Imperial College London
Department of Mathematics

Assessing Live Performance in Professional Football Players

Francesco Ventura

CID: 02488337

Supervised by Dr. Adam Sykulski and Dr. Tim Blackmore

August 30, 2024

Submitted in partial fulfilment of the requirements for the MSc in Statistics at Imperial College London

The work contained in this thesis is my own work unless otherwise stated.

Signed: Francesco Ventura Date: August 30, 2024

Acknowledgements

First, I would like to thank my supervisors, Dr. Adam Sykulski of Imperial College London and Dr. Tim Blackmore of Sportlight Ltd. They dedicated their expertise to shape this project by providing insightful suggestions and thoughtful feedback.

Second, I wish to express my gratitude to Sportlight Ltd for providing the data that served as the foundation of this study. I am especially thankful to Dr. Ian Cowling for his generous assistance.

Last but not least, I wish to thank my family, my friends, and my girlfriend. Their constant support is what got me through this journey and ultimately made me achieve this important milestone.

Contents

Abstract

Statistics is already playing a crucial role in gaining a competitive edge in professional football. This study explores a novel methodology to assess the physical performance of football players, both during and post-game. The methodology was applied to highly accurate data regarding seasons 2022/2023 and 2023/2024 of one Premier League club. The main focus is to analyse the extreme values of absolute acceleration, which are then compared to the ones of the player's nearest counterpart on the opposing team in order to identify under- or overperformance in real-time or post-game. The extreme observations were selected using Peak Over Threshold (POT) analysis and compared using the Kolmogorov-Smirnov two-sample test. This methodology offers a tool to facilitate the evaluation of the physical performance, which can lead to more informed decision-making processes. Some concrete live decisions it could help with are player substitutions or tactical changes, while it could help long-term with adjustments of the player's role in the team or their skill development plan.

1 Introduction

Football (Soccer) is the most popular sport worldwide, which engages and brings together almost half of the earth's population. It is estimated that the number of football fans for summer 2024 was around 3.5 billions, around a billion more than Cricket, the second sport in popularity (Remitly, Inc., 2024). This incredible following, coupled with entrepreneurship abilities, made big football clubs become companies with hundreds of millions of euros in yearly revenue, Figure 1.

This wealth is undoubtedly one of the main factors contributing to the increasing budget dedicated to data analysis by football clubs as described in Sportlight Technology Ltd (2024). These investments created a market for companies like *Sportlight Technology Ltd* (https://www.sportlight.ai), a leading sports tech company providing patented

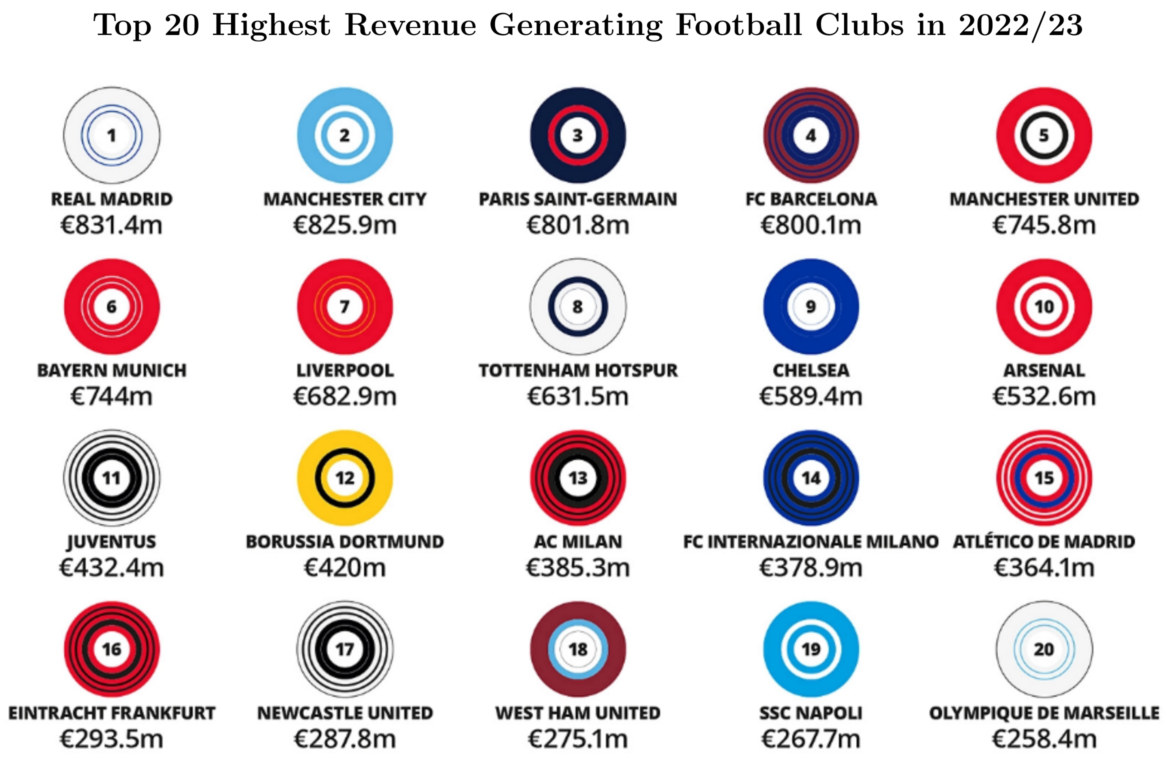


Figure 1: Football clubs revenue of 2022/2023 in millions of euros, Deloitte LLP (2024)

and military-grade technology regarding players' movements to professional sports organisations. The company is currently working with the majority of the English Premier League, producing hyper-accurate, relevant, and rich insights to promote data-led decision-making. This research stems from a partnership between Imperial College London and Sportlight Technology Ltd, which generously provided mentorship and the data upon which this research is based. The data will be described in more detail in Section 2.1.

1.1 Aim of the Research

Sportlight as a company is heading towards developing new data products that could benefit their customers. In particular, they focus on providing meaningful live data analysis to help the coaching staff make in-game decisions. To this end, in this report,

we describe a methodology to provide live insights on a player's performance compared to its nearest neighbour from the opposing team in a short time window. Furthermore, we described and gave an interpretation of the results obtained by applying the aforementioned methodology.
The coaching staff can use these results to make in-game substitutions, quick tactical adjustments, or even long-term changes in a player's role or position within the team.

1.2 Performance measurement

Measuring football players' fatigue/performance is a hot topic in the literature, especially in recent years (Evans et al., 2022; Lourenço et al., 2023). Football is a complex sport that requires technical ability, strategic thinking, and intense athletic performances. Furthermore, a single player's contribution cannot only be measured by their individual performance, but it must also consider their impact on the team's overall performance, both from a tactical and a psychological point of view. One metric, however complex it may be, will never be enough to capture these aspects fully. This research aims to provide coaches and their coaching team with another arrow on their quiver rather than a guide that must be followed blindly.

To provide useful insights, we focused on monitoring a player's individual performance throughout the game. One of the most common metrics other studies rely on is absolute acceleration, i.e., acceleration and deceleration treated indistinguishably, see Wilson (2022). Antonio Pintus, a Real Madrid fitness coach, mentioned how "Analyzing the data after a match is not that easy: sometimes you run more, and it does not mean that you will win. There are statistical studies that show running more does not always lead to success. What matters are the accelerations and decelerations: the ability to sprint, brake, and change direction. Running more is a different matter." translated from an interview by Sky international AG (2024). Furthermore, from Thoseby et al. (2023), "The data suggest that the development of acceleration and repeat effort capacities is crucial in youth players for them to transition into professional competi- tion", which testifies how crucial acceleration is as a performance indicator.

The two main ways in which absolute acceleration has been analysed is through its average over time or distance (Wilson, 2022), or through its extreme values (Daykin, 2023). Together with Sportlight Technology Ltd, we decided to focus on the absolute acceleration extreme values.

The choice was dictated by the consideration that the footballers analysed are all high-level professionals and therefore elite-level athletes. Their average physical performance is expected to be above a minimum standard and relatively constant throughout the game, see Figure 2, making the detection of a change in performance harder. Furthermore, the high volume of data points makes the extreme values analysis feasible. Other studies on high-intensity acceleration have shown how a higher frequency of accelerations and decelerations correlated with a better game outcome (Rhodes et al., 2021). The study, though, is limited to one team playing in English League Two, which is the fourth-highest tier of English Football.

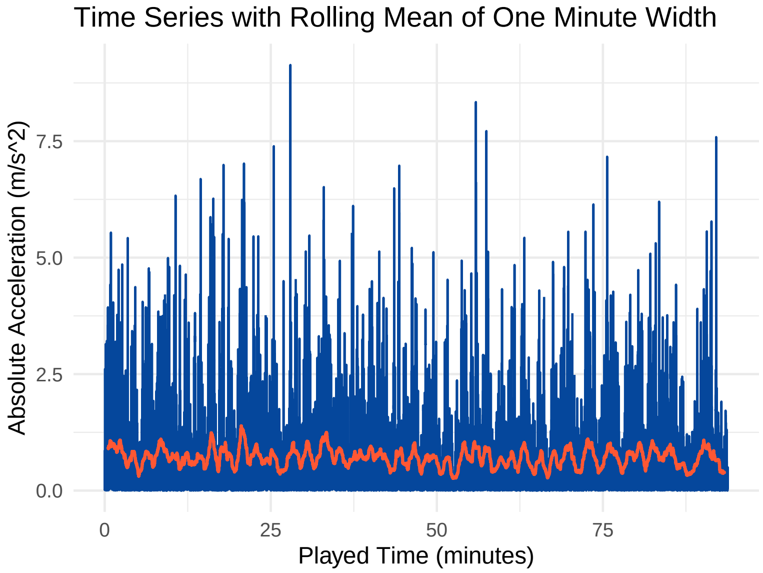


Figure 2: Time series of the absolute acceleration with its rolling mean (one-minute window) for a right back during a Premier League game.

It is worth noting that in football, different teams have different game tactics, which require very different actions depending on the role covered. More specifically, mental fatigue is much more common for goalkeepers than physical one, (Abbott et al., 2023).

Midfielders will generally need to make more turns and tackles than Wingers or Fullbacks, which are also highly exhausting activities. Depending on their characteristics, Center Forwards and Central Defenders may need to battle each other physically more than sprinting often. All this is to say that acceleration may not always be the leading cause of fatiguing/performance reduction, but it is undoubtedly one of its effects. This implies that analysing absolute acceleration is sensible for all the roles except for Goalkeepers, even though coupling acceleration with other role-specific metrics may give more accurate results. This idea is explored further by Toni Modric and Liposek (2019).

2 Materials and Methods

2.1 Data

The data used for the analysis have been fully provided by Sportlight Technology Ltd. It is composed of a main data set with in-game measurements for every player on the pitch and many additional data sets reporting indicators and statistics computed by the company, which will not be central for this analysis.

2.1.1 Data Description

The main data set refers to a Premier League team and contains data for games across all the competitions they participated in in season 2022/2023 and 2023/2024. For each game, the main data set contains the the positions of each player on the pitch, expressed as xy coordinates, with a frequency of ten times per second. The data set included the following variables that turned out to be key for the analysis:

- **section_id**: ID code to identify the considered game;
- **player_id**: ID code to identify the considered player in the section;

- **stamp**: time stamp at which the observation was taken expressed in nanoseconds passed from 1 January 1970 00:00:00 UTC;
- **y**: y coordinate on the pitch of the player considered in meters (m);
- **x**: x coordinate on the pitch of the player considered in meters (m);
- **speed**: speed of the player considered measured in meters per second (m/s). It is obtained by computing the euclidean distance from the previous observation and dividing by the time difference between them (1/10 s). It is null for the first observation. In symbols:

$$speed_n \, m/s = \begin{cases} \frac{\sqrt{(x_n - x_{n-1})^2 + (y_n - y_{n-1})^2}}{0.1} \, m/s & \text{if } n > 1 \\ 0 \, m/s & \text{if } n = 1 \end{cases} \quad ;$$

- **acceleration**: acceleration of the player considered measured in meter per second squared (m/s^2). It is obtained by computing the difference between the speed in the previous observation and the current one and afterward dividing it by the time difference between them (1/10 s). It is null for the first observation. In symbols:

$$acceleration_n \, m/s^2 = \begin{cases} \frac{speed_n - speed_{n-1}}{0.1} \, m/s^2 & \text{if } n > 1 \\ 0 \, m/s^2 & \text{if } n = 1 \end{cases} \quad .$$

The technology used to measure the players' coordinates on the pitch is called LiDAR, which stands for *Light Detection And Ranging*, and it is a highly accurate tracking method that was originally developed for military purposes. A detailed description of the technology is provided by Synopsys, Inc. (2024); furthermore, in Bampouras and Thomas (2022) are presented the advantages of this technology with respect to other methods like GPS.

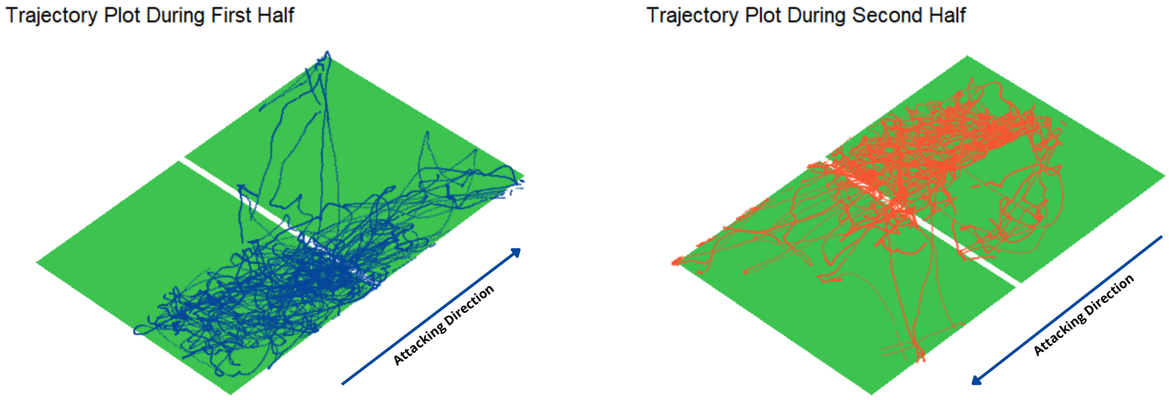


Figure 3: Trajectory plot of a right-back on a schematic football pitch of a Premier League game during first half (left) and second half (right).

2.1.2 Missing Data

Even if the technology is extremely accurate, there are cases in which several players are very close to one another and neither one of the two sensors can clearly distinguish the position of each player individually. These missing observations may even represent up to 10% of the player's data set for one game. The gaps are generally isolated; i.e., it is rare to have missing data for more than a few seconds. From a temporal point of view, the missing entries appear to be randomly distributed, Figure 4 (left) and Figure 5 (left). On the other hand, the missing stamps seem to be clustered spatially, in particular the proportion of missing values inside the box appears higher than in the other areas of the pitch. This can be appreciated when comparing Figure 4 (right) and Figure 5 (right) with Figure 3.

For the analysis, we decided not to impute the missing entries, because when there are multiple bodies close to one another it is unlikely to register high values of absolute acceleration.

Missing Positional Data in a Premier League game, First Half

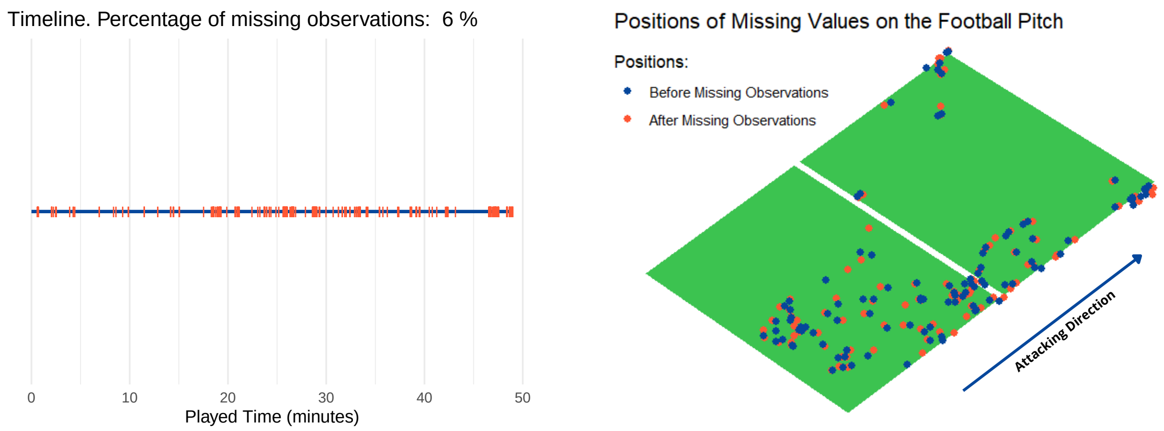


Figure 4: Blue timeline with red bars corresponding to the missing entries (left). Football pitch schematic representation with player's positions before and after the missing entries (right). Both figures regard a right-back in a Premier League game during the first half, in the same game as Figure 5.

Missing Positional Data in a Premier League game, Second Half

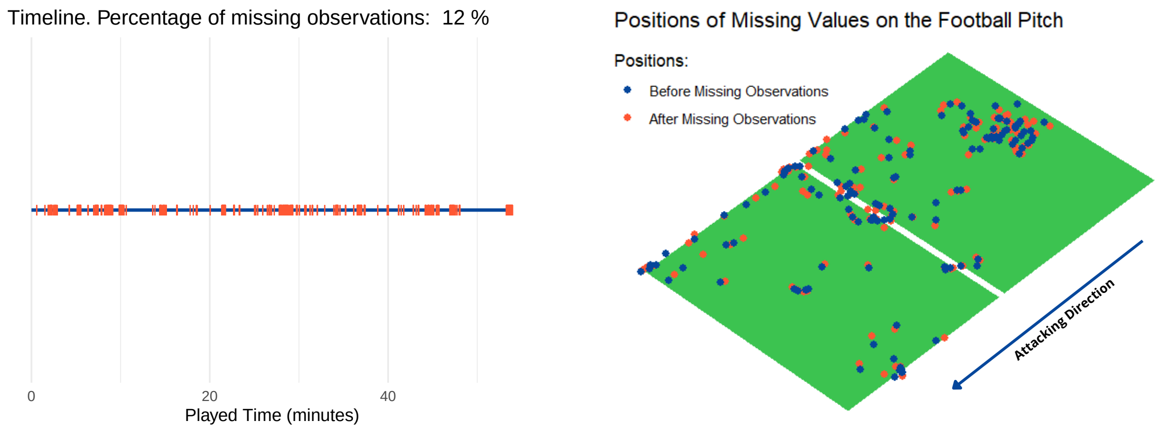


Figure 5: Blue timeline with red bars corresponding to the missing entries (left). Football pitch schematic representation with player's positions before and after the missing entries (right). Both figures regard a right-back in a Premier League game during the second half, in the same game as Figure 5.

2.2 Peak Over Threshold (POT)

As previously mentioned in Subsection 1.2, this report's analysis focuses on extreme absolute acceleration. The two classical methods to categorise observations as extreme are Block-Maxima (BM) and Peak Over Threshold (POT). Essentially, the standard BM consists of dividing the data set into batches and considering the maxima of each batch as extreme values. On the other hand, POT consists in fixing a threshold ($u_0 \in \mathbb{R}$) and considering as extreme values, called exceedances ($x_j \in \mathbb{R}$, where $x_j > u_0$ for $j = 1, 2, \ldots, n$ and $n \in \mathbb{N}$ is the total number of observations), the observations that exceeded the threshold and as excesses ($y_j = x_j - u_0$ for $j = 1, 2, \ldots, n$) the difference between the exceedances and the threshold itself. As a result, the excesses, and consequently the exceedances, follow a Pareto distribution (Pickands, 1975). Depending on the nature of the data, one approach may be more suitable than the other (Szubzda and Chlebus, 2019). I chose POT over BM because of the non-stationary nature of a football game rhythm. By utilising BM, I could have categorised as extreme values that are small in modulus simply due to a period of lower game intensity. This would make the detection of a decay of the peaks due to a worse overall performance much harder.

The main issue with the POT method is the choice of an appropriate threshold. Too low of a threshold would be a source of bias because we would consider values that are not extreme. This would break the assumption that the excesses would be Pareto distributed. Too high of a threshold can result in a small number of excesses, causing a high sample variance, which makes the results not generalisable. There is no analytical solution to select a threshold for a perfect bias-variance trade-off, but some empirical solutions are proposed by Coles (2001). The proposed empirical methods all exploit the fact that if the excesses y_j are Pareto distributed, with scale parameter σ and shape parameter $\xi < 1$, their expectation $\mathrm{E}[(y_1, \ldots, y_n)] = \mathrm{E}[(x_1, \ldots, x_n) - u_0] = \sigma/(1 - \xi)$ is linear in $u, \forall u > u_0$ (Coles, 2001). This means that, by plotting the excesses average against the thresholds, the ideal u_0 would be the smallest value, after which there is a sign of linearity in the curve. This process is quite computationally expensive; therefore, we opted for other threshold choices.

- **Threshold** $= 3m/s^2$: It is common practice in the industry to consider accelerations over $3m/s^2$ as high-intensity, see Daykin (2023); Dalen et al. (2016). We decided to consider it as one of the thresholds to give continuity to the current literature, choosing the threshold that could be meaningful in the context.

- **Player Specific Threshold**: The reasoning behind performing the analysis with a player-specific threshold was that different positions on the pitch require different acceleration intensities. We picked each player's 99^{th} percentile respectively to ensure that the excesses would follow a Pareto distribution. The percentile is computed up to the last game played before the one under analysis; therefore, it is constantly updated during the season.

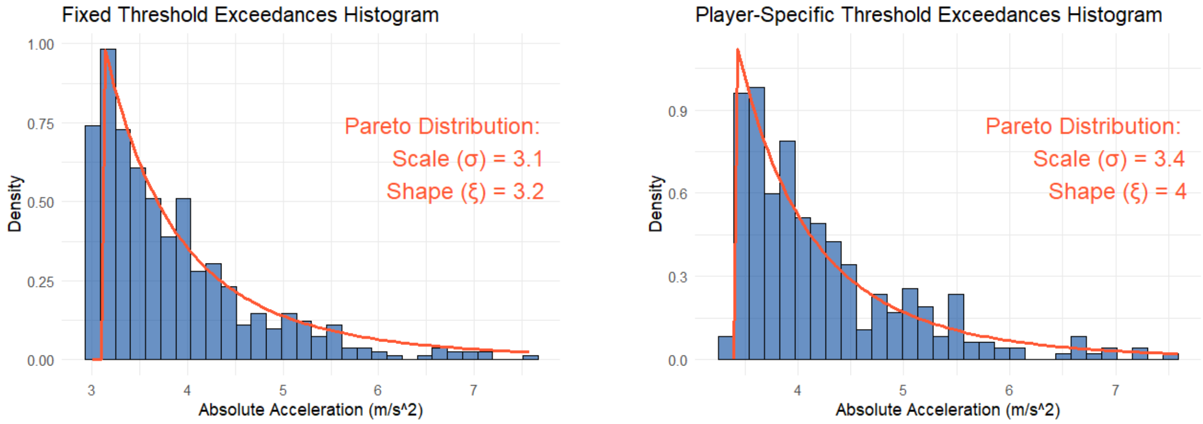


Figure 6: Comparison between the empirical densities of absolute acceleration exceedances, with the curve of an arbitrary Pareto distribution, when using the Fixed Threshold (left) and the Player-Specific Threshold (right). It refers to the absolute acceleration of a right-back during a Premier League game.

2.3 The Kolmogorov-Smirnov Two-Sample Test

The method I used to compare the distributions of the Absolute Acceleration peaks is the Kolmogorov-Smirnov (KS) two-sample test (Definition 2.1) because it compares the entirety of the empirical distributions, providing more meaningful conclusions than the tests that only compare the mean values. Among these tests, we chose the KS test due to the distribution of the peaks. They follow a Pareto distribution as described in

Subsection 2.2 and the KS test is shown by Chu et al. (2019) to be the most powerful nonparametric test when performing Goodness of Fit (GOF) tests for Pareto distributions. Even though this analysis doesn't strictly fall into the GOF category, we am still comparing two empirical distributions that belong to the Pareto family, therefore, the argument by Chu et al. (2019) of the KS test to be the most powerful among the nonparametric tests is still valid in this situation.

Definition 2.1. The **Kolmogorov-Smirnov two-sample test** (sometimes called simply the Smirnov test) is based on the maximum difference between the two Empirical Cumulative Distribution Functions (ECDFs) $F_m(t)$ and $G_n(t)$ of the two samples of sizes $m \in \mathbb{N}$ and $n \in \mathbb{N}$ respectively. Specifically, the test statistics are given by

$$D_{mn} = \max_t |G_n(t) - F_m(t)| \tag{1a}$$

$$D^+_{mn} = \max_t [G_n(t) - F_m(t)] \tag{1b}$$

$$D^-_{mn} = \max_t [F_m(t) - G_n(t)]\,, \tag{1c}$$

where $\max_t$ denotes the maximum over all t.
Equation (1a) is called the two-sided statistic since the absolute value measures differences in both directions, and Equations (1b) and (1c) are called the one-sided statistics. The null hypothesis is always $H_0 : F = G$, while the alternative hypothesis H_1 are $F \neq G$, $G \geq F$ and $G \leq F$ for Equations (1a), (1b) and (1c) respectively. In comparing the two ECDFs, the equality and inequality signs have to be interpreted in a stochastic sense (Pratt and Gibbons, 1981).

The process to analytically find the null distributions for the one-sided and the two-sided statistics when $m = n$ is described in Pratt and Gibbons (1981). The results, for

$k = 1, 2, \ldots, n$ are:

$$P\left(D_{nn}^{+} \geq k/n\right) = \binom{2n}{n-k} \Big/ \binom{2n}{n} = (n!)^2 / \left[(n+k)!(n-k)!\right]$$

$$P\left(D_{nn}^{-} \geq k/n\right) = \binom{2n}{n-k} \Big/ \binom{2n}{n} = (n!)^2 / \left[(n+k)!(n-k)!\right]$$

$$\begin{aligned} P\left(D_{nn} \geq k/n\right) &= 2\left[\binom{2n}{n-k} - \binom{2n}{n-2k} + \binom{2n}{n-3k} - \cdots\right] \Big/ \binom{2n}{n} \\ &= 2\sum_{i=1}^{[n/k]} (-1)^{i+1} \binom{2n}{n-ik} \Big/ \binom{2n}{n} \end{aligned}$$

The statistical tables reporting the rejection regions can be found in Harter and Owen (1970).

In this analysis, we applied only the one-sided versions for better interpretability of the results. In particular, the two-sided test would detect the difference between the acceleration peaks distributions but not whether this difference is positive or negative, not differentiating an overperformance from an underperformance.

2.4 Time Division and Nearest Neighbours

We decided to conduct our analysis, considering every game's half-time as a separate game. This choice was dictated purely by the context of a football game. In fact, after the first half, the players return to the locker rooms to rest for 15 minutes and adjust the game tactics with their coaches. Even though most of the players on the pitch during the second half may be the same as during the first, it can be seen as a sort of reset of the game conditions since the context of the game can change drastically. This is what ultimately determined our choice.

As discussed in Subsection 1.1, one of the goals of the investigation is to build an in-game analysis. To this end, we decided to divide each half into smaller periods and perform the KS two-sample tests in each one. Past research on the topic, such as Escuret (2023), subdivided the regular game into fifteen-minute windows and the extra time. Extra time in football is given after the 45^{th} minute in first-half and/or after the 90^{th} in second half to compensate for non-playing moments during the regular 90 minutes. After discussing the matter with Sportlight, we reduced the window size to 5 minutes to provide coaches and their teams with faster test results. Furthermore, we aggregated the extra time to the corresponding half and considered it a stand-alone window irrespective of its duration. It is worth noting that shortening the window size makes the result of each test less accurate and more sensitive to noise, i.e., less reliable. The concrete implications of this will be discussed in Section 3.

To make a comparison that would consider the game conditions, we compared the players from the analysed team with their nearest neighbor among the opposition, within each time window. By window nearest neighbour we mean the player from the opposing team that, on average, is closest to the analysed player from our team within one window. The type of distance we used is the classical euclidean distance in meters (m) computed using the xy coordinates of each player in symbols:

$$distance\, m = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}\, m \qquad .$$

2.5 Multiple Testing Issues

The small size of the window implies that the number of tests carried out is relatively high, even considering every half as a separate game. There are 9 five-minute windows of regular time plus 1 extra time, and for each of them, we carry out two different KS tests assuming that the analyzed player played throughout the game; we have a total of 20 tests per half per player. Now, let us consider a standard Type I error $\alpha = 0.05$ and assume independence between the tests; the probability of making at least one Type

I error is $1 - 0.95^{20} \simeq 0.64$. Therefore, it is necessary to adopt some methodology to control the overall Type I error.

2.5.1 Standard Benjamini-Hochberg Method

The chosen method is the Benjamini-Hochberg (BH) method, first introduced in Benjamini and Hochberg (1995). This method does not control the Family-Wise Error Rate (FWER) as other popular methods like the Bonferroni method (Dunn, 1961), but instead controls the False Discovery Rate (FDR), which is the expectation of the False Discovery Proportion (FDP). These are the formal definitions of the quantities mentioned above.

Definition 2.2. The **Family-Wise Error Rate** (FWER) is defined as the probability of having one or more incorrect rejections in the family of tests. In symbols:

$$\mathrm{FWER} = \mathrm{P}[V \geq 1] \quad , \tag{2}$$

where $V \in \mathbb{N}$ is the number of incorrect rejections.

Definition 2.3. The **False Discovery Proportion** (FDP) is defined as the proportion of incorrect rejections. In symbols:

$$\mathrm{FDP} = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases} \quad , \tag{3}$$

where $V \in \mathbb{N}$ is the number of incorrect rejections and $R \in \mathbb{N}$ is the total number of rejections.

Definition 2.4. The **False Discovery Rate** (FDR) is defined as the expected value of the FDP. In symbols:

$$\mathrm{FDR} = \mathrm{E}[\mathrm{FDP}] = \begin{cases} \mathrm{E}[V/R] & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases} \quad , \tag{4}$$

where $V \in \mathbb{N}$ is the number of incorrect rejections and $R \in \mathbb{N}$ is the total number of rejections.

The BH method, which can also be denoted as BH(α) where α is the Type I error for a single test, consists of rejecting all the null hypotheses for which the p-value is smaller or equal to the Benjamini-Hochberg threshold $L \in \mathbb{R}$.

Example 2.5. Let us imagine that $m \in \mathbb{N}$ hypothesis tests have been carried out and consider the ordered vector of the resulting m p-values $P_{(1)} \leq \ldots \leq P_{(m)}$. Define

$$l_i = \frac{i\alpha}{m}, \text{ and } \hat{k} = \max\left(i : P_{(i)} \leq l_i\right) \text{ for } i = 1, 2, \ldots, m.$$

If $\exists\, \hat{k}$, the Benjamini-Hochberg threshold L is defined as $L = P_{(\hat{k})}$, otherwise $L = 0$. The null hypotheses H_{0i} for which $P_i \leq L$ are rejected.

In Theorem 2.6, it is stated how the BH(α) method controls the FDR at the level α under mutual independence of the p-values.

Theorem 2.6. Benjamini & **Hochberg**.

Let $m \in \mathbb{N}$ be the number of hypothesis tests carried out and suppose that all the p-values are mutually independent. If BH(α) is applied, then regardless of how many null hypotheses are true and regardless of the distribution of the p-value when the null hypothesis is false, FDR is controlled at level α. Specifically,

$$\text{FDR} = \text{E}[\text{FDP}] \leq \frac{m_0}{m}\alpha \leq \alpha \quad , \tag{5}$$

where $m_0 \in \mathbb{N}$ is the unknown number of true null hypotheses.

A proof of the Theorem can be found in Ferreira and Zwinderman (2006). If the mutual independence condition between the p-values is not satisfied, Theorem 2.7 provides a level at which the BH(α) controls the FDR.

Theorem 2.7. Benjamini & Yekutieli.
Let $m \in \mathbb{N}$ be the number of hypothesis tests carried out and $S(m) := 1 + \frac{1}{2} + \cdots + \frac{1}{m}$. Under dependence of the p-values, BH(α) controls the FDR at level $\alpha S(m)$, specifically

$$\mathrm{FDR} \leq \frac{m_0}{m} \alpha S(m) \leq \alpha S(m) \quad , \tag{6}$$

where $m_0 \in \mathbb{N}$ is the unknown number of true null hypotheses.

A proof of Theorem 2.7 can be found in Benjamini and Yekutieli (2001).
The dependence of the p-values within the same half will be analysed further in Section 3. It is worth mentioning that in our case, the maximum number of tests carried out every half is 20. Therefore, the FDR will be controlled at a level that ranges from α, in case of mutually independent p-values (Theorem 2.6), to $\alpha \cdot 3.43$ in case of complete dependence (Theorem 2.7). For a standard Type I error $\alpha = 0.05$ it means that FDR is controlled at a level between 0.05 and 0.18. Furthermore, the fact that one player is not necessarily compared to the same opponent throughout the game makes the p-values potentially more independent of one another, improving the control level of FDR.

2.5.2 Online Benjamini-Hochberg Method

The standard BH method requires all the p-values of the tests to be known in advance before concluding whether to reject a test or not. This requirement can be satisfied when analysing a past game, but it prevents the methodology from providing in-game insights. For this reason, I also considered an online version of the BH method in which only the p-values up to the last completed time window are considered, every time a new time window is completed, the list of p-values is updated and the BH method is re-ran. I will refer to it as the *Online BH* for simplicity.
The decision to reject or not a specific test may change due to the upcoming availability of new p-values that will impact the individual rejection regions. As a general statement, if the new p-value is larger than the current, the relative threshold for rejection l_i in Example 2.5 will be smaller. On the other hand, a smaller p-value follows a greater or

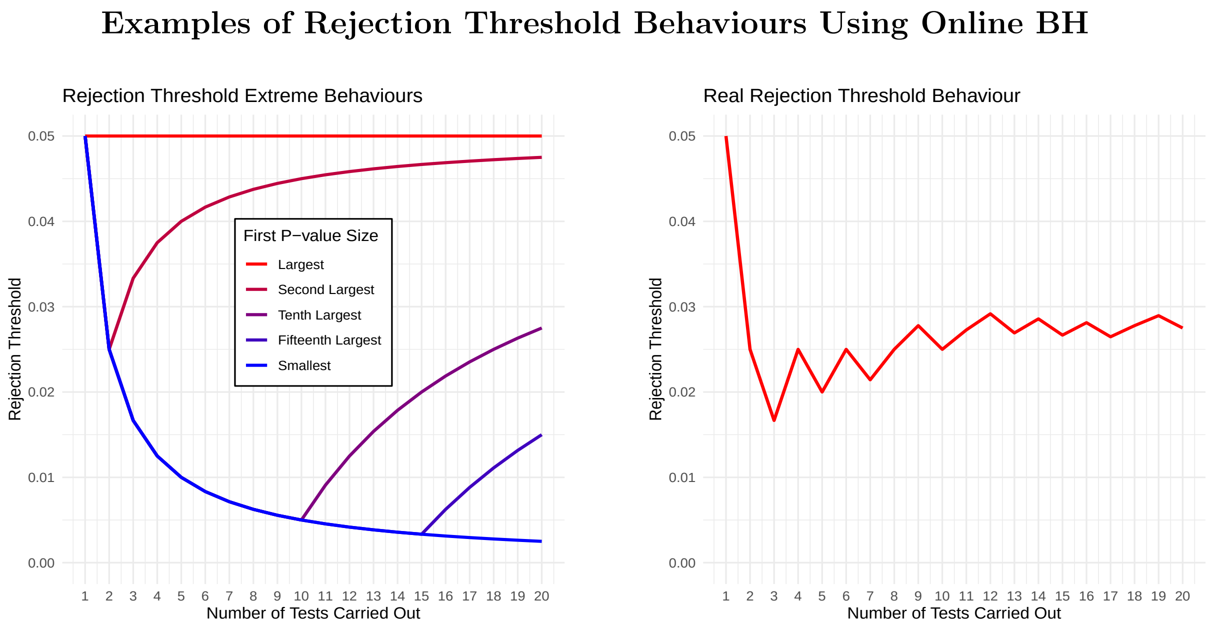
equal rejection threshold that will never exceed the chosen Type I error α. This change in the individual thresholds could potentially move the overall threshold L in the same direction as the individual ones.

Example 2.8. Let us imagine that $m = 5$ hypothesis tests have been carried out, the chosen Type I error is $\alpha = 0.05$, and consider the smallest p-value, $p = P_{(1)} = 0.013$, in the ordered vector of m p-values $P_{(1)} \leq \ldots \leq P_{(5)}$. By following Example 2.5, the individual rejection threshold to compare it with is $l_1 = \frac{1 \cdot 0.05}{5} = 0.010 \leq p$. In this case if p-values $P_{(2)} \leq \ldots \leq P_{(5)}$ were above their respective l_is, there would be no rejections among the m tests and $L = 0$.

Let us now assume that a new hypothesis test has been carried out $(m = 6)$, with a resulting p-value $q = 0.008$. Since $q \leq p$, $q = P_{(1)}$ and $p = P_{(2)}$ and the individual rejection threshold to compare p with becomes $l_2 = \frac{2 \cdot 0.05}{6} = 0.016 \geq p$. By still considering the other p-values $P_{(3)} \leq \ldots \leq P_{(6)}$ to be above their respective l_is, $L = p = 0.013$ and both hypotheses tests corresponding to the p-values p and q will be rejected.

To get a better idea of the behaviour of the individual thresholds, I plotted the actual behaviour of one in Figure 7 (right). Then I analysed the individual thresholds of some possible extreme scenarios and plotted them together in Figure 7 (left). The scenario in which 20 tests are carried out, and the first observed p-value is the smallest. The scenario in which 20 tests are carried out, and the first observed p-value is the largest. Additionally, I analysed three more individual thresholds: the one for the second largest, the tenth largest, and the fifteenth largest, assuming that the following p-values were initially all larger, making the thresholds decrease, and then all smaller making the threshold increase again.

A lowering of the individual rejection threshold makes a test harder to reject. This means that the initial decision could have potentially been less conservative than if we had all the p-values at our disposal, which implies an initial Type I error rate higher than appropriate. Vice versa, an increase of the individual rejection threshold causes a


Examples of Rejection Threshold Behaviours Using Online BH

Rejection Threshold Extreme Behaviours

Real Rejection Threshold Behaviour

Figure 7: Individual rejection thresholds (l_is) behaviour using Online BH method: extreme cases (left) vs. real data from a right-back in a Premier League game (right).

rejection to be more likely, implying a potentially too conservative initial decision.

To put the threshold changes in context, an overly conservative approach would fail to detect a sign of under- or overperformance on the first time carrying out the test. On the other hand, an approach that is not conservative enough is more likely to flag a difference in the absolute acceleration peaks due to noise or randomness rather than an actual difference in their distributions. The conclusion of one single test may be misleading, however, the severity of error is quickly alleviated by the following tests. In fact, the first few tests that follow correct the single rejection threshold much more than the last few, as can be be observed in Figure 7. The practical consequences of the Online BH on the methodology use will be discussed further in Section 3.

3 Results

3.1 Structure of the Overall Product

The analysis we designed is embedded in a data product that retrieves all p-values from the KS two-sample tests and highlights those indicating a rejection of their null hypothesis. Its steps are described in Algorithm 1.

It is important to remember that each half-time is considered a separate game.

Algorithm 1 Analysis Workflow

```
Input: Player Dataset, Opposing Team Dataset, Absolute Acceleration Threshold
Workflow:
Divide datasets into five-minute windows plus extra time        ▷ Up to 10 windows
for each window do
    Compute Euclidean distance of each opposing player from the analyzed player
    Identify the nearest neighbor among the opposing players
end for
Filter Opposing Team Dataset to include only the nearest neighbor in each window
Filter both datasets to include only absolute accelerations over the threshold
for each window do
    Apply KS two-sample test using both one-sided alternative hypotheses
    Store the p-values
end for
Apply the BH(0.05) method to the p-values to determine test rejections    ▷ Up to 20
p-values
Output: Visual Representation of Test P-values
```

We made the analysis live by updating the two inserted datasets during the ongoing game, switching from the BH method to the Online BH method. As a result, it retrieves test results up to the last completed five-minute/extra-time window. Examples of the graphical result are presented in Figure 8 and Figure 9.

To put Figure 8 and Figure 9 in context, I analysed the empirical distributions of the absolute acceleration peaks relative to the first two rejections in Figure 8. The graphical representation is in Figure 10. Even though it is not the classical shape expected from

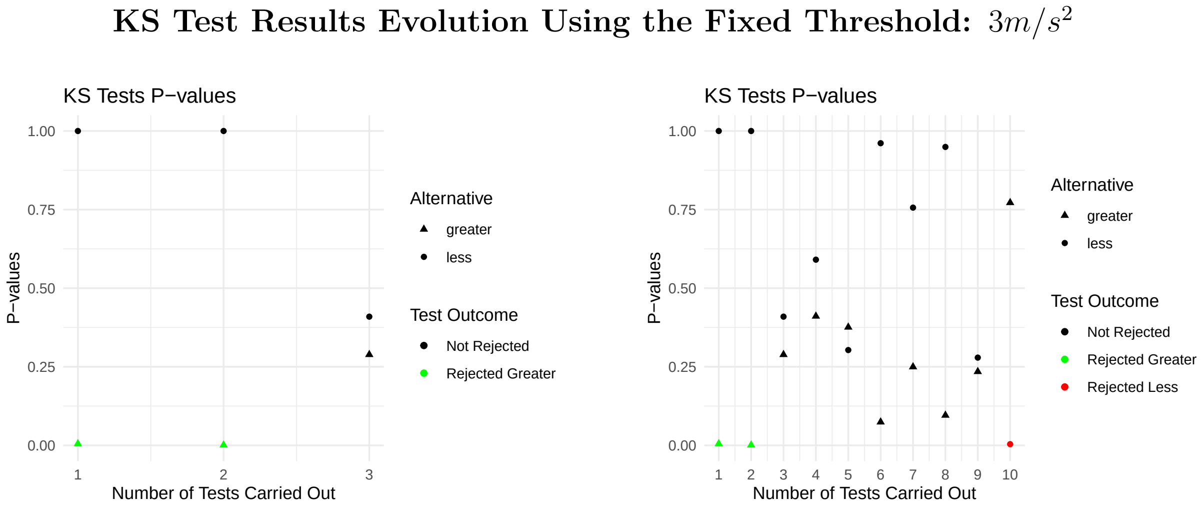


Figure 8: Results of Online Analysis using Fixed Threshold: after 15 minutes (left) vs. end of half (right). They are referred to a right-back during the first half of a Premier League game, the same game as Figure 9.

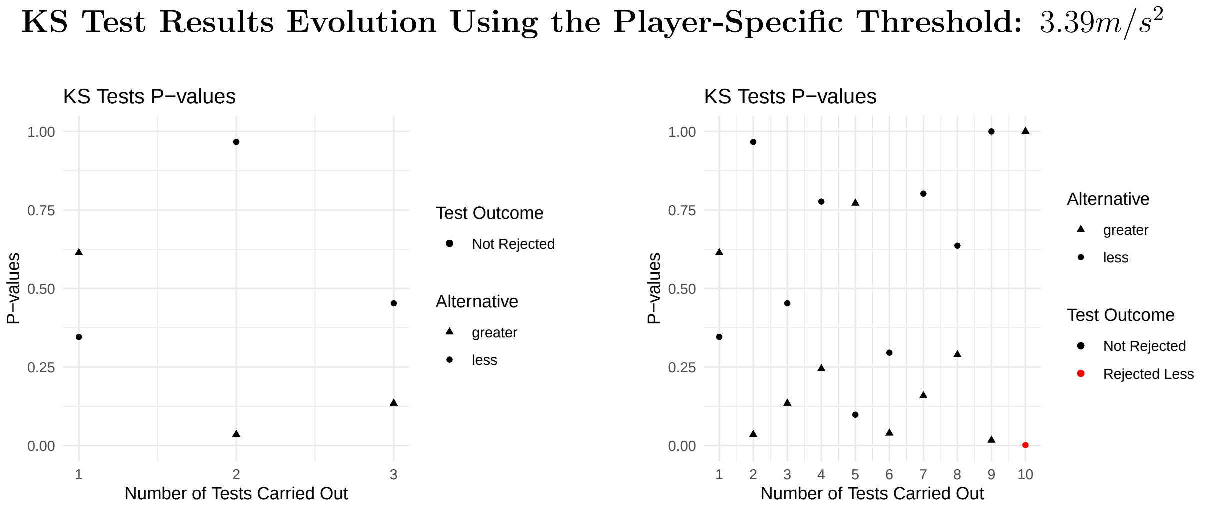


Figure 9: Results of Online Analysis using Player-Specific Threshold: after 15 minutes (left) vs. end of half (right). They are referred to a right-back during the first half of a Premier League game, the same game as Figure 8.

two stochastically different distributions (example in Figure 11), the average of the analysed player is greater in both time windows.

To understand the FDR control level of the methodology, we analysed the dependence structure of the p-values shown in Figure 8 (right). Within the same window, the two p-values are dependent on one another. This has to be attributed to the intrinsic

Densities of the Peaks Using the Fixed Threshold: $3m/s^2$

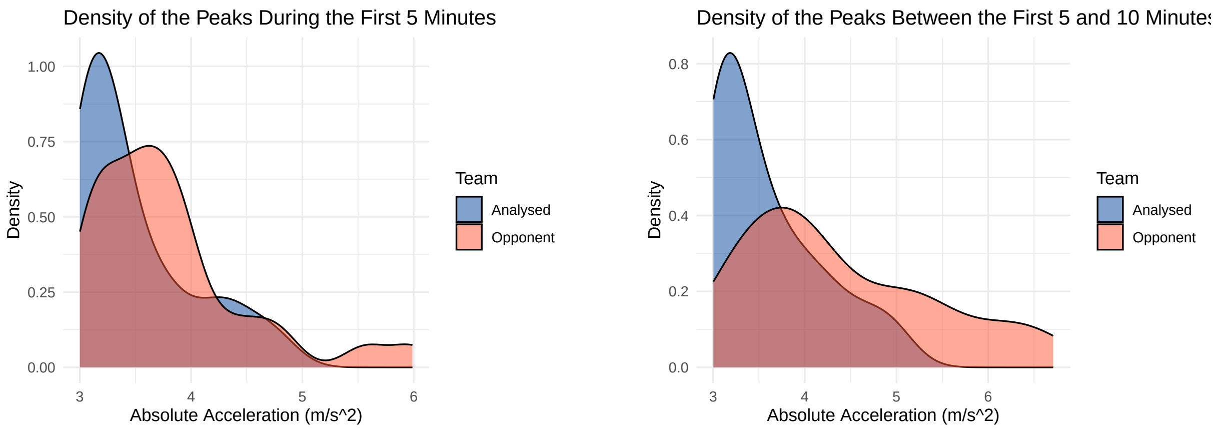


Figure 10: Density plots of the absolute acceleration exceedances of a right-back during the first half of a Premier League game. The threshold used is the Fixed Threshold of $3m/s^2$. The left plot shows the density for the time window of 0-5 minutes, and the right plot shows the density for the 5-10 minutes window.

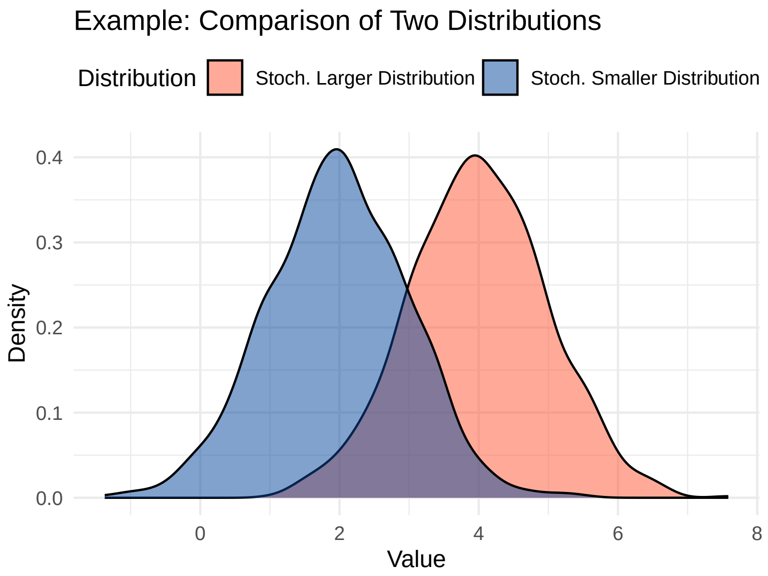


Figure 11: Standard example of two distributions which are stochastically different.

interpretation of the p-values, which is the likelihood of observing as extreme of a result or more under the null hypothesis. When it is very unlikely to observe one distribution as stochastically greater than the other, it would be likely not to observe it and vice versa. Therefore, within the same window, a high p-value in one test implies a low p-value in the other, while a p-value in the middle implies a similar p-value for the other test. Across the different time windows, I analysed each vector of p-values from the two different KS alternatives separately. To perform this analysis, I observed their Autocorrelation Function (ACF), Figure 12 and Figure 13. It can be noticed that at the first lag, there is no significant autocorrelation. It is worth mentioning, though, that all the significance bands are very high in modulus (0.5) due to the small sample size of the p-values vector. Therefore, only an incredibly strong dependence (both positive and negative) could have been classified as significant.

Autocorrelation Function of the P-values Using the Fixed Threshold: $3m/s^2$

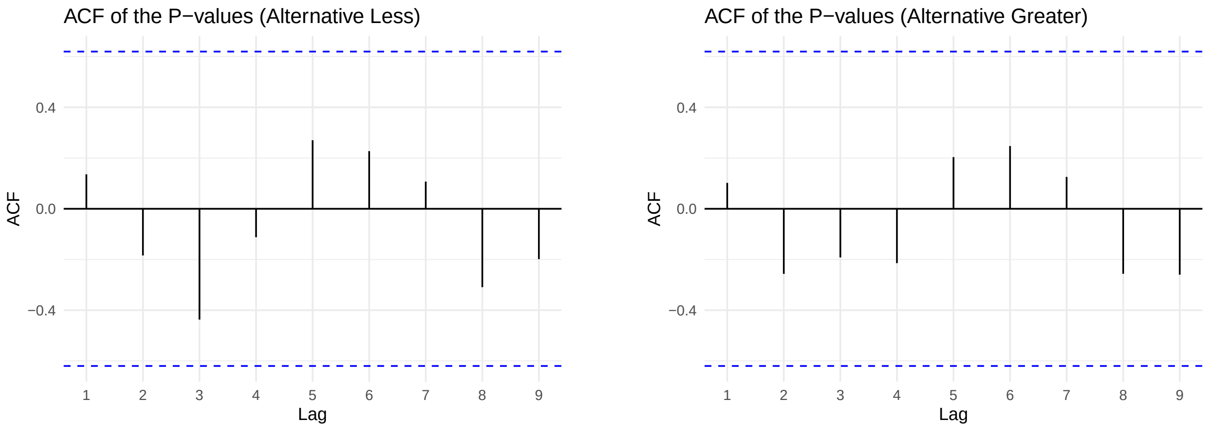


Figure 12: Visualisation of the results from the Online analysis using Fixed Threshold after 15 minutes (left) and after the half was over (right). The analysed player is a right-back during the first half of a Premier League game, the same game as Figure 9.

All in all, considering the possible control level for the FDR, that ranges from 0.05 to 0.18 as mentioned in Subsection 2.5, the Offline method (when having all the p-values at our disposal) is likely to control the FDR at level in the upper quartiles of the range. As discussed in Subsection 2.5.2, the Online BH does not have the theoretical guarantees of the standard BH method. However, from an application point of view, committing a Type I error in-game is not as dangerous as it may be in other fields. This has to do with the fact that even true rejections should not cause an immediate player substitution or an overall tactical change. They should instead be interpreted as a flag, which, coupled with other factors, can give a more complete picture of the game. In fact, true rejections

Autocorrelation Function of the P-values Using the Player-Specific Threshold: $3.39m/s^2$

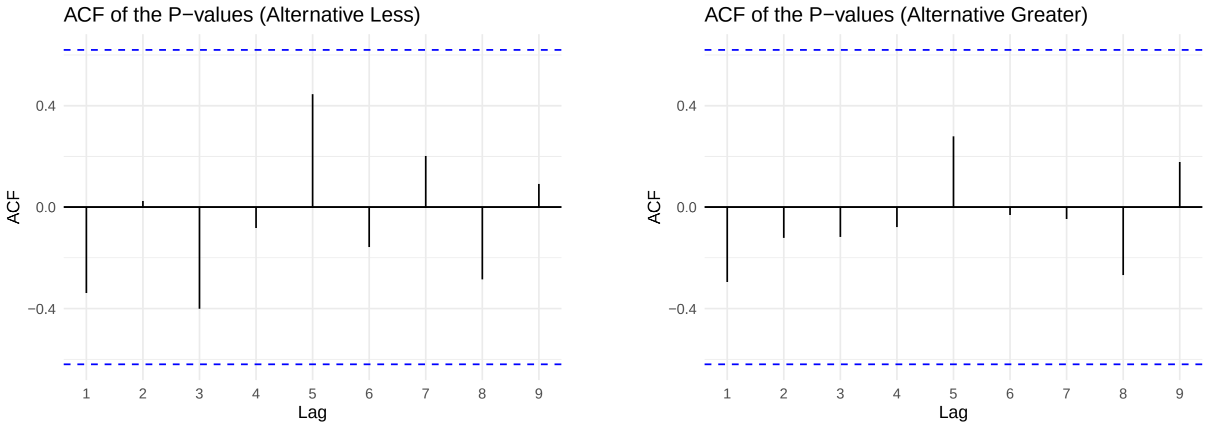


Figure 13: Visualisation of the results from the Online analysis using Player-Specific Threshold after 15 minutes (left) and after the half was over (right). The analysed player is a right-back during the first half of a Premier League game, the same game as Figure 8.

can be caused by factors unrelated to physical performance since football is much more than an acceleration competition. Furthermore, the high frequency of the tests makes the results quickly converge to the standard BH method, as discussed in Subsection 2.5.2.

3.2 Comparison Between the Two Thresholds

The analysis we carried out using two different thresholds yielded some interesting results. To compare the Fixed Threshold (FT) with the Player-Specific Threshold (PST), for every game and each half, we collected the number of tests carried out on all the players in the team and the corresponding number of rejections. Then, we normalised the number of rejections by the number of tests carried out during that time frame, creating a vector of four key numbers per game for both thresholds:

- Normalised rejections during the first half, KS alternative: less;
- Normalised rejections during the first half, KS alternative: greater;
- Normalised rejections during the second half, KS alternative: less;

- Normalised rejections during the second half, KS alternative: greater;

I compared the means of those four critical numbers for FT against PST using all alternatives of a Welch t-test (Welch, 1938), which resulted in a vector of 12 p-values. I applied BH(0.05) to it and displayed the results in Table 1. I also displayed box plots of the normalised rejections for every alternative and every half of the game to better understand their distribution in Figure 14 and Figure 15.

Game Half and KS Alternative Hypothesis	**P-value**	**T-test Alternative Hypothesis**	**BH(0.05) Conclusion**
First Half,	0.0005	PST $\leq$ FT	Rejected
Alternative: Less	0.9995	PST $\geq$ FT	Not Rejected
	0.0010	PST $\neq$ FT	Rejected
First Half,	0.8222	PST $\leq$ FT	Not Rejected
Alternative: Greater	0.1778	PST $\geq$ FT	Not Rejected
	0.3556	PST $\neq$ FT	Not Rejected
Second Half,	0.0040	PST $\leq$ FT	Rejected
Alternative: Less	0.996	PST $\geq$ FT	Not Rejected
	0.0080	PST $\neq$ FT	Rejected
Second Half,	0.9903	PST $\leq$ FT	Not Rejected
Alternative: Greater	0.0098	PST $\geq$ FT	Rejected
	0.0195	PST $\neq$ FT	Rejected

Table 1: Comparison between the p-values and hypotheses between Player Specific Threshold (PST) and Fixed Threshold (FT) key numbers mean. The methods used are the Welch Two Sample t-test to compare the means and the BH(0.05) to draw conclusions.

From the box plots in Figure 14 and Figure 15, we can observe a smaller number of rejections when applying the BH method than when we did not apply it, confirming its theoretical guarantees.

Interestingly enough, both from Figure 14, Figure 15 and from Table 1, it appears that

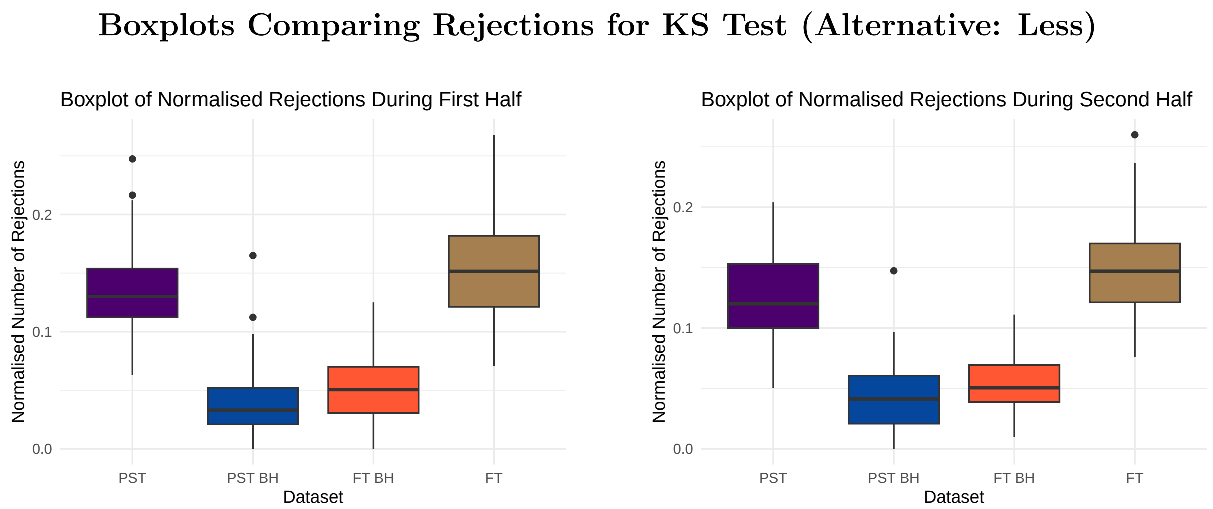


Figure 14: Boxplots comparing the number of rejections, normalised by the number of tests carried out, for KS Test (Alternative: less) across different thresholds with and without applying BH.

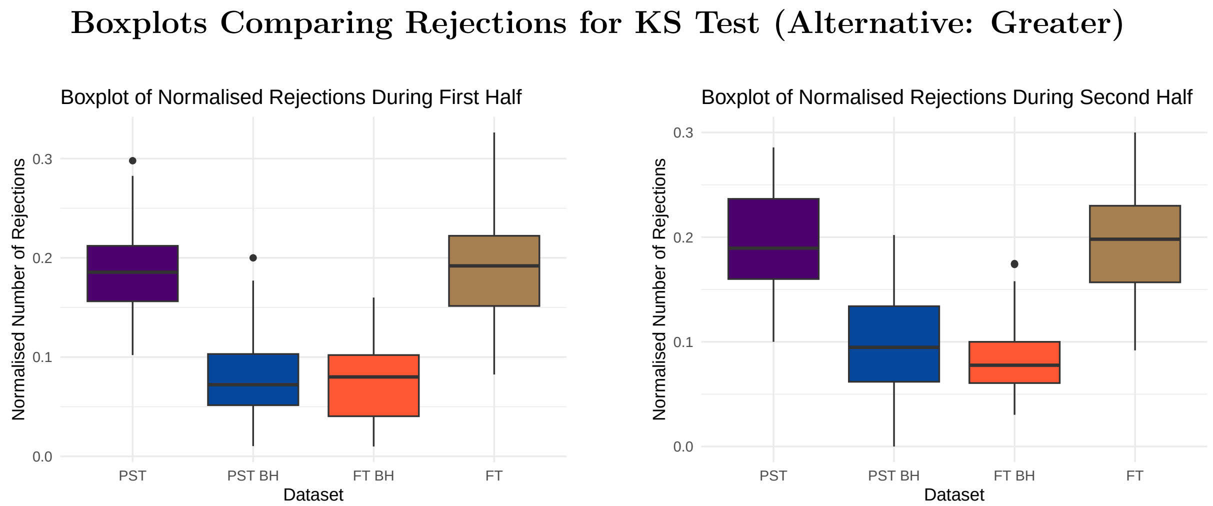


Figure 15: Boxplots comparing the number of rejections, normalised by the number of tests carried out, for KS Test (Alternative: greater) across different thresholds with and without applying BH.

when using the Player Specific Threshold, the tests with the KS alternative hypothesis *less* are on average rejected significantly more often than when using the Fixed Threshold. On the other hand, when considering the KS alternative hypothesis *greater*, there are no significant differences in the average normalised number of rejections. This result is consistent on both halves. A possible cause can be that the PST is generally higher in modulus than the FT, which is set to $3m/s^2$. This makes it more likely for the excesses, and consequently for the exceedances, to follow a Pareto distribution under which, as discussed in Subsection 2.3, the KS test is the most powerful among the nonparametric tests.

The drawback of the PST may be that eventual correlations between the number of rejections and other positive or negative effects (depending on the alternative of the rejected test), may generalise to other teams, making it necessary to run the analysis from scratch again. In this regard, I analysed the correlation between the normalised number of rejections and the game result. I assigned a score to each possible result: 3 for a win, 1 for a draw, and 0 for a loss. I chose this system to mimic the points given in a championship match. This assigning system could be improved to reflect the unique importance of every game better, but it is not necessary for the sake of this argument. Especially considering that the resulting correlation between the normalised number of rejections (both positives and negatives) and the game result is close to zero, Table 2. From the displayed correlations, we can conclude that the normalised number of rejections alone is not enough to predict the result of a game accurately.

Tests	**Correlation with Result**
Sum of Normalised Rejections in the Two Halves with PST, Alternative: Less	0.1218
Sum of Normalised Rejections in the Two Halves with PST, Alternative: Greater	0.0696
Sum of Normalised Rejections in the Two Halves with FT, Alternative: Less	0.0740
Sum of Normalised Rejections in the Two Halves with FT, Alternative: Greater	0.0094

Table 2: Pearson Correlation between Rejected Tests and Points Gained.

4 Discussion

All in all, we started the analysis with the accelerations and the positions of the players of a Premier League team and the teams they were facing, which were recorded at a rate of ten observations per second. We then divided the scope of the analysis into two halves; within each half, we divided the time into 9 five-minute windows of regular time and 1 window of extra time. For each window, we considered only a players couple, the one under analysis from our team and the nearest neighbour in the window frame among the opposing players. We then filtered both their datasets, keeping only the peaks of absolute acceleration that exceeded a threshold that was either fixed at $3m/s^2$ or at the 99^{th} percentile of our team player's absolute acceleration up to the previous game. Afterward, we compared the filtered datasets on each window, testing whether one was stochastically greater than the other using the Kolmogorov-Smirnov two-sample test. Ultimately, we applied the Benjamini and Hochberg method to account for multiple testing and control the overall Type I error before deciding whether to reject a test. We went beyond the post-game analysis, proposing and implementing the in-game version of the methodology, for which we discussed the potential theoretical downsides and what those translate to in a real scenario.

Within the methodology, there are some possible improvements. For instance, choosing carefully a player specific threshold which is low enough to include more than 1% of the data and simultaneously make the excesses follow a Pareto distribution; this would make the results more generalisable while maintaining a high accuracy.

The methodology is not without limitations, mainly because it does not provide an unambiguous result upon which to base decisions. This is due to the fact that acceleration is only one of the many key components that determine the individual performance of a player, not to mention the performance of the overall team. This is reflected in the absence of a positive correlation between a good result in the game and the increasing number of moments in which a significantly greater absolute acceleration intensity could be detected teamwise. The same holds true for a negative correlation between a good result and the number of moments of a teamwise significantly smaller absolute accel-

eration intensity. Therefore, this instrument should not be the only reference for any tactical choice but rather an additional information to be put in context and properly used by the team's coaching staff. To this end, a potential improvement could come from considering other role-specific features suitable for extreme value analysis, like the number of turns for midfielders.
Another big limitation of the methodology is that it has no predicting power. A big improvement could come from developing a clever model to predict absolute acceleration peaks while considering the game's requirements for better prediction accuracy.

Nevertheless, this methodology provides an easily interpretable tool to detect under- or over performances from an acceleration standpoint at a five-minute frequency. The coaching staff can use it during or after the game to have a better overall picture of the performance of each player in the team, revealing insights that are hard to detect just by observing the game. This can lead to a tactical change in the role of a specific player to better suit their characteristics, to a meaningful substitution during a critical moment of the game, or to a quick tactical adjustment to capitalise more on some opponent flaws. Observing the results of teams using this powerful technology would be exciting.

5 Endmatter

The code used for the analysis can be found in a private GitHub repository

The repository is private due to the data's nondisclosure agreement with Sportlight Technology Ltd. Nevertheless, Imperial users can request access via email at francesco.ventura23@imperial.a or ventura.f.2023@gmail.com.

References

Abbott, W., Thomas, C., and Clifford, T. (2023). Effect of playing status and fixture congestion on training load, mental fatigue, and recovery status in premier league academy goalkeepers. *Journal of Strength and Conditioning Research*, 37(2):375–382.

Bampouras, T. M. and Thomas, N. M. (2022). Validation of a lidar-based player tracking system during football-specific tasks. *Sports Engineering*, 25(1):8.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, 29.

Chu, J., Dickin, O., and Nadarajah, S. (2019). A review of goodness of fit tests for pareto distributions. *Journal of Computational and Applied Mathematics*, 361:13–41.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer-Verlag London, 1 edition.

Dalen, T., Ingebrigtsen, J., Ettema, G., Hjelde, G. H., and Wisløff, U. (2016). Player load, acceleration, and deceleration during forty-five competitive matches of elite soccer. *Journal of Strength and Conditioning Research*, 30(2):351–359.

Daykin, C. (2023). Maximum acceleration and deceleration – metric considerations and uses. https://pro.statsports.com/maximum-acceleration-and-deceleration-metric-considerations-and-uses/. Accessed: 2024-08-17.

Deloitte LLP (2024). Deloitte football money league 2024. https://www.deloitte.com/uk/en/services/financial-advisory/analysis/deloitte-football-money-league.html. Published 2024-01-25.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Escuret, H. (2023). Detecting fatigue in professional football. Master's thesis, Imperial College London.

Evans, D. A., Jackson, D. T., Kelly, A. L., Williams, C. A., McAuley, A. B. T., Knapman, H., and Morgan, P. T. (2022). Monitoring postmatch fatigue during a competitive season in elite youth soccer players. *Journal of Athletic Training*, 57(2):184–190.

Ferreira, J. A. and Zwinderman, A. H. (2006). On the benjamini–hochberg method. *The Annals of Statistics*, 34(4).

Harter, H. L. and Owen, D. B. (1970). *Selected tables in mathematical statistics. Sponsored by the Institute of Mathematical Statistics. Edited by H. L. Harter and D. B. Owen.* Markham series in statistics. Markham, Chicago.

Lourenço, J., Élvio Rúbio Gouveia, Sarmento, H., Ihle, A., Ribeiro, T., Henriques, R., Martins, F., França, C., Ferreira, R. M., Fernandes, L., Teques, P., and Duarte, D. (2023). Relationship between objective and subjective fatigue monitoring tests in professional soccer. *International Journal of Environmental Research and Public Health*, 20(2):1539. Epub ahead of print, 2023 Jan 14.

Pickands, J. I. (1975). Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics*, 3(1):119 – 131.

Pratt, J. W. and Gibbons, J. D. (1981). *Kolmogorov-Smirnov Two-Sample Tests*, pages 318–344. Springer New York, New York, NY.

Remitly, Inc. (2024). The most popular sports in the world for summer 2024, ranked by country. https://blog.remitly.com/lifestyle-culture/popular-sports-in-the-world-for-summer/. Published 2024-06-07 - 09:53.

Rhodes, D., Valassakis, S., Eaves, R., Bortnik, L., Harper, D., and Alexander, J. (2021). The effect of high-intensity accelerations and decelerations on match outcome of an elite english league two football team. *International Journal of Environmental Research and Public Health*, 18.

Sky international AG (2024). Real madrid, il preparatore pintus spiega gli allenamenti. video. https://sport.sky.it/calcio/champions-league/2024/04/10/real-madrid-antonio-pintus-allenamento-video. Published 2024-04-10 - 09:31.

Sportlight Technology Ltd (2024). How sports tech evolved from a luxury to a necessary investment. https://www.sportlight.ai/post/how-sports-tech-evolved-from-a-luxury-to-a-necessary-investment.

Synopsys, Inc. (2024). What is lidar? https://www.synopsys.com/glossary/what-is-lidar.html#. Accessed 2024-07-17.

Szubzda, F. and Chlebus, M. (2019). Comparison of block maxima and peaks over threshold value-at-risk models for market risk in various economic conditions. *Central European Economic Journal*, 6(53):70–85.

Thoseby, B., Govus, A. D., Clarke, A. C., Middleton, K. J., and Dascombe, B. J. (2023). Peak match acceleration demands differentiate between elite youth and professional football players. *PLOS ONE*, 18(3):e0277901.

Toni Modric, Sime Versic, D. S. and Liposek, S. (2019). Analysis of the association between running performance and game performance indicators in professional soccer players. *International Journal of Environmental Research and Public Health*, 16(20):4032.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29.

Wilson, J. (2022). Acceleration load, acceleration density and acceleration density index. https://support.catapultsports.com/hc/en-us/articles/360001559976-Acceleration-load-acceleration-density-and-acceleration-density-index. Accessed: 2024-08-17.