigger tournaments (higher K-factor), and baseball ratings updating more for larger margins of victory (scaling the K-factor by winning margin).

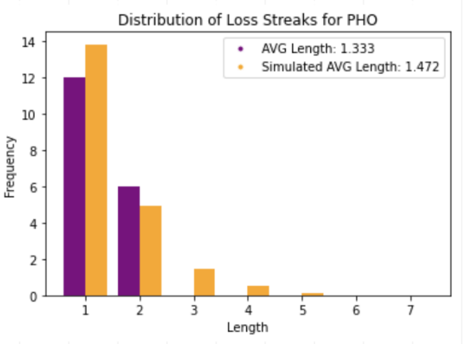
According to FiveThirtyEight, an optimal K-factor for NBA data is 20. For comparison, chess ratings use K-factors of 30, 15, or 10, popular soccer models use K-factors between 20 and 60, and baseball models utilize single digit K-factors. A credible NBA Elo model will likely maintain a K-factor close to this "optimal" value. Looking at our streak measures, one stands out for having values quite close to 20: the gap measure. Thus, one appealing possibility is to set the K-factor for a team to exactly the gap measure for a given team. In particular given a game result of S_A and expected result of E_A, we will update Team A's rating with the following formula

$$R'_A = R_A + K(S_A - E_A) \text{ with } K = \sqrt{\sum_{i=1}^{n} \left[\left(\sum_{j=1}^{i} w_j \right) - i * p \right]^2}$$

4.3 NBA Predictions

First, in order to further justify the aforementioned gap measure adjustment, consider simulating the season with a constant Elo rating for each team, say their rating at the end of the season. Below, we have a particularly interesting case study of the 2021-22 Phoenix Suns, who finished with the best record in the league at 64-18, with several long winning streaks throughout this run.





Seemingly, this static rating significantly underestimates the Suns streakiness, demonstrating the need for some hot hand adjustment to the Elo model.

With our gap adjustment justified, let's look at a comparison of a few "vanilla" Elo ratings with their gap-adjusted equivalents. In particular, let's consider the cases of the 2021-22 Boston Celtics and Golden State Warriors, who we also investigated in the **NBA Streak Measures** section. From the streak measures, we know these teams were relatively streaky, with Boston having a huge mid-season turnaround and Golden State having the highest gap measure. For one additional interesting case, we will look at the 2021-22 Portland Trail Blazers, who were a bad team throughout the season but finished significantly worse (on an 11 game loss streak) as a result of injuries (namely to Damian Lillard) and subsequent tanking.

Team	Vanilla Elo	Vanilla Loss	Gap Elo	Gap Loss
BOS	1626.029890	-0.125794	1640.973560	-0.128246
GSW	1556.567220	-0.134005	1552.793733	-0.132737
POR	1321.488577	0.185259	1305.674761	0.185031

We can see that the Gap Elo for Boston is significantly higher than the Vanilla Elo rating, and significantly lower for Portland. Both of these are desirable results, as the Gap Elo better captures Boston's rise that ultimately led to the finals and Portland's outright tanking at the end of the season. The difference is far less pronounced for Golden State, perhaps reflecting their streakiness across the season rather than specifically leading into the playoffs.

Moreover, we calculated the loss of both models for each of these teams by finding the mean difference between the Elo model predictions and the actual game results. The loss is very similar regardless of the gap adjustment, which is a good result, as it demonstrates similar long-term performance of the Gap Elo model to the proven Vanilla Elo model. One potential area for further adjustment is that the loss is negative for the two “good” teams (underestimating performance) and positive for the “bad” team (overestimating performance). Nevertheless, the overall results of this experiment are positive and demonstrate some significant strengths of the Gap Elo in our cases of interest.

5 Conclusions

In the end, this project was quite informative and successful, with historical research and literature review yielding conclusive affirmation of the existence of the hot hand and powerful methods for observing it in real NBA data. The simulations of these measures appeared to detect empirical underlying trends, as explained in our case studies, and when the measures were applied to our Elo model, the ratings seemed to more accurately reflect short-run performance.

In the future, there is lots of room for expansion of this project in all components, including

- Mathematical derivations of additional streak measures that account precisely for hot hand characteristics (autocorrelation and/or non-stationarity), perhaps by studying the corresponding deviations from the Binomial distribution
- More nuanced Elo models with adjustments for each of the measures, not just those that can be easily extrapolated to K-values
- Elo models with more adjustments beyond streakiness, e.g. home/away or injuried players, similar to the work of FiveThrityEight
- Experimentation with different types of models, such as those using a Bayesian framework of statistical priors or machine learning models
- Application of the hot streak principles to individual player performance, more akin to the idea of "streaky shooters" like GVT

Overall, I learned a lot throughout this project, both from a technical standpoint, as I was exposed to new research and statistical methods, and from a project management standpoint, as I had to turn an original idea into a viable project according to my own timelines and goals. While not without its struggles and delays, I am quite happy with the end result. I am probably most pleased with extensible and generalizable codebase and interactive website, which could be easily modified for any of the aforementioned improvements and expansions. Moreover, it could be transferred to any sport or league, as very little was specific to the NBA or the sport of basketball. Perhaps comparison across sports will inspire the measures and models that truly unravel the mystery of the hot hand once and for all.

Ultimately, the results of this project have more than justified further research into these topics, and I look forward to continuing to narrow down more accurate and expansive sports models in the future. And with that, I just want to thank my advisors, professors, friends, and family that supported me along the way! I hope you enjoyed the journey and results as much as I did!

6 References

[1] *NBA Logo*. URL: https://www.nba.com.

[2] *Cal Ripken Quote*. AZ Quotes. URL: https://www.azquotes.com/quote/712065.

[3] Bethlehem Shoals. *The 25 Most Unbreakable Streaks in Sports*. GQ. URL: https://www.gq.com/gallery/greatest-sports-streaks.

[4] Raghav Mittal. *What is an ELO Rating?* Medium. URL: https://medium.com/purple-theory/what-is-elo-rating-c4eb7a9061e0.

[5] Nate Silver and Reuben Fischer-Baum. *How We Calculate NBA Elo Ratings*. FiveThirtyEight. URL: https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/.

[6] *How Our NFL Predictions Work*. FiveThirtyEight. URL: https://fivethirtyeight.com/methodology/how-our-nfl-predictions-work/.

[7] ESPN Analytics. *ESPN's Basketball Power Index*. ESPN. URL: https://www.espn.com/nba/story/_/page/Basketball-Power-Index/espn-nba-basketball-power-index.

[8] Mike Lynch. *SRS Calculation Details*. Sports Reference. URL: https://www.sports-reference.com/blog/2015/03/srs-calculation-details/.

[9] Sharon Katz. *How ESPN's NFL Football Power Index was developed*. ESPN. URL: https://www.espn.com/nfl/story/_/id/13539941/how-espn-nfl-football-power-index-was-developed-implemented.

[10] Sascha Wilkens. "Sports prediction and betting models in the Machine Learning Age: The case of tennis". In: *Journal of Sports Analytics* 7.2 (2021), pp. 99–117. DOI: 10.3233/jsa-200463.

[11] Joshua B. Miller and Adam Sanjurjo. "Surprised by the Gambler's and Hot Hand fallacy? A Truth in the Law of Small Numbers". In: *Econometrica* 86.6 (Dec. 2018), pp. 2019–2047. DOI: 10.2139/ssrn.2627354.

[12] Thomas Gilovich, Robert Vallone, and Amos Tversky. "The hot hand in basketball: On the misperception of random sequences". In: *Cognitive Psychology* 17.3 (1985), pp. 295–314. DOI: 10.1016/0010-0285(85)90010-6.

[13] Yao Zhang, Eric T. Bradlow, and Dylan S. Small. "New Measures of Clumpiness for Incidence Data". In: *Journal of Applied Statistics* 40.11 (2013), pp. 2533–2548. DOI: 10.1080/02664763.2013.818627.

[14] Patrick R. McMullen. "Standardization of winning streaks in sports". In: *Applied Mathematics* 08.03 (2017), pp. 344–357. DOI: 10.4236/am.2017.83029.

[15] Jim Albert. *Streaky and Consistent Teams?* Sept. 2017. URL: https://baseballwithr.wordpress.com/2017/09/18/streaky-and-consistent-teams/.