Estimating Player Impact over Time:

Hierarchical Models for Scouting & Performance Evaluation in English Football

Tahmeed Tureen <tureen@umich.edu>
Irena Chen <irena@umich.edu>

1 Introduction

Since the introduction of the expected goals (xG) metric in football (soccer), the development and distribution of advanced analytics has changed how the game is played and consumed by supporters, professionals, and clubs [3]. Football broadcasters and pundits now regularly include advanced metrics such as shots on target, goal involvements, and carries in their match analyses [13, 14, 18]. Most of these metrics are probabilistic measures that quantify the chance of an action occurring at any given moment in a football match [3]. For example, the xG models assign a score between zero and one to any observed shot in a match. The philosophy being that: given a set of shot characteristics (i.e. distance and angle), the model estimates a probability of a goal from the observed shot. Different expected models have different sets of predictors depended on the response variable being modeled. As a matter of fact, expected goals can vary depending on the choice of predictor variables as well. Commonly used predictors are shot location, distance to goal, shot angle, type of play, body part used to shoot, shot type, and shot technique [1, 12]. Newer models have also started to incorporate predictors beyond shot characteristics, such as goalkeeper location, defender location, received pass type, and whether the shot-taker is under pressure [6, 12, 16, 19].

These models are typically developed using event-level football data like the datasets which Hudl StatsBomb provide [9, 15]. Event data is collected by tracking players over the course of a football match and logging their actions such as shots, passes, and tackles. There are other types of football data such as freeze frame data, physical/wearable data, tracking/positional data, video broadcast, and match sheet data [8, 20, 21]. In recent years, researchers have been working on creating a common data format to enable users to analyze and build models using the data with ease. With access to such rich data, football researchers and scientists have continually advanced the field of football analytics through books, academic journals, conferences (virtual/physical), and industry panels [1, 12, 15].

Naturally, as an applied science to sport, one of the most interesting aspects of football analytics for both football supporters and professionals is the scouting and evaluation of players. A very nuanced trait of most productionalized expected models is that they do not account for players who take a football action. In statistical terms, there is no player predictor in the model. This

absence of the player in models seems contradictory to the nature of football, where a player's skill influences the success of an action. It is reasonable to assume that a striker would have a higher chance of scoring a goal as opposed to a centre back. Moreover, not all strikers have the same goal-scoring capabilities. However, in an xG model that does not adjust for players, two separate shots that have the same measures for the model predictors will be assigned the exact same xG regardless of who is taking the shot. This is arguably a limitation of current xG models. The inclusion of players in xG models along with their shot characteristics can provide information about the skills of the shot-taker as well. Rather than a general xG measure, an individualized approach can estimate the player's effect on goal-scoring which represents their contribution to the success of a goal. Popular models used to create these event-level models are classification models like logistic regression and extreme gradient boosting (XGBoost). For these models, scientists can use the player as an individual predictor in the model. However, there are issues in doing so: (i) categorical predictors require a dummy/reference level and (ii) the number of unique players can lead to a high cardinality issue and a sparse predictor space. Therefore, neither type of model has the ability to estimate "an impact" of the player on a football action. In football, a key assumption is that players have different skill sets and are different from each other. Any model that does not account for the player, will inherently ignore this assumption and treat every player as the same in the dataset.

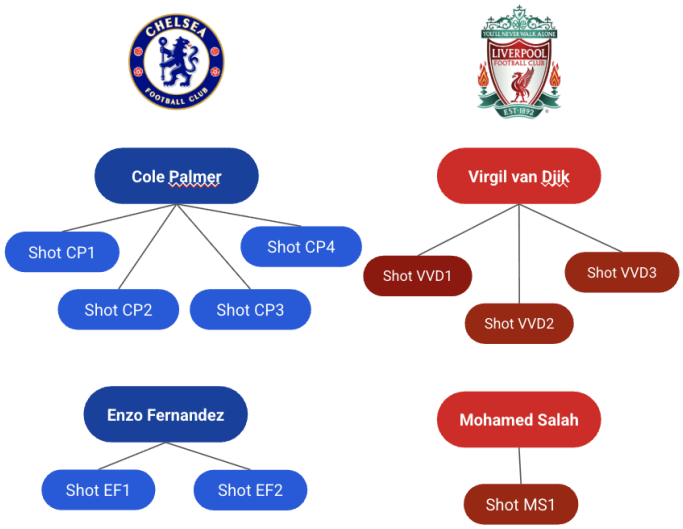


Figure 1. This diagram illustrates the hierarchical structure of event-level data for four different players collected from Chelsea's 3-1 win over Liverpool on May 4th, 2025. In this match, Cole Palmer (CP) took 4 open play shots, Enzo Fernandez (EF) took 2 shots, Virgil van Dijk (VVD) took 3 shots, and Mohamed Salah (MS) only took 1 shot.

Secondly, football data, especially event-level data, has a hierarchical structure. Since players make several passes, engage in a number of tackles, and take multiple shots over the course of a match, they can show up more than once in the data for a single match. All of these events are logged under the players' names. As a result, any given player can be ***repeated*** in the dataset especially when the data history spans multiple seasons. In statistical terms, the events are nested under a player hierarchy in the data. This hierarchy is illustrated in **Figure 1**.

This data hierarchy suggests that the events that are associated with any specific player, are statistically correlated with each other. In other words, two actions taken by the same player are not independent of each other as there is a common link between the two actions. This is because individual subjects in data have idiosyncrasies that are latent and unmeasured by the data. Therefore, this repeated measures structure violates one of the fundamental assumptions of most statistical and machine learning models: "all observations are independent of each other". Popular probabilistic models used in football analytics like logistic regression and tree-based boosting models do not account for this within-player correlation. However, the application of these models to repeated measures can lead to improper inferences [4]. If the primary aim of a model is prediction, this data hierarchy has a miniscule impact when robust non-parametric models like XGBoost are used. However, if the analysis revolves around player scouting and evaluation, then it is reasonable to argue that appropriate statistical inference should also be a priority.

Tureen & Olthof (2022) unpacked this phenomenon in even greater detail emphasizing how it provides researchers a unique opportunity to develop player-adjusted models. They demonstrated the use of hierarchical statistical models to account for the repeated measures and player idiosyncrasies in event data using a multi-level parameter modeling framework. The models are called hierarchical or multilevel models because (i) the model is fit on data that has a hierarchy and (ii) the model itself has a hierarchy in terms of its parameters [17]. The particular subset of hierarchical models used in their research are called Generalized Linear Mixed Models (GLMM). They created player-adjusted xG models for both the men's and women's game and demonstrated how they can be used to derive player impact scores coined as **"Estimated Player Impact" (EPI)**. Additionally, their models provide interpretable predictor effects on xG with statistically appropriate confidence intervals. This type of analysis is not possible for tree-based models as they are black-box models [7]. As for logistic regression or single-level generalized linear models, the standard errors for predictor effects are biased because they do not account for the within-player correlation [4, 17]. An obvious extension of Tureen & Olthof's paper is to incorporate multiple seasons of data to longitudinally analyze players **over time**. Time in itself can add another level of hierarchy to the data where each season might have an impact on how individual players perform. Many factors such as incoming/outgoing transfers at a club, managerial and structural changes, and age can alter the expected performance of a player. One can also argue that certain players may actually be robust to these changes and their performances may not change all that much. An anecdotal example would be how Jose Mourinho's core players from his first tenure at Chelsea (i.e. Frank Lampard, John Terry, Didier Drogba, and Petr Cech) continued to perform at the top level regardless of the constant change in managers at the club.

For the purposes of this research, the original work by Tureen & Olthof is extended to a longitudinal study by incorporating five seasons' worth of English Premier League data from Hudl StatsBomb and by re-formulating the modeling framework. Secondly, this research paper demonstrates how hierarchical models can be applied to other football actions such as progressive carries and passes into the attacking penalty box. Finally, the statistical models are applied to a hypothetical football

case study to illustrate how statistical inference can be used to supplement player scouting and evaluation.

2 Research Aims

2.1 Data

Hudl StatsBomb event & 360 freeze frame data are used in this research. The datasets contain information from 1,900 matches spanning five seasons of the English Premier League from the 2020/21 season to the 2024/25 season. The events in the data refer to football actions such as shots, passes, and dribbles etc. Each event is labeled with different characteristics of a particular football action, including the player, their position on the field at the time of the event (in x-y coordinates) and granular information about the action itself. The 360 data contains the locations of every player in the freeze frame taken during the logging of a football event. **Table 1** illustrates how the data looks using dummy measures. In this example, Enzo Fernandez has two shots from the same match. The shot characteristics, however, differ for each shot. A single premier league season has 380 total matches where a player can play at most 38 matches. Therefore, the sheer volume of data points for any given player can explode very quickly depending on how involved they are in their respective matches. In this research, there are five seasons worth of data creating an additional layer of hierarchy where the matches themselves are nested under a season. A detailed illustration of the player level hierarchy is provided in **Figure 2**. The hierarchy in the figure can be expanded based on the levels available for study such as the match, the matchweek, the football club, and the season etc.

Table 1. Snapshot of hierarchical data structure of shot event data.

Event ID	Match ID	Season	Player	...	Location	Body Part	Shot Outcome
abc1	cheliv25	24/25	E. Fernandez	...	(x, y)	Right Foot	Goal
hjq2	cheliv25	24/25	E. Fernandez	...	(x, y)	Head	No Goal
...	...	...	...	...	...	...	...
jyt242	livcry25	24/25	C. Gakpo	...	(x, y)	Head	No Goal
cxy190	livcry25	24/25	M. Salah	...	(x, y)	Left Foot	Goal

2.2 Motivation

The models developed by Tureen et al. demonstrate how GLMMs can be carefully applied to football data to draw inferences on player impacts on goal scoring opportunities as well as deriving associations between the predictor and response variables. For example, the paper identified that Heung-min Son (Tottenham Hotspur FC) & Vivianne Miedema (Arsenal Women's FC) were the best ranked players in terms of player impact on goal scoring. In terms of predictor-response associations, they statistically quantified that a defender pressuring a shot-taker can reduce their

odds of scoring a goal by approximately 15%, on average in the Premier League. On the other side of things, they derived that when a player opens up the angle of their shot by roughly 16 degrees they increase the odds of a goal by 57%, on average in the Premier League. Their study samples are from two seasons of the Premier League and three seasons of the Women's Super League (WSL). As a consequence, the inferences drawn from their models are contextualized within those seasons only. These are, however, interesting findings that allow football practitioners to better understand the empirical analysis of how goals have been scored in the Premier League and WSL. Some of the findings are very intuitive such as the defender pressure example. But, an empirical number allows a football practitioner, such as a defensive coach in the youth academy, to emphasize how important it is for a defender to make sure to pressure a shot-taker.
The original work, however, did not demonstrate how extending the parameters in the model allows the researchers to also assess the longitudinal change in player impact from one season to the next. The takeaways were provided from a cross-sectional perspective where they did not account for the fact that some players had played multiple seasons of matches. Additionally, their work only focuses on the build of an xG model. In the discussion section of their paper, the authors emphasize that event level data enables researchers to expand their work to other types of football actions such as passes, crosses, and tackles.

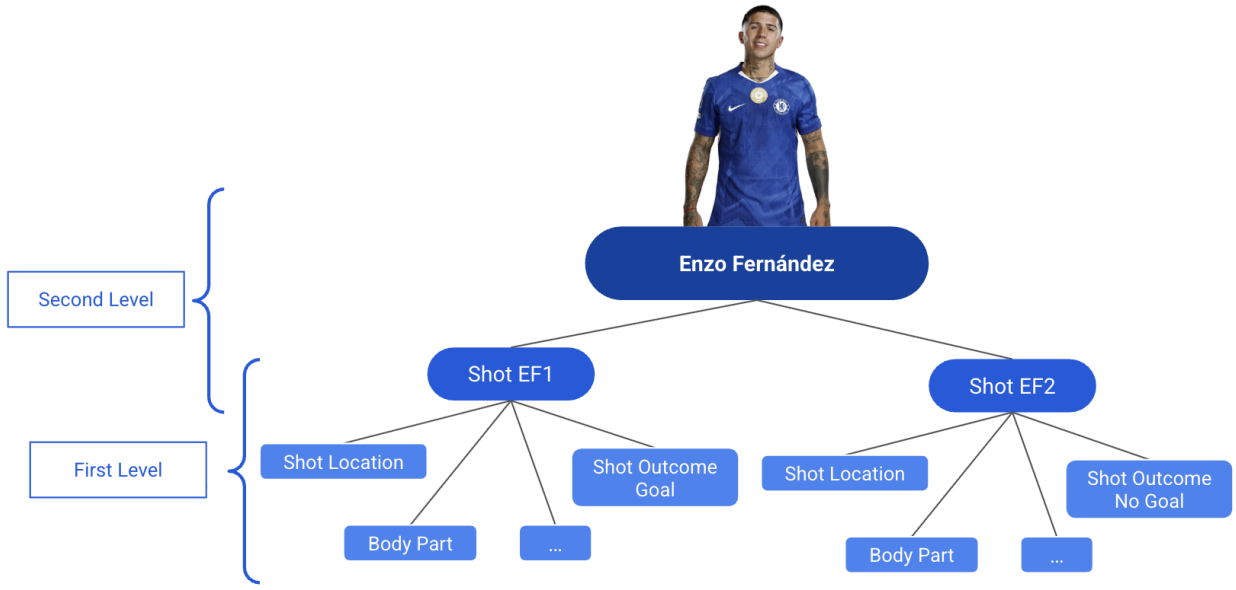


Figure 2. Illustration of hierarchical data structure for shot event data using the player as the highest level. The hierarchy in the figure can be expanded based on the levels available for study such as the match, the matchweek, the football club, and the season etc.

2.3 Aims

Motivated by the work of Tureen & Olthof and the access to five seasons worth of data, the aims of this research are to: (i) build GLMMs that allow for the longitudinal analysis of football players, (ii) apply the hierarchical model framework to two additional football actions of interest, (iii) estimate the player impact on football actions of interest and allow for player-adjustment, and (iv) draw generalizable inferences about the relationships between the predictor and response variables in a

football context. The focal point of this research is to re-iterate the capabilities of hierarchical models as well as demonstrate to the analytics community how to design a statistically sound longitudinal study using event-level data. Unlike the original paper, this paper does not investigate the predictive capabilities of hierarchical models in football.

---

3 Methodology

Hierarchical models can be approached from either a frequentist or a Bayesian perspective [4, 11]. The original paper used a frequentist framework in their model design; therefore, this research takes a frequentist approach as well. However, we discuss the powerful capabilities of the Bayesian framework in the discussion section of this paper to motivate future research related to football hierarchical models. The hierarchical models are built on three types of events: (i) shots, (ii) carries into the penalty box, and (iii) passes into the penalty box.

3.1 Shots for Expected Goals (xG)

Naturally, any model that estimates the probability of a goal from a set of shot characteristics is an expected goals (xG) model. The response or target variable is binarized into two categories (1: Goal, 0: No Goal). The predictors used in our xG model are as follows:

(i) **Three continuous predictors:** distance to goal, angle of shot, season[1] since the players' data debut
(ii) **Two binary predictors:** presence of goalkeeper in the shot triangle, whether shot-taker is under pressure
(iii) **One multi-class predictors:** body part with which the shot was taken
(iv) **One count predictor:** count of defenders in the shot triangle,

The shot characteristics and their summary statistics are provided in **Table 2** in **Section 5**. The predictor variables were selected based on literature review and investigation of multi-collinearity. The study sample is restricted to only shots that did not result from a direct penalty or free kick. The models are also stratified such that one model is built using only forwards (strikers and wingers) and another model that includes all position types. The season predictor is included to adjust for any population-level relationship with the target: "does playing in the league for more seasons increase goal scoring odds, on average?". Additionally, it is statistically inappropriate to exclude the time predictor when performing a longitudinal data analysis.

3.2 Progressive Carries into the Box

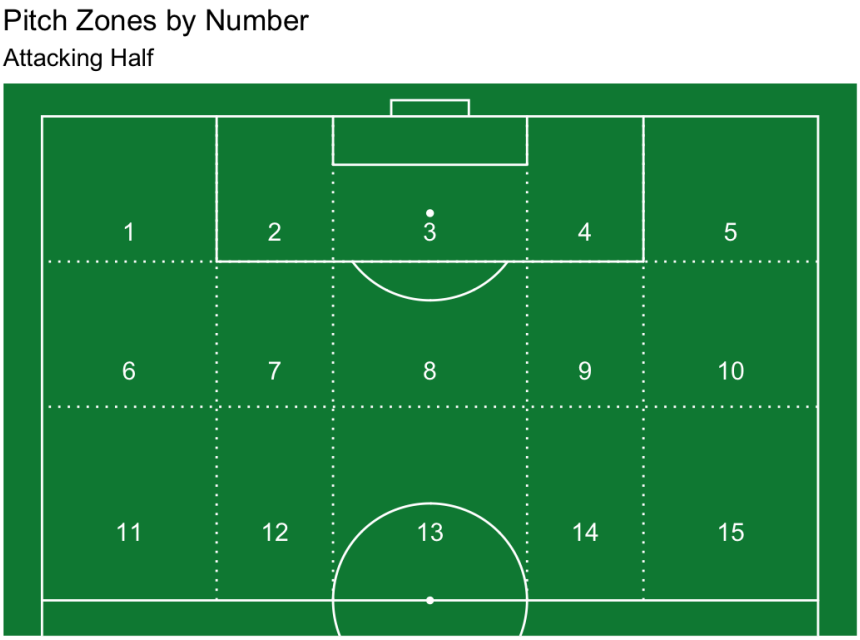
In football analytics, a "carry" refers to the action where a player moves the ball from one location to another while maintaining control/possession of it. Hudl StatsBomb defines it as an action where a

[1] season is converted to a numeric variable that represents the number of full seasons a player has played in the premier league according to the data. A value of 0 represents their first season in the data, a value of 2 represents their third season in the data. Not all players have their real life debut season in the data.

player moves the ball at least 5 metres while maintaining possession. In addition, the carry ends with another on-ball action such as a pass, shot, dribble, or loss of possession etc. A carry is different from a dribble in the sense that a dribble involves the player attempting to beat an opposition player in a clear one-versus-one.

For this paper, we define the response variable as a binary variable that reflects whether it was a successful or unsuccessful carry ***into the opposition penalty box***. A carry is defined as a "success" when it results in a shot on target, completed dribble, completed pass, or a won foul. The response is defined as an "unsuccessful" carry when it results in a shot off target, blocked shot, failed dribble, failed pass, turnover, or a committed foul. The model predictors are relatively straightforward: (i) duration of carry (in seconds), (ii) distance of carry, (iii) whether the ball carrier was under pressure by a defender during the action, (iv) whether the ball carrier was under pressure right after the carry, (v) whether the carry started from a pass-reception or (vi) after a dribble by the player themselves, and (vii) season[1]. The final predictor is the season predictor as described in **Section 3.1**. The study sample is limited to carries that start from the attacking half excluding the penalty box: carries initiated from zones 2, 3, and 4 in **Figure 3** are excluded. The output of this model is interpreted as the estimated probability that a carry results in a successful action.

Figure 3. Visual of the attacking half of a football pitch divided into numbered zones



3.3 Passes into the Box

Expected passes (xP) are an interesting suite of models that quantify the probability of a pass completion. As one can imagine, pass probabilities can change dramatically based on the area of the pitch a player makes the pass as well as where the pass is intended to go. In this research, we focus on two types of "**attacking"** passes: (i) passes into the box from zones 1, 5, 6, and 10 and (ii) passes into the box from zones 7, 8, 9, and 11 through 15 (see **Figure 4**). Zones 12-14 are technically

not considered "wings", however, we include them for the convenience of analysis. The models are stratified based on the pass types to delineate the inferences made about them. The first type of pass reflects passes that are driven into the box from the width of the pitch and the second type of pass reflects passes that are pinged into the box from the midfield. This stratification is applied because we hypothesize that there is a statistical difference (particularly in magnitude) in how the pass outcome is influenced due to inherent differences in pitch zones. This hypothesis is tested using the hierarchical models. StatsBomb data binarizes the pass events as either "Complete" or "Incomplete". The output of this model is interpreted as the estimated probability that a pass results in a completion (i.e. reception by a teammate in the penalty box). The predictors for the two pass models are summarized in **Tables 5**.

3.4 Generalized Linear Mixed Models

All three types of response variables in this research paper are binary variables. As a result, the appropriate model choice is the multi-level logistic regression. The formal name for this model is Generalized Linear Mixed Model (GLMM) with binary outcomes. This model linearly learns the relationship between a set of predictors and the target. A trained model can then be used to calculate probabilities based on its predictors. To describe the formulation of the model, we will use the xG model as an example. Equation 1 illustrates the GLMM framework used by Tureen & Olthof:

$$g(\pi_{[ij]}) = \beta_0 + \beta_1 x_{[ij],1} + \beta_2 x_{[ij],2} + \ldots + \beta_p x_{[ij],p} + \delta_{0j}$$

Equation 1. Generalized linear mixed model with a random intercept specification on players indexed by j

In this equation, the i refers to the i-th shot in the study sample. The g() represents the logit link function and the π term is the odds of a goal for the i-th shot. The choice of the logit function is what makes this GLMM a multi-level logistic regression. The x_i terms are the model predictors and the associated β terms represent the predictor effects. To appropriately adjust for the players in the data, a parameter for the players can be incorporated to create a Generalized Linear Mixed Model (GLMM). The j index refers to the j-th player in the data. The δ term is a statistical parameter that represents the random effect associated with the j-th player. As the model is trained, it computes the values of the parameters in the equation including the δ . As a result, each player in the dataset will have their own estimated δ measure. This implies that each *j* player will have their own unique intercept or baseline ($\beta_0 + \delta_j$) for a shot that they take. The δ parameter can be included in the framework because the data has a hierarchy. With access to five seasons worth of data, we can account for player-specific changes over time. To do this, the original model specification can be extended by including the season as a fixed effect predictor and including a **random slope** on the season. **Equation 2** illustrates the new GLMM framework:

$$g(\pi_{[ij]}) = \beta_0 + \beta_s season_{[ij]} + \beta_1 x_{[ij],1} + \beta_2 x_{[ij],2} + ... + \beta_p x_{[ij],p} + \delta_{0j} + \delta_{1j} season_{[ij]}$$

Equation 2. Generalized linear mixed model with a random intercept-slope specification where the random intercept is on the j-th player and the random slope is on season played by the j-th player

The δ_{ij} parameter represents the random slope. The season is included as a fixed effect predictor to ensure that a population-level temporal trend can be estimated. Whereas, the random slope enables the estimation of individual level deviations from the average trend. Not all players have all five seasons worth of data. So, the season predictor is re-coded such that each players' first season (in the data) is treated as their respective baseline season. It is important to finally note that β_0 represents the model intercept or baseline. This is the estimated model output when all predictors are at their baseline level and the random effect and slopes are zero. The zero value for the random effect represents the average player and for the random slope represents the average change since their debut season. The GLMMs also estimate uncertainty measures for the predictor effects in the form of 95% Wald Confidence Intervals (CIs) and the standard deviations for the random effects and slopes.

3.5 Estimated Player Impact

From a statistical perspective, the random effect can be interpreted as the baseline change on the target variable that is attributable to the players: this change is unmeasured by the rest of the predictors in the model. However, explained in football terminology, the measure can be used and interpreted as the "estimated player impact" (EPI) on the outcome of a football action. Consider the xG model as an example. The unique skills from each player can now be statistically estimated by deriving their δ which quantify the effects of players on xG. The δ is a continuous measure and is assumed to follow a Gaussian distribution [10]. This implies that certain players have positive effects on the xG while others have negative effects. A positive value for a player would imply that they increase the xG value in the estimation. Whereas, a negative value for a player would decrease the xG estimation. Any player with an EPI value that is either zero or close to zero is considered the "average" player in the study sample. It is important to note, however, that because these GLMMs are fit using a logit link, these effects actually have a multiplicative impact on the xG scale. The player impacts derived in the models are only comparable between the players in their respective analytical samples. This suggests that the random effects from a forward only model is not comparable to a centre-back only model. The EPI (or δ) can be interpreted in a similar fashion when it comes to the progressive carries and passes models.

3.6 Random Slope

From a football perspective, the random slope tells us how a player's performance changes from season to season relative to the league as a whole. A positive random slope value reflects that a player's impact of achieving a successful football outcome increases over seasons at a **faster** rate than the league average. In simpler terms, these players have empirically demonstrated growth. In the xG context, we can infer that a player has become "better", not necessarily efficient, at scoring. Conversely, a negative random slope value reflects a declining trend over time suggesting that the

player's performance grows slowly, flattens, or decreases relative to the average trend. This suggests that these players are not keeping up with the rest of the league. However, it is important to interpret these slopes in conjunction with the random intercepts. The random intercept reflects the players' baseline level for EPI. Some of these players may begin their first season at a much higher baseline (e.g. an established elite striker) while others start lower but improve faster over the seasons (e.g. academy player). Collectively, the intercept and slope describe whether a player not only begins above or below the population average, but also whether they continue to deviate away from the rest of the league or converge toward the average player over time. A high EPI player with a "small" yet negative slope can still be viewed as a high performing player.

3.7 Limitations & Work-Arounds

Although random slope models are theoretically justifiable when working with large volumes of longitudinal data, random-slope specifications remain relatively uncommon in applied research due to their computational demands, convergence challenges, and substantial data requirements [2]. Barr et al. [2] emphasize the theoretical value of maximal random-effects structures for valid inference. However, later methodological work has demonstrated that such models are often difficult to implement in observational performance settings [22, 23]. These limitations are particularly relevant in football analytics, where player responsibilities can change drastically due to tactical changes by a manager, transfer of the player themselves, or the arrival of new players at the club.

Despite the access to five seasons' worth of event data, football data are often highly imbalanced. In the context of shots, forwards have more shot data on average compared to defenders and defensive midfielders. As a result, a large subset of players may not have enough information in their repeated measures for slope estimation. Additionally, in an ever-changing game like football, there exists structural shifts in season to season trends. These temporal dynamics can induce instability in slope estimation and, in some cases, lead to convergence issues even with extensive data coverage.

Given these limitations, we propose a two-stage modeling approach. First we attempt to fit a random intercept and random slope model where feasible. When intercept-slope models fail to converge or produce variance estimates that are unstable, we provide a pragmatic work-around. We stratify the models by season and compare the estimated player impacts across seasons. This sensitivity analysis assesses whether the individual players consistently deviate from the population baseline. Players that consistently have a positive or negative intercept for all of their seasons can be viewed as players who have sustained the same type of performance in the respective season. Conversely, players with very different EPIs can be viewed as players who have demonstrated sporadic performances in their time in the English Premier League.

This season-stratified approach is less statistically powerful because it effectively breaks a single longitudinal dataset into several independent sub datasets. This reduces the sample size for each model and restricts the models from learning about trends across seasons. In a full mixed-effects

framework, random slopes and intercepts are estimated simultaneously using partial pooling, allowing shared information across players and time periods to improve parameter stability and reduce estimation variance (Gelman & Hill, 2007). Furthermore, the stratified approach will innately treat each season as independent rather than recognizing potential continuity in player performance across seasons. While this stratified approach is **less statistically powerful** than the intercept-slope specification and cannot directly model individual trends over time, it still offers a robust empirical solution to enable player scouting and evaluation

---

4 Data Analysis & Engineering

Prior to model development, multiple data processing and filtration steps are taken. All of the target variables were re-coded such that a successful outcome was represented by a numeric "1" and an unsuccessful outcome was represented by a "0". For each player, we estimated their preferred foot as well as most common position using pass event data. The frequency of passes taken by both feet are measured and the foot that led to more passes is assigned as the player's preferred foot. The same approach is taken when assigning a player's most common playing position. However, the position groups were simplified to four outfield categories: (i) Forwards which include strikers and wingers, (ii) Midfielders, (iii) Centre Backs, and (iv) Fullbacks/Wingbacks. Additionally, human intervention was required for certain edge cases where wingbacks were assigned as forwards. These players were re-categorized into the "Fullback/Wingback" category. For all three types of models, the continuous predictors are centered by the sample mean and scaled by the sample standard deviation. The multi-class predictors were dummy-encoded such that one of the classes is the reference level. The binary predictors are one-hot encoded and the count variables are treated as is unless otherwise stated.

4.1 Shots

The response variable for the xG model is whether a shot results in a goal. Penalty and direct free-kicks are dropped from the analysis because they represent a direct shot from a dead-ball situation. This resulted in 47,076 unique shots from a total of 985 players across five seasons. Because we are building longitudinal random effect-slope models, we add additional filters. Players who did not play for more than one season are dropped from the analysis. Additionally, players who had zero goals and less than 25 shots across all of their seasons were dropped from the analysis. These ad-hoc steps are added to ensure there is at least some variability within each players' data points. This led to a study sample of 39,859 shots from 403 unique players and 26 clubs. Players from Norwich City were dropped as they did not fall into these categories. For the forwards only xG model, the study sample consisted of 20,576 shots from 153 forwards and 24 clubs. Luton Town, Norwich City, and West Bromwich Albion forwards did not pass the filtration process.

Exploratory analysis of the study sample shows that forwards, on average, take more shots as opposed to midfielders and defenders (see **Figure A1** in appendix). This phenomenon is unsurprising; however, it gives us the opportunity to build a forwards only model without

convergence issues. Due to the lack of variability in response variables as well as low sample sizes for individual defenders and midfielders, stratified models for these position groups lead to convergence issues. Refer to the appendix for additional visuals of the distribution of sample sizes for players by position groups.

4.2 Progressive Carries into the Box

As mentioned in the 3.2, this paper also focuses on progressive carries into the penalty box originating from all of the zones excluding 2, 3, and 4 in **Figure 3**. If a carry originated on the border of two zones, the carry was assigned to the zone with the lower number. Hudl StatsBomb data provides information on which related events occur before and after the carry action. For each carry, we identified how the carry began: dribble, ball receipt, or neither. We also used the related events to identify whether the player was under pressure by the defender during and after the carry action. Finally, the related events structure of the data allowed us to categorize the outcome of the carry into a successful or unsuccessful carry using the rules defined in section 3.2. Similar to the shots data filtering process, we filtered for players who have at least one season's worth of data, at least 30 carries, and at least one successful carry across all of their seasons. This led to a study sample of 17,501 carries from 179 unique players and 26 clubs.

4.3 Passes

Similar to the carries variable, the progressive passes in the box are filtered for attempted passes into the penalty that originated from the attacking half. Any pass originating from the penalty box was dropped. Additionally, passes from zones 7 to 9 and 11 to 15 are categorized as **"attacking" passes from midfield"** and passes from zones 1, 5, 6, and 9 are categorized as **"attacking" passes from the wing**. Direct passes from dead-ball situations such as corners, free kicks, throw-ins, and goal kicks were dropped. Hudl StatsBomb 360 Freeze Frame data was also appended to the pass data and any observation missing freeze frame information was dropped from the analysis. Players with less than 2 seasons' worth of data were dropped. Any player that did not have at least one completed pass and at least 50 attempted passes were also dropped. The final study sample consisted of 21,261 midfield attacking passes from 198 players and 27,598 wing attacking passes from 210 players for all 27 clubs in the data [**CIT**].

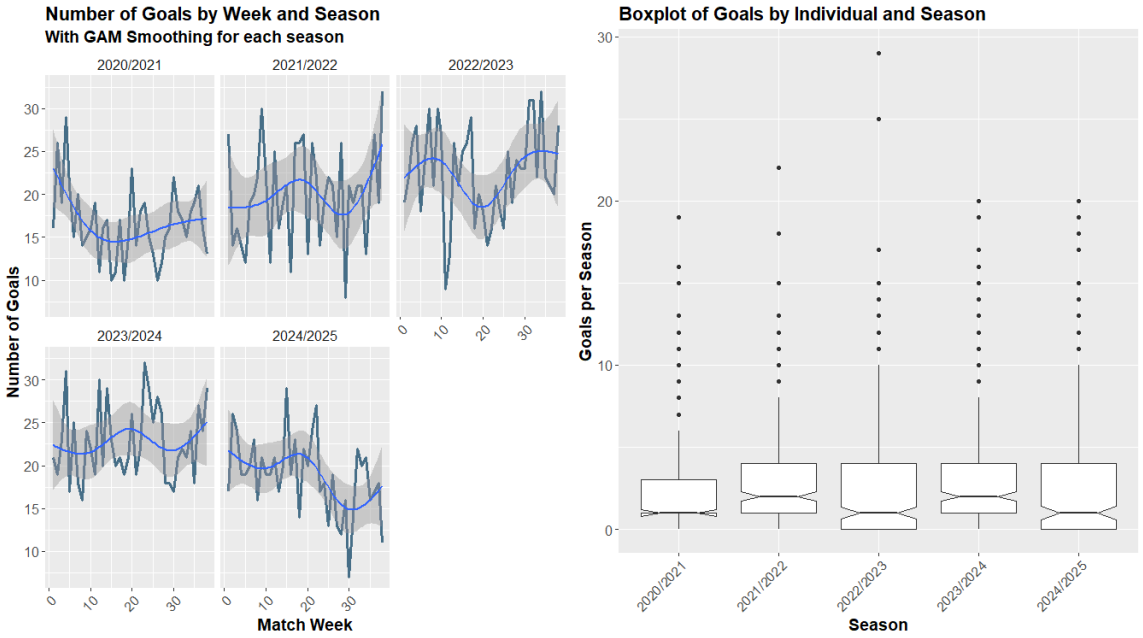
---

5 Descriptive Statistics

In this section, descriptive statistics for the study samples for analyses and model building are provided. Before discussing the actual study samples, we use goals scored in the Premier League to demonstrate that there are trends in the league that can be modeled and that there are players who perform at a level above the average trend. In **Figure 4,** the plot on the left shows the number of goals scored across all matches for each season, with Generalized Additive Model (GAM) smoothing to better show the overall trend across matches. There are clear fluctuations in the number of goals scored over the course of a given season. The right plot of Figure 4 shows the same data, but in boxplot form. We can see that there are several individual players who score far

above the average amount of goals each season.

Figure 4: On the left, the plot displays total goals scored per match week across all five available EPL seasons. On the right, the boxplot displays the total number of goals by each individual in the dataset across all five seasons.



5.1 Shots

The descriptive statistics for the predictors used in the xG models are provided in **Table 2.** A few interesting statistics to note are: (i) approximately 30% of shots are under pressure by an opposition player, (ii) the frequency of shots taken by the head and the less preferred foot are roughly the same per percentages, and (iii) the average number of seasons players have spent in the Premier League in this sample are ~1.48 seasons.

Table 2. Descriptive statistics for the expected goals (xG) models. The mean (standard deviation) are provided for continuous variables and count (percentage) are provided for categorical/binary variables **(continued onto next page)**

Predictor	Type	Statistic	
Body Part Preferred Foot Other Foot Head Other	Categorical	 25,337 (63.58%) 7,357 (18.46%) 6,997 (17.56%) 159 (0.4%)	 13,098 (63.68%) 4,209 (20.46%) 3,183 (15.48%) 78 (0.38%)
Defenders In Shot Triangle	Count	1.09 (1.11)	0.85 (0.96)
Distance To Goal (SB^2 units)	Continuous	17.31 (7.48)	16.3 (6.81)

[2]StatsBomb pitch units

Predictor	Type	Full Model	Forwards Only
Goalkeeper In ShotCone	Binary	38,593 (96.85)	19,736 (95.95%)
Shot Angle (radians)	Continuous	27.07 (15.89)	27.75 (16.17)
Under Pressure	Binary	11,926 (29.93)	6,136 (29.83%)
Season	Continuous	1.48 (1.27)	1.44 (1.28)

When comparing the descriptive statistics between the full and forwards only models, we find some interesting differences: (i) the difference in other foot shots and headers is ~5%, (ii) the average shot distance decreases by roughly 1% in this analysis but does not change by a lot for angle, and (iii) forwards take shots with fewer average defenders in front of them compared to all players (including forwards). These findings in themselves can be converted into research questions related to model stratifications.

5.2 Carries

The descriptive statistics for the progressive carries model are summarized in **Table 3**. Some interesting discoveries to note are: (i) the standard deviation for carry distance is roughly 10.12 StatsBomb pitch units, suggesting there is a lot of variability in distances, (ii) pressure is higher during the carry than after the carry, on average, and (iii) most of the carries start after a ball receipt rather than dribbles. Other types of events right before a carry are balls won from an interception, a ball recovery, and a 50-50 battle for the ball. These account for the remaining percentage for what happens to the carrier right before the carry event.

Table 3. Descriptive statistics for the progressive carries model. The mean (standard deviation) are provided for continuous variables and count (percentage) are provided for categorical/binary variables.

Predictor	Type	Midfield Statistic
Duration (seconds)	Binary	0.67 (0.79)
Carry Distance (SB^3 units)	Continuous	16.49 (10.12)
Under Pressure During Carry After Carry	Continuous	 9,564 (54.65) 6,591 (37.66)
Event before Carry Ball Receipt Dribble	Continuous	 14,149 (80.85) 1,580 (9.03)
Season	Continuous	1.57 (1.3)

[3]StatsBomb pitch units

5.3 Passes

The descriptive statistics for both types of attacking passes are provided in **Table 4**. Interesting patterns that we would like to note are as follows: (i) passers from midfield are under pressure more than the wing (12.73% versus 8.3%), (ii) the variability (per standard deviation) is much higher for wing passes, (iii) ground and high passes from midfield are evenly split in terms of percentages, but high passes account for the majority when they are from the wings. The third note is a finding that should be intuitive to any football fan as we usually see crosses from the width. However, we can see that roughly 27% of the passes are on the ground which begs the question of is there a significant difference when it comes to their effects on a successful pass.

Table 4. Descriptive statistics for attacking passes models. The mean (standard deviation) are provided for continuous variables and count (percentage) are provided for categorical/binary variables.

Predictor	Type	Midfield Statistic	Wing Statistic
Under Pressure	Binary	2,707 (12.73%)	2,292 (8.3%)
Pass Length (yards)	Continuous	25.77 (13.09)	27.27 (10.87)
Pass Angle (radians)	Continuous	-0.03 (0.66)	-0.04 (1.33)
Distance to Closest Defender (SB[4] units)	**Continuous**	3.93 (2.35)	4.44 (2.43)
Number of Defenders towards Goal (Scaled)	**Count**	7.08 (2.25)	5.08 (2.91)
Pass Height Ground Low[5] High	**Categorical**	 9,928 (46.7%) 1,817 (8.55%) 9,516 (44.76%)	 7,545 (27.34%) 4,006 (14.52%) 16,047 (58.15%)
Body Part Preferred Foot Other Foot Head Other	**Categorical**	 18,632 (87.63%) 2,013 (9.47%) 552 (2.6%) 64 (0.3%)	 23,697 (85.86) 3,712 (13.45) 178 (0.64) 11 (0.04)
Season	**Continuous**	1.58 (1.27)	1.56 (1.31)

[4]StatsBomb pitch units
[5]Low pass: Ball comes off the ground but is under shoulder level at peak height.

6 Model Results

For each model, a random intercept for every individual player in the respective study sample was included in the model build. A random slope on the season was incorporated into the xG and passes models. However, the carries model did not achieve convergence with the random slope. As a work around for the longitudinal analysis, we stratified the study sample data by the season and built a model for each to perform sensitivity analysis at the player level. For the predictor-level analysis, all available seasons were used.

6.1 Shots

The xG fixed effects for both forwards only and full models are provided in **Table 7**. The effects in the table are exponentiated and the results are interpretable as odds ratios. Effects greater than one are interpreted as having a positive multiplicative effect on the odds of a shot conversion. Whereas, effects that are less than one have a negative multiplicative effect. These effects are adjusted for all of the predictors in the model. 95% Wald confidence intervals (CI) are also provided for each of the predictor effects. A CI containing a value of one suggests that the relationship between the predictor and response is not statistically significant (similar to p-value > 0.05). Shot Angle has a positive predictor effect. The remaining variables in the predictor set have negative effects. However, it is important to note that the multi-categorical predictors are relative to their reference level. For example, the body part predictors demonstrate a negative effect or association with scoring a goal because they are compared to shots being taken by a player's preferred foot.

The following interpretations are provided for each type of a predictor for the reader's reference. The interpretation style is the same for all the predictors in all the other models in this paper as they are all GLMMs with binary outcomes.

Shot Angle (Continuous Predictor): For a given player, an increase of one standard deviation (15.89 degrees) from the average angle (27.07 degrees) is estimated to increase the xG value by 65% ([1.65 - 1.00] x 100 = 65%) while adjusting for other predictors.

Body Part; Other Foot (Categorical Predictor): For a given player taking a shot with their non-preferred foot, the xG is estimated to decrease by 21% ([0.89 - 1.00] x 100 = -21%) relative to taking it with their preferred foot while holding other predictors constant.

Defenders in Shot Triangle (Count Predictor): For a given Premier League player, the appearance of an additional defender in the shot triangle is estimated to decrease the xG value by 34% ([0.66 - 1.00] x 100 = -34%) while holding other predictors constant. It is important to note that these "effects" are interpreted as statistical associations and not causal effects.

Table 5. The odds ratios (predictor effects) and their respective 95% confidence intervals (CI) from the Expected Goals (xG) models. **Bolded*** numbers represent effects that demonstrate a statistically significant relationship with the target variable. The effects associated with scaled continuous predictors are interpreted on their respective standard deviation scale. (**continued on next page**)

Predictor	Full Model Estimate	Full Model 95% CI	Forward Model Estimate	Forward Model 95% CI
Body Part Preferred Foot Other Foot Head Other	 ref. **0.79*** **0.31*** **0.35***	 **[0.719, 0.863]** **[0.278, 0.344]** **[0.229, 0.539]**	 ref. **0.79** **0.32** **0.45**	 **[0.704, 0.886]*** **[0.275, 0.367]*** **[0.248, 0.797]***
Defenders In Cone	**0.66***	**[0.638, 0.691]**	**0.61**	**[0.578, 0.651]***
Distance To Goal (Scaled)	**0.60***	**[0.562, 0.645]**	**0.67**	**[0.611, 0.728]***
Goalkeeper In ShotCone	**0.45***	**[0.387, 0.515]**	**0.44**	**[0.37, 0.524]**
Shot Angle (Scaled)	**1.65***	**[1.571, 1.731]**	**1.71**	**[1.601, 1.828]**
Under Pressure	**0.76***	**[0.705, 0.826]**	**0.82**	**[0.738, 0.905]***
Season	**1.03***	**[1.00, 1.063]**	**1.05**	**[1.007, 1.089]***

6.2 Carries

The predictor effects from the carries model is summarized in **Table 8**. We can see that duration and pressure during carry have positive relationships with the response variable. The latter association is quite interesting and unintuitive. However, the data here may suggest that the pressure on the carrier actually leads the player to make an on-ball decision that leads to a successful carry outcome. It would be of interest to test how the association would change if the definiton was altered to something else. Predictors with negative significant associations with the target variable are pressure on the carrier right after the carry, the carry starting after a dribble, and the distance of the dribble.

Table 6. GLMM results for the fixed effect predictors for the carry outcome model. All coefficients and 95% confidence intervals have been exponentiated. **Bolded*** numbers represent effects that demonstrate a statistically significant relationship with the target variable. The effects associated with scaled continuous predictors are interpreted on their respective standard deviation scale.

Predictor	Estimate	95% CI
Duration (Scaled)	**1.68***	**[1.61, 1.75]**
Carry Distance (Scaled)	**0.88***	**[0.86, 0.91]**
Under Pressure During Carry After Carry	 **1.21*** **0.88***	 **[1.11, 1.32]** **[0.80, 0.96]**
Event before Carry Ball Receipt	 0.98	 [0.86, 1.09]

Dribble	**0.80***	**[0.69, 0.93]**
Season	0.98	[0.96, 1.00]

6.3 Passes

The fixed effects for the the passes models are summarized in **Tables 9 and 10.** The interpretations for the predictors are left for the readers to draw. However, we want to note how some of the fixed effects differ between modles both in terms of magnitude and directional relationship with the response variable. For example, the the length of a pass has a negative relationship with the response for passes from midfield but it has a positive association with passes from the wing.This empirical difference in the findings can be a valuable insight for football practitioners especially coaches looking to optimize possession in different attacking zones of the pitch. A few interesting findings are that: (i) low and high passes decrease the odds of a completed pass as opposed to a ground pass and (ii) headed passes from the wing increase the odds of a completed pass. The finding about pass height is interesting because it may motivate coaches to prioritize ground passes from the wing rather than crossing in the ball. The latter finding is quite unintuitive but this may a consequence of a low sample size for headed passes. Additionally, the fixed efffect has quite a wide confidence interval.

Table 7. The odds ratios (predictor effects) and their respective 95% confidence intervals (CI) from the Attacking Midfield Passes model. **Bolded*** numbers represent effects that demonstrate a statistically significant relationship with the target variable. The effects associated with scaled continuous predictors are interpreted on their respective standard deviation scale.

Predictor	Estimate	95% CI
Under Pressure	0.92	[0.842, 1.012]
Pass Length (Scaled)	**0.79***	**[0.756, 0.815]**
Pass Angle (Scaled)	1.00	[0.972, 1.03]
Distance to Closest Defender (Scaled)	**1.09***	**[1.057, 1.131]**
Number of Defenders towards Goal (Scaled)	1.00	[0.963, 1.028]
Pass Height Ground Low High	 Ref. **0.35*** **0.32***	 **[0.317, 0.391]** **[0.29, 0.341]**
Body Part Preferred Foot Other Foot Head Other	 ref. 0.91 0.83 **0.62**	 [0.823, 1.004] [0.493, 1.387] **[0.508, 0.748]**

Season	1.017	[0.991, 1.043]

Table 8. The odds ratios (predictor effects) and their respective 95% confidence intervals (CI) from the Attacking Wing Passes model. **Bolded*** numbers represent effects that demonstrate a statistically significant relationship with the target variable. The effects associated with scaled continuous predictors are interpreted on their respective standard deviation scale.

Predictor	Estimate	95% CI
Under Pressure	**0.83***	**[0.752, 0.923]**
Cut Back	1.14	[0.975, 1.341]
Pass Length (Scaled)	**1.19***	**[1.145, 1.228]**
Pass Angle (Scaled)	0.98	[0.949, 1.008]
Distance to Closest Defender (Scaled)	1.00	[0.967, 1.025]
Number of Defenders towards Goal (Scaled)	**1.15***	**[1.118, 1.185]**
Pass Height Ground Low High	 ref. **0.23*** **0.19***	 **[0.21, 0.25]** **[0.178, 0.208]**
Body Part Preferred Foot Other Foot Head Other	 ref. **0.82*** **1.64*** 2.89	 **[0.753, 0.889]** **[1.193, 2.264]** [0.842, 9.934]
Season	**1.03***	**[1.010, 1.058]**

7 Estimated Player Impact Rankings

The EPI measures are derived from the GLMMS and they represent players' influence on the response variable of interest. For the xG models, the players are ranked by their position groups: (i) Forwards, (ii) Midfielders, (iii) Centre Backs, and (iv) Full Backs & Wing Backs. The random slopes are also provided for the longitudinal aspect of the analyses. Additional relevant metrics are also provided for each set of rankings for context. It is very important to note that these rankings do not suggest that one player is **outright** better than the other. Rather it provides the analyst a data-driven measure of how much impact a player has on the football action of interest compared to the rest of the study sample. The same can be said about the random slope. We encourage the readers to challenge our findings by running their own analyses on players and also demonstrate

how to improve the GLMMs.

7.1 Expected Goals

Table 9. Forward Estimated Player Impact (EPI) rankings from forwards only expected goals (xG) model. The difference in goals and StatsBomb xG is provided for additional insight.

Player	EPI	Slope	Goals (xG)	xG Difference
Heung-Min Son	0.42	-0.001	69 (48.4)	20.59
Phil Foden	0.29	+0.002	53 (34.3)	18.71
Harry Kane	0.21	+0.002	57 (45.7)	11.32
Marcus Rashford	0.20	-0.003	42 (32.1)	9.86
Harvey Barnes	0.18	0.001	42 (29.7)	12.28

Findings from this table suggest that Heung-Min Son has the highest player impact on converting shots into goals. We also note that his slope is roughly zero further suggesting that his goal scoring impact does not change at a different rate than the population average of forward players in the study. For all five of the top players here, they demonstrate very small slope measures. Examples of players who demonstrated high positive slope values are (i) Bryan Mbeumo (+0.014), (ii) Chris Wood (+0.014), and (iii) Callum Hudson Odoi (+0.009).

"Where is Erling Haaland?" We expect that the reader may have paused and wondered about this specific question. Haaland is ranked 12th in this analysis which includes forwards from the last five seasons. His EPI value is 0.14 (slope: -0.001) which can be interpreted as a +0.14 increase in the log-odds of converting a shot into a goal or 15% increase in goal odds. To explain why Son is ranked higher than Haaland, we first look at the StatsBomb xG for both of the players. Haaland has scored 68 goals from a total xG of 61.75, a difference of 6.26. Son has scored 69 goals from a total xG of 48.4, a 20.59 difference. This suggests that Son has scored from lower xG opportunities as opposed to Haaland. This ***can be one*** of the explanations behind our model findings. However, the more interesting justification lies within how the GLMMs work. The predictors in a regression based model are used to explain the observed variance in the response variable. It is empirically rare to be able to fully explain the variance and there is leftover variance that is unexplained. In the particular case of a hierarchical model, the random intercept attempts to explain some of the remaining variance: the player predictor explains differences in shot conversion. The findings in our analysis suggest that Heung-Min Son as a player can explain more of the variability in the response than Haaland. In other words, Son is actually scoring goals, on average, from shots that have low-odds in terms of predictor effects. We can see this trend by comparing Haaland and Son's descriptive statistics for their shots in **Table 11** below.

Table 10. Summary statistics of shot predictors for both Heung-Min Son and Erling Haaland used in the expected goals model. n represents the number of shots by each player. The predictors are sorted by the magnitude of the fixed effect.

Predictor	Forward xG Model Fixed Effect	Son n = 365	Haaland n = 336
Shot Angle (radians)	1.71	24.68 (11.51)	**36.02 (18.3)**
Body Part Preferred Foot Other Foot Head Other	 ref. 0.79 0.32 0.45	 208 (57.0%) **141 (38.6%)** 15 (4.1%) 1 (<0.3%)	 186 (55.4%) 48 (14.3%) **100 (29.8%)** 2 (<0.2%)
Goalkeeper In Shot Cone	0.44	**358 (98.1%)**	307 (91.4%)
Defenders In Shot Triangle	0.61	**0.83 (0.9)**	0.59 (0.75)
Distance To Goal (SB[6] units)	0.67	**17.43 (6.5)**	12.39 (5.13)
Under Pressure	0.82	95 (26.3%)	**130 (38.69%)**

We can see that Son's average shot distance is 17.43 (standard deviation: 6.5) whereas Haaland's shot distance is lower. The fixed effect for distance is 0.67 which implies that longer distances decrease goal-odds by a large amount. Regardless of this association, Son still shoots from a longer average distance and manages to score goals based on his finishing skill.

Additionally, we also observe that Son has shot with his "Other Foot" **38.6%** times but Haaland mainly shoots with his preferred foot or with his head. Headers are usually a consequence of the received pass whereas the choice of feet in shooting usually is left up to the player. We can also see that Haaland has a higher average shot angle than Son. This particular predictor significantly increases the odds of a goal whereas Son manages to score from average angles of approximately 25 degrees (in radians). The fixed effects analysis suggests that there are population-level trends that players can follow to improve their goal scoring opportunities. In this particular example, it looks like Haaland's chances are, on average, more optimized for a goal than Son's chances. This type of analysis is what makes the random intercepts a very powerful tool for player evaluation because it can capture latent behavior in how players choose to play the game of football. Son is a player who is comfortable shooting with both legs from tight angles whereas Haaland is a player who converts his high xG chances. Two different types of players whose latent impact on goal

[6]StatsBomb pitch units

scoring is quantified by the hierarchical models. Building models using data from undertapped or undervalued leagues can lead to discovering talent who are not being tracked by classical models.

Table 11. Forward Estimated Player Impact (EPI) rankings from full expected goals (xG) model. The difference in goals and StatsBomb xG is provided for additional insight.

Player	EPI	Slope	Goals (xG)	xG Difference
Heung-Min Son	0.512	-0.0310	69 (48.4)	20.59
Phil Foden	0.358	-0.0154	53 (34.3)	18.71
Marcus Rashford	0.257	-0.0143	42 (32.1)	9.86
Harry Kane	0.252	-0.0112	57 (45.7)	11.32
Harvey Barnes	0.23	-0.0104	42 (29.7)	12.28

The top five rankings for the forwards barely changed with Rashford and Kane switching spots. The change in their EPI estimate is rather miniscule. However, what is more interesting here is the slope measures. When compared to every shot-taker, these players all demonstrate a decrease in their EPI over time at a higher magnitude than when compared to shot-takers. This finding suggests that there is indeed a trend difference when comparing players within position groups when it comes to xG EPI.

Table 12. Midfielder Estimated Player Impact (EPI) rankings derived from the full expected goals model.

Player	EPI	Slope	Goals	xG Difference
James Maddison	0.321	-0.0176	37 (24.8)	12.16
Kevin De Bruyne	0.267	-0.0141	32 (21.4)	10.58
Matheus Cunha	0.222	-0.010	27 (15.6)	11.37
Bruno Guimarães	0.217	-0.012	21 (14.3)	6.73
Jesse Lingard	0.199	-0.013	10 (5.3)	4.69

Current active players in the premier league who would break into the top 5 are Dejan Kulusevski (EPI: 0.183, Slope: -0.001) and Martin Odegaard (EPI: 0.181, -0.0083).

Table 13. Centre back Estimated Player Impact (EPI) rankings from the full expected goals (xG) model.

Player	EPI	Slope	Goals	xG Difference
Michael Keane	0.20	-0.008	11 (5.4)	5.57
John Stones	0.14	-0.006	10 (5.1)	4.85
Thiago Silva	0.12	-0.006	8 (4.8)	3.21

Kurt Zouma	0.11	-0.007	11 (7.0)	4.01
Trevoh Chalobah	0.11	-0.006	7 (3.8)	3.17

The two highest goal scorers by volume are Gabriel (12 goals) from Arsenal and Virgil van Dijk (12 goals) from Liverpool. Both of them have xG values of 14.3 and 11.1 xG respectively suggesting that they are usually converting the higher xG chances, on average when attempting to score a goal. A reasonable takeaway is that players here are individuals who improve the odds of lower xG chances. However, we want to make the point that individual-level analysis on centre backs through frequentist hierarchical models is not as statistically sound due to low sample sizes for the players.

Table 14. Fullback/Wingback Estimated Player Impact (EPI) rankings from the expected goals (xG) model.

Player	EPI	Slope	Goals (xG)	xG Difference
Stuart Dallas	0.14	-0.010	9 (5.4)	3.58
Olaoluwa Aina	0.124	-0.005	5 (1.6)	3.40
Jack Hinshelwood	0.121	-0.007	8 (4.5)	3.46
Joško Gvardiol	0.08	-0.005	9 (7.4)	1.63
Pedro Porro	0.066	-0.004	8 (4.9)	3.07

7.2 Carries into Box

Table 15. Estimated Player Impacts of Carriers into the Box. OBV refers to StatsBomb's On-Ball value metric. +EPI refers to a positive EPI value.

Player	EPI	Successful Carries (% Total)	+EPI Seasons (Total)	Post-Carry OBV per match
Jack Grealish	0.399	207 (55%)	5 (5)	0.22
Martin Ødegaard	0.393	80 (60%)	5 (5)	0.10
Jadon Sancho	0.388	107 (58%)	3 (4)	0.15
Bernardo Silva	0.376	113 (58%)	5 (5)	0.11
Emile Smith Rowe	0.372	36 (64%)	4 (4)	0.18
Heung-Min Son	0.343	138 (53%)	5 (5)	0.19
Callum Hudson-Odoi	0.322	88 (57%)	4 (4)	0.14
Emmanuel Dennis	0.316	41 (55%)	1 (2)	0.13
Amad Diallo	0.28	45 (52%)	2 (2)	0.16
Willian	0.255	41 (57%)	3 (4)	0.08

The carries model did not converge for the random slope. As a consequence, we stratified by season and ran a separate model for each of the available seasons in the dataset. Jack Grealish

ranks as the best progressive carrier into the box even with the largest volume of attempted carries. He also had a positive EPI value in all of his seasons which were spent playing at Aston Villa and Manchester City demonstrating he is a robust player when it comes to progressive carries into the box. Emile Smith Rowe is also an interesting player in this table as he has played for Arsenal but departed for Fulham due to lack of playing minutes.

7.3 Passes into Box

Table 16: Estimated Player Impact rankings for Attacking Passes into the Box from Midfield

Player	EPI	Slope	Completed Passes (% Total)	Shots Assists (per match)
Jack Grealish	0.165	0.039	100 (76%)	21 (0.31)
Rodri	0.138	0.078	133 (57%)	16 (0.16)
Granit Xhaka	0.108	-0.001	84 (62%)	20 (0.29)
Declan Rice	0.094	0.043	97 (62%)	21 (0.23)
Moisés Caicedo	0.069	0.031	69 (61%)	14 (0.21)

Jack Grealish ranks as number one in this category on top of the carries rankings. This makes him a very interesting figure as he has recently departed for Everton (on loan) from Manchester City in 2025/2026. Both Declan Rice and Moisés Caicedo appear in this table. These two players were two of the most expensive transfers in recent history and both have a positive slope from the model emphasizing their improvement over time.

Table 17: Estimated Player Impact rankings for Attacking Passes into the Box from Midfield

Player	EPI	Slope	Completed Passes (% Total)	Shots Assists (per match)
Adam Smith	0.328	-0.054	43 (47%)	21 (0.43)
Mohamed Salah	0.319	-0.079	170 (47%)	38 (0.27)
Leandro Trossard	0.296	-0.026	127 (45%)	29 (0.22)
Wilfried Zaha	0.274	-0.004	73 (58%)	14 (0.23)
Jadon Sancho	0.258	-0.0192	67 (60%)	17 (0.30)

Adam Smith from AFC Bournemouth ranks as the number one player when it comes to EPI on attacking passes into the box from the wing. This is a finding that can be viewed as an anomaly by the regular football supporter. However, a finding like this can lead to interesting data-driven investigations. We leave it to the reader to challenge or support this discovery.

8 Case Study: Oliver Glasner's Crystal Palace

Oliver Glasner took over as the manager of Crystal Palace Football Club before matchweek 26 in the 2023/2024 season. Since then, his side have won two trophies by beating Pep Guardiola's Manchester City in the 24/25 FA Cup final and Arne Slot's Liverpool, the defending Premier League Champions, in the 2025 FA Community Shield. On top of those two trophies, Palace are also competing in their debut European campaign in the 2025/2026 season. Considering Crystal Palace's history, all of these achievements are truly remarkable for all parties involved at the club from South London.

Motivated by their recent successes and the departure of Eberechi Eze, we perform a data driven case study on Crystal Palace using the hierarchical models in this paper. The purpose of this case study is to demonstrate how one could leverage these advanced but intuitive models into their player scouting and evaluation process. We do not make any claims that the models we have built here are the best nor that they should be the gold standard. However, we hope that this case study gives football practitioners, researchers, journalists, and especially students inspiration to add hierarchical models into their data analysis toolkit. As a matter of fact, we encourage anyone to take our work and improve upon it.

8.1 Attacking Data Trends

The primary focus of this case study will be on the attacking side of Glasner's side particularly how they attack the opposition penalty box. The carries and pass models in this paper will be handy tools in this study. Based on descriptive data analysis in Table CP, we can observe that Glaser has started in a 3-4-2-1 formation 31 times in the 24/25 season. Even during the matches, Glasner has opted to make a tactical shift to the 3-4-2-1 a total of 23 times which accounts for 46% of all of the formation switches as tracked by StatsBomb's data

Table CP1. Starting formation statistics for Crystal Palace FC (2024/2025 season)

Starting Formation	Count
3-4-2-1	31
3-4-3	6
3-5-2	1

Table CP2. Tactical Formation shift statistics for Crystal Palace (2024/2025 season)

Tactical Shifts during Match Play	Count (% of Shifts)
3-4-2-1	23 (46%)
3-4-3	6 (12%)
3-5-2	6 (12%)
4-4-2	5 (10%)

3-5-1-1	5 (10%)
4-2-3-1	4 (8%)
4-3-3	1 (2%)

By nature of this formation, it is reasonable to assume that Glasner leverages his wing backs in the attacking phases of play. This is supported by an analysis of the open play touches by all of the Crystal Palace fullbacks and wing backs in Figure CP1. We can see that the wing backs also venture into the right sided half spaces quite a bit as well.

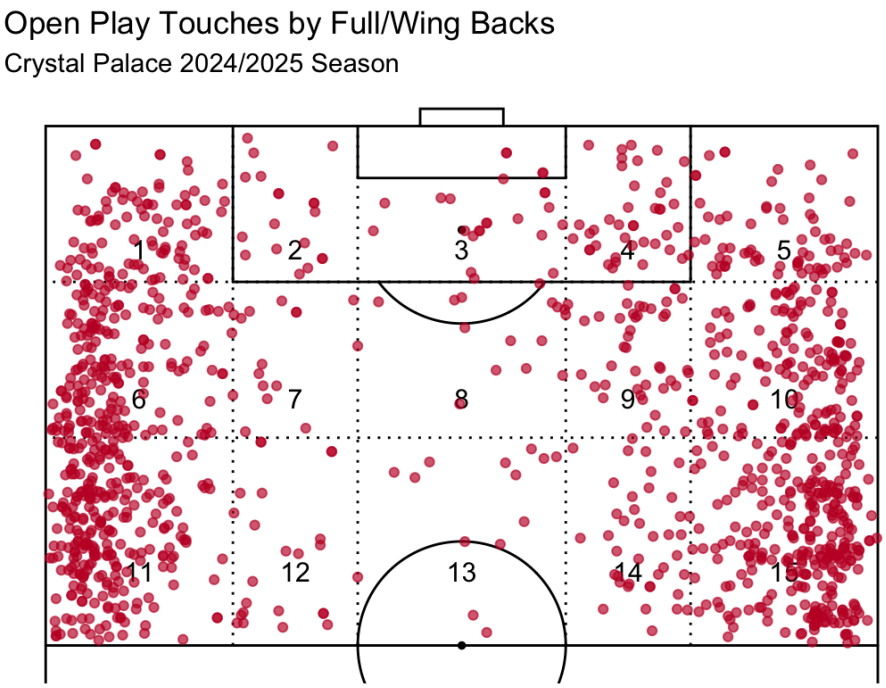


Figure CP1. Visualization of open play touches in the attacking half of the football pitch by players who fall under the fullback or wingback category at Crystal Palace (2024/2025 season).

Table CP3 breaks down the touch statistics by position group and the players in each group. Crystal Palace midfielders rank as the highest when it comes to overall open touches in the opposition half. However, we can see that Daniel Muñoz and Tyrick Mitchell rank as the 2nd and 4th players in terms of touch count. This further supports the involvement of the wing backs in Glasner's attack.

Table CP3. Tactical Formation shift statistics for Crystal Palace (2024/2025 season)

Position Group	Open Play Touches	Top 3 Players	Open Play Touches
Midfielder	2,828	E. Eze I. Sarr J. Lerma	634 540 380

Wing/Full Back	1,250	D. Muñoz T. Mitchell N. Clyne	615 526 60
Forward	834	J-P Mateta E. Nketiah O. Édouard	518 272 25
Centre Back	744	M. Guehi M Lacrois C Richard	347 164 114

In terms of attacking passes, Crystal Palace tends to attack the box from the wings more often than other parts of the pitch. This can be seen in **Figure CP2**. Zones 1 and 5 have the highest number of passes and the black arrows represent the most common passes into the box from those zones. We can see that the passes are pinged into the middle of the box for teammates to receive or shoot the ball. **Figure CP3** further supports the wingbacks' involvement by showing that Mitchell and Munoz rank in the top two for most passes attempted into the box. Additional inferences that we can draw from this plot is that Ismaila Sarr and Eze are also highly involved in the attacking passing phases of play. After Eze, Adam Wharton attempted 56 passes into the box which is a 30% drop from Eze's number.

Table CP 4 summarizes the breakdown of passes by the top five zones for attacking passes. For zone 1, Mitchell accounts for the majority of passes and also leads the club in terms of shot assists (which is a measure of a pass being converted into a shot). However, for Zone 5, Daniel Munoz does not rank as number one instead we see that Sarr, a midfielder, accounts for 39 of the total passes. Daniel Munoz does rank as the highest passer from Zone 10 which is also part of the wing. Zones 1 and 5 led to the highest number of shot assists for Palace demonstrating the importance of these zones and the wingbacks in Glasner's system.

Figure CP2. Visualization of the attacking passes into the box by Crystal Palace players in the 2024/2025 Premier League season. The coral orange arrows represent each individual pass. The black arrows represent the average passes from zones 1 and 5. The darker the green color for the zone, the more passes originated from that zone and vice versa.

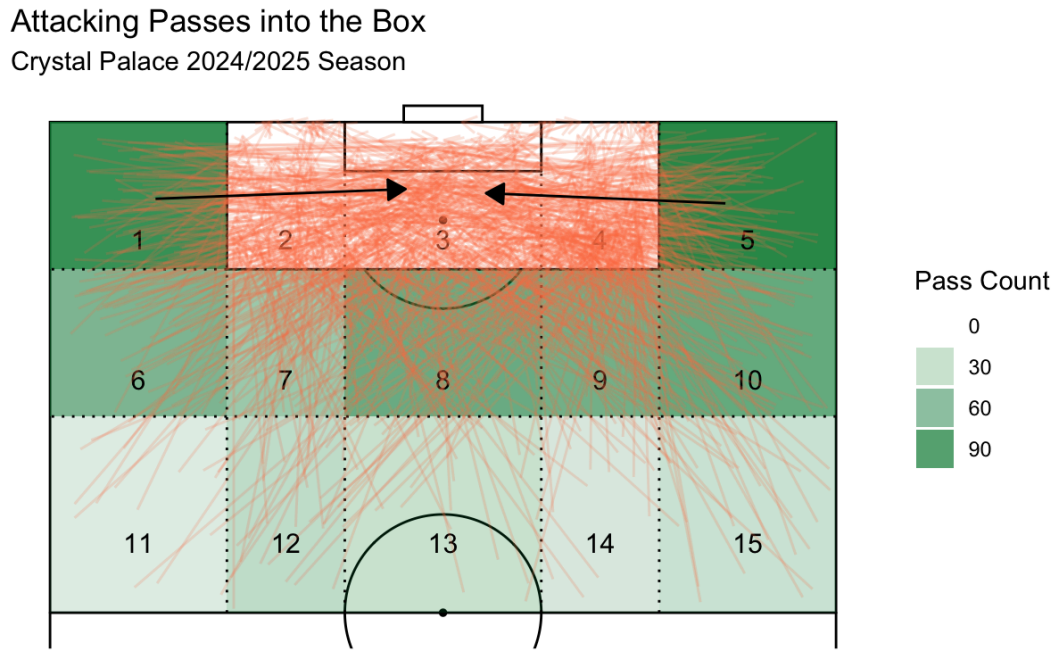


Figure CP3. Frequency plot summarizing the top 10 players who attempt passes into the opposition penalty box. The plot is arranged by pass attempts and the number in the parenthesis represents the count of passes converted into shots.

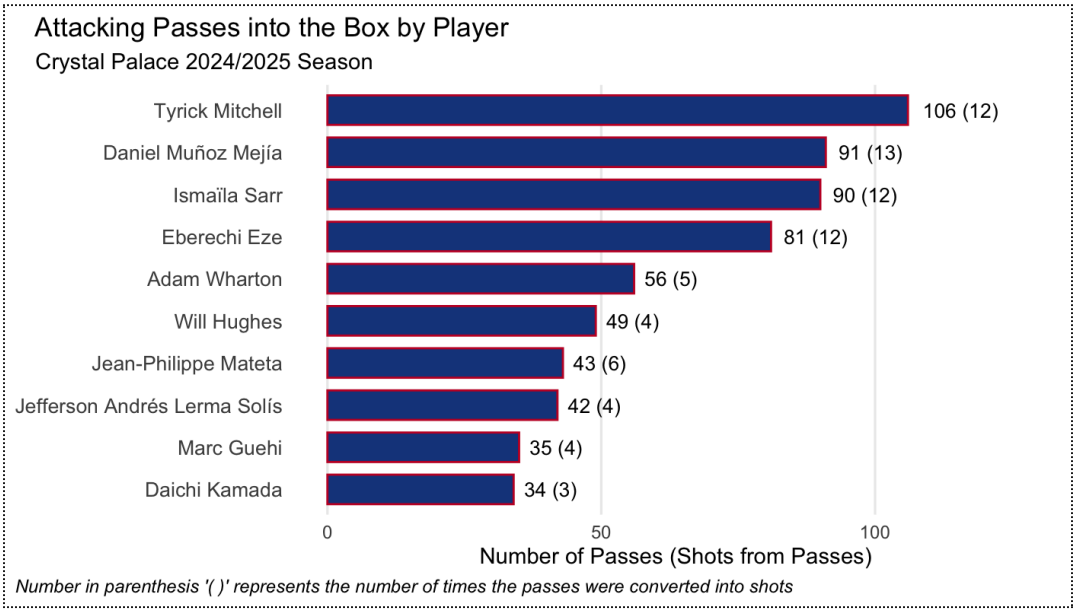


Table CP4. Statistical breakdown of passes, pass completion rate (success %), and shot assists by zone and by players for Crystal Palace during the 2024/2025 season

Zone	Passes (Success %)	Shot Assists	Player	Passes (Success %)	Shot Assists
Zone 1	106 (37%)	19	**T. Mitchell** J-P Mateta E. Eze	60 (38%) 10 (20%) 8 (50%)	10 0 2
Zone 5	116 (34%)	18	I. Sarr **D. Munoz** J-P Mateta	39 (33%) 34 (41%) 9 (22%)	5 9 1
Zone 8	79 (61%)	14	I. Sarr E. Eze J-P Mateta	15 (80%) 14 (57%) 8 (88%)	4 3 4
Zone 9	75 (52%)	9	I. Sarr A. Wharton W. Hughes	16 (75%) 15 (33%) 12 (50%)	3 0 1
Zone 10	82 (31%)	6	**D. Munoz** I. Sarr A. Wharton	35 (26%) 11 (18%) 10 (40%)	1 0 2

As for progressive carries into the box, the most common carry zones are zones 7, 8, and 9. Figure CP and Table support this claim. Both Eze and Sarr served as no. 10's in Glasner's team and they accounted for the most carries into the box by Palace players (see **Figure CP5**). Most of Eze's carries are from zone 7 and Sarr's carries are from zone 10. The black arrows illustrate the average carries from both of these zones. We can see that the average carries do not go deep as into the box as the passes in **Figure CP2.** This suggests that both of these players initiate a different action as soon as they enter the box. Zone 8 is also a common origin for carries into the box for Palace which both Eze and Sarr rank in the top three. StatsBomb On-ball value (OBV) measures are included in Table **CP4** to provide an additional layer to the analysis. OBV is a value that is assigned to an on-ball action based on the action's impact on the probability of scoring a goal for the team.

Figure CP5. Frequency plot summarizing the top 10 players who performed progressive carries into the opposition penalty box. The plot is arranged by frequency and the numbers in the parenthesis represent the Hudl StatsBomb on-ball value (OBV) after the carry action and the net OBV after the carry action respectively.

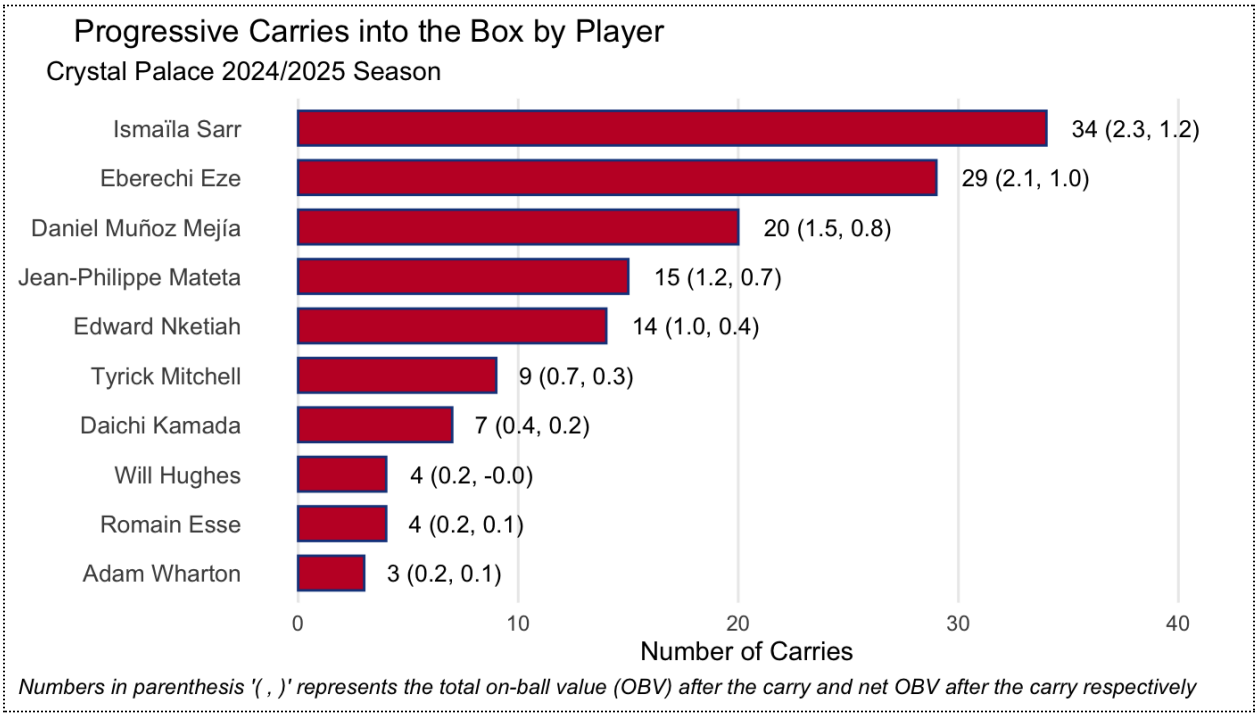


Figure CP4. Visualization of the progressive carries into the box by Crystal Palace players in the 2024/2025 Premier League season. The coral orange arrows represent each individual carry. The black arrows represent the average carries from zones 8 and 9. The darker the green color for the zone, the more carries originated from that zone and vice versa.

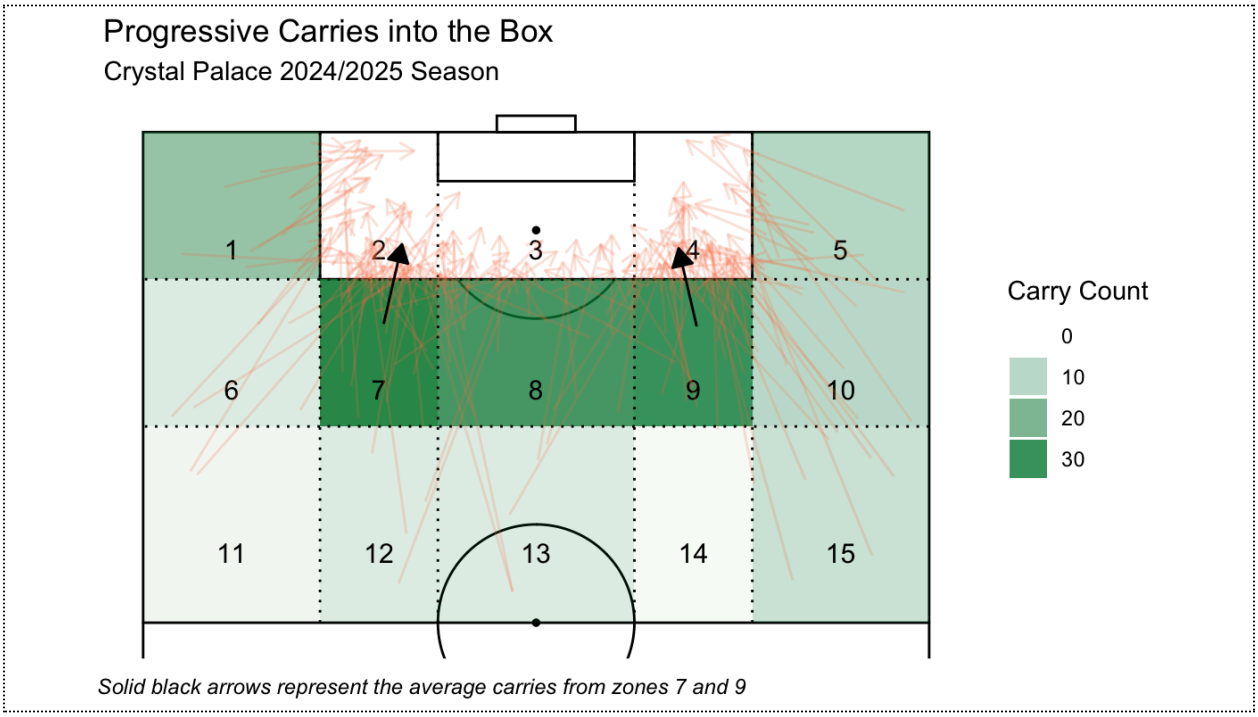


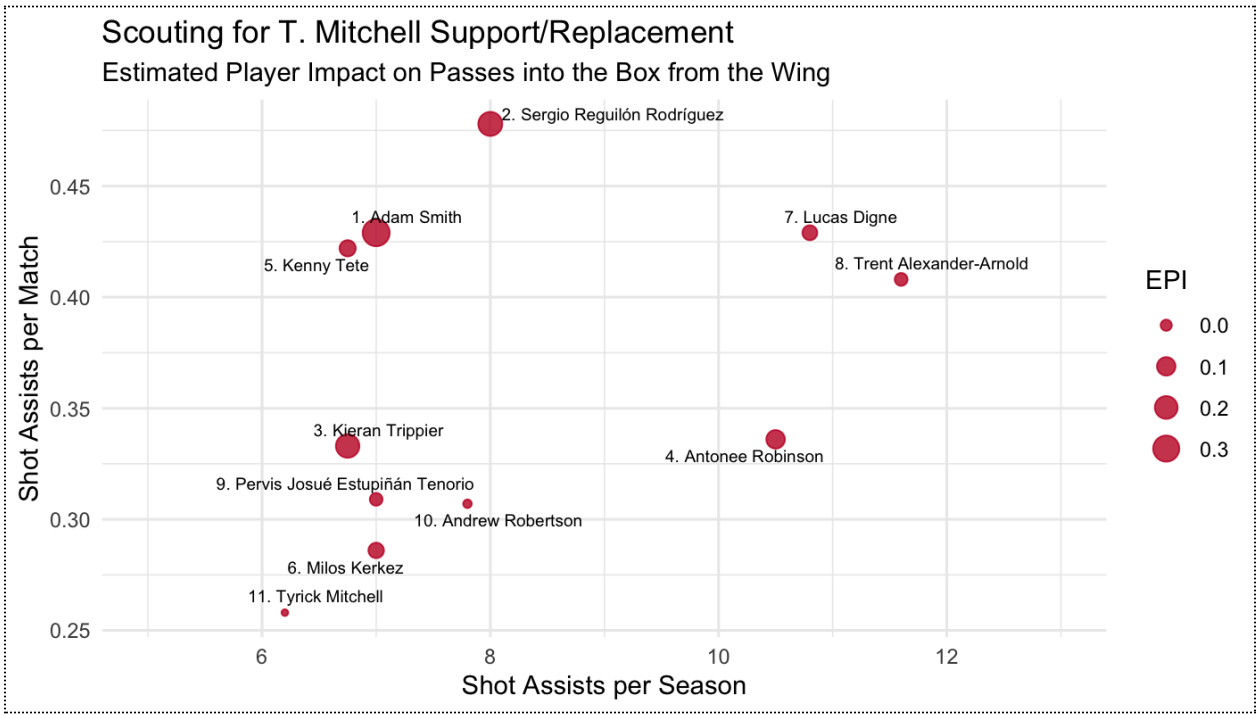
Table CP5. Statistical breakdown of carries by zone and by players for Crystal Palace during the 2024/2025 season

Zone	Carries	OBV (Net OBV)	Player	Carries	OBV (Net OBV)
Zone 7	33	2.37 (1.16)	**E. Eze** J-P Mateta T. Mitchell	10 7 4	0.73 (0.35) 0.54 (0.29) 0.36 (0.19)
Zone 8	28	2.48 (0.95)	**E. Eze** E. Nketiah I. Sarr	6 4 4	0.59 (0.23) 0.31 (0.13) 0.43 (0.21)
Zone 9	30	2.09 (0.94)	**I. Sarr** D. Munoz D. Kamada	9 7 3	0.56 (0.25) 0.52 (0.25) 0.22 (0.11)

8.3 Estimated Player Impacts for Scouting

Motivated by these findings, we use the carries and pass models in our paper to provide complementary analyses for the Crystal Palace scouting team. Naturally, as an academic paper, we did not work with real-world scouts for this hypothetical case study. Therefore, we make some assumptions to demonstrate **how the models can be used**. It is important to remind the reader that the models and analysis can be tweaked based on what the football practioner or client is truly interested in. For this study, we make the scenario such that the scouting team at Palace is specifically looking for players who would complement Mitchell and Munoz as competition for the upcoming season. They are interesed in Premier League players who have demonstrated higher shot assists than both Munoz and Mitchell. To help with this scouting, we filter for players who have higher shot assists per match, per season, and higher EPI values for attacking passes into the box from the wing. The analysis is summarized in Figure CP 6. The scouts and analysts can then use this visual to pefrom additional tactical analysis or scouting on players they deem interesting or reasonable for Glasner's side. For example, extensive analysis on Sergio Reguillon could be of interest since he has already played for three different Premier League clubs but he demonstrates higher measures than Mitchell on all accounts in this specific analysis.

Figure CP6. Scatterplot of players who are possible replacements or competition for Tyrick Mitchell at Crystal Palace. The y-axis represents shot assists per match, the x-axis represents shot assists per season, and the size of the dot represents the magnitude of the estimated player impact for attacking passes into the box.



Similar analyses are done for Eze who departed the club this season. The findings are summarized in Figure **CP7.** The two additional metrics used in this demo analysis are OBV after the carry per season and per match. Additionally, the EPI on carries into the box are included in the analysis. Based on this analysis, we can actually see that Ismaila Sarr actually has higher measures for all three metrics than Eze. This suggests that Palace already have a capable player in Sarr to play the Eze role from a carries perspective. It is interesting to note that all the other players that demonstrated higher values than Eze are players who play for the "Big Six" clubs in England. This make them less feasible alternatives for an Eze replacement. So, we relax the requirement and look at players who are within 0.04 units of Eze's OBV measures. This leads to identifying Emile Smith Rowe at Fulham. Depending on the needs, the analyst can relax the filter further.

Figure CP6. Scatterplot of players who are possible replacements or competition for Tyrick Mitchell at Crystal Palace. The y-axis represents shot assists per match, the x-axis represents shot assists per season, and the size of the dot represents the magnitude of the estimated player impact for attacking passes into the box.

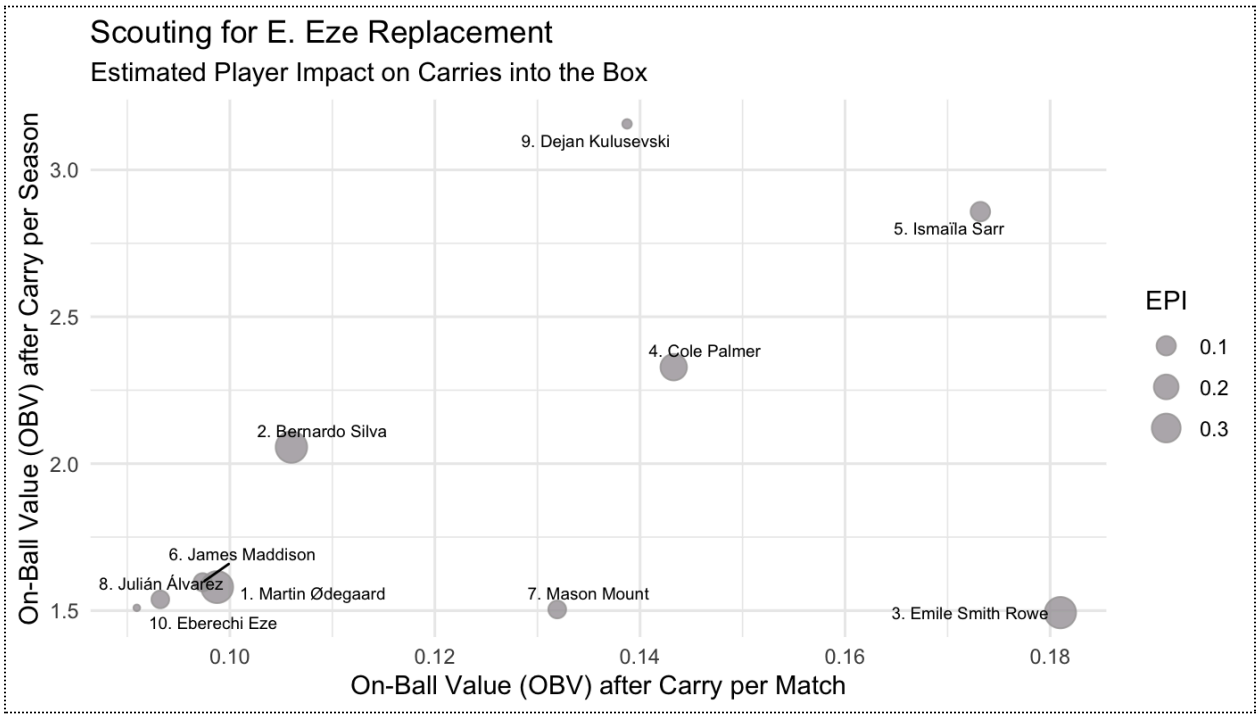
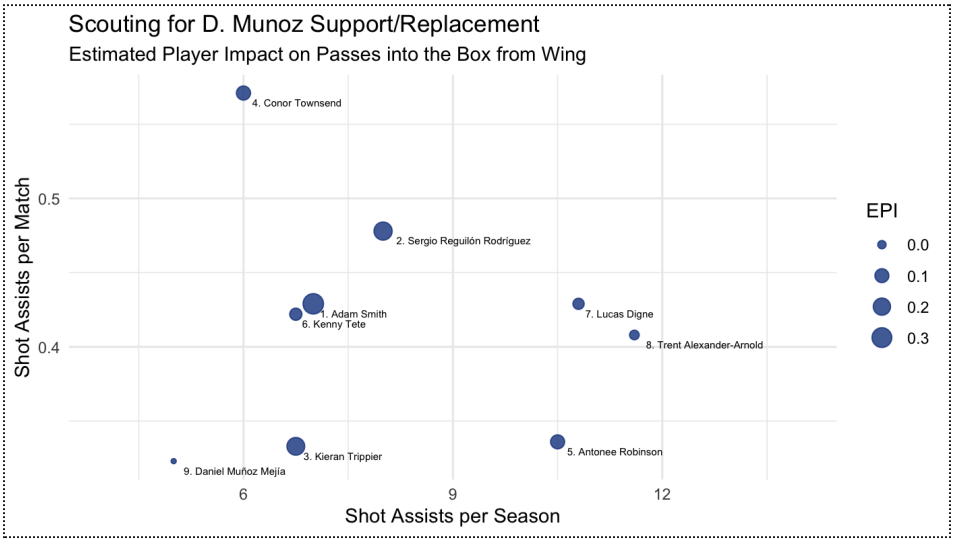


Figure CP6. Scatterplot of players who are possible replacements or competition for Tyrick Mitchell at Crystal Palace. The y-axis represents shot assists per match, the x-axis represents shot assists per season, and the size of the dot represents the magnitude of the estimated player impact for attacking passes into the box.



We also provide an additional analysis for Daniel Munoz's competition to further demonstrate that these models can be adjusted for different types of analysis based on the football needs of the scouting team or coach.

9 Discussion

(i) build GLMMs that allow for the longitudinal analysis of football players, (ii) apply the hierarchical model framework to two additional football actions of interest, (iii) estimate the player impact on football actions of interest and allow for player-adjusted predictions, and (iv) draw generalizable inferences about the relationships between the predictor and response variables in a football context. The focal point of this research is to re-iterate the capabilities of hierarchical models as well as demonstrate to the analytics community how to design a statistically sound longitudinal study using event-level data.

9.1 Findings

In this research, we successfully built and implemented GLMMs that enable the longitudinal analysis of football players using event-level data. Through this framework, we were also able to estimate player-specific impacts on football actions of interest and adjust for both individual player variability and temporal trends across seasons through random parameters.

Furthermore, we extended the hierarchical modeling approach to two additional football actions, demonstrating the flexibility of this framework in capturing football analytics beyond just xG evaluation. The models provided robust, generalizable inferences about the relationships between key predictor and response variables within a football context as well. This highlights how multilevel structures can effectively account for player skill differences and repeated measures over time.

Finally, we conducted a case study to illustrate how football analysts can leverage hierarchical models in practical applications such as player evaluation and scouting. This exercise reinforced the practical utility and interpretability of the approach, emphasizing how GLMMs can bridge the gap between academic statistical modeling and applied football analytics.

9.2 Future Work

We would like to extend our analysis in several directions. The event data can be analyzed either at the match level, the week level or the season level, depending on the research question. Analyzing data on the match level would allow us the most granular view of individual player trajectories, but it is also computationally expensive, especially with the number of players and matches in the dataset. An extension to the GLMM framework is the Generalized Additive Model, which is a semi-parametric (or even non-parametric) framework that can "smooth" out the noise from large volumes of data to detect the overall trends. Careful selection of the specific functional forms for the GAM is essential, which will require us to do extensive model comparisons across training and test datasets to ensure that we have fit the data well with our selected models.

9.3 Bayesian methods

Another aspect we would like to explore further is how to better quantify the uncertainty surrounding both the data and our resulting estimates. Although the HUDL data is incredibly

precise at the match level, we know that the nature of the sport means that the probability of shot conversion or a successful pass is highly variable and can change quickly, depending on multiple factors during a game. This means that it may be equally useful to have uncertainty intervals around our estimates so that we have some sense of the upper and lower bound on the probability of shot conversion (or successful pass). Bayesian methods provide a natural way for quantifying this uncertainty, since the final output is a *distribution* of possible values, rather than a single point estimate. In particular, Bayesian methods could potentially alleviate the random slope convergence issues we faced with the carry models, since we can place a weakly informative prior on the distribution of the random slope, which should help with estimating the individual differences [5]. Due to time constraints, we were not able to fully explore this research direction, but we leave it open for future research.

Acknowledgements

We would like to thank Hudl StatsBomb for providing the data and for giving us the platform to share our football research at the 2025 Hudl Performance Insights Conference in London, United Kingdom. We would also like to thank the following people for their support during the research and conference preparation process.

Kailee Luddy
Katie Slade Howell
Scott Johnson

Contact Information

We encourage anyone to reach out to us, the authors, for any questions related to our work. We love the beautiful game as much as you do!

Tahmeed Tureen <tureen@umich.edu>
LinkedIn: tahmeed-tureen

Irena Chen <irena@umich.edu>
LinkedIn: irena-chen

References

[1] Gabriel Anzer and Pascal Bauer. 2021. A Goal Scoring Probability Model for Shots Based on Synchronized Positional and Event Data in Football (Soccer). *Front. Sports Act. Living* 3, (March 2021). https://doi.org/10.3389/fspor.2021.624475

[2] Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* 68, 3 (April 2013), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

[3] Jonty Colman. 2025. Expected goals: What is xG in football and how does it work? *BBC Sport*. Retrieved November 4, 2025 from https://www.bbc.com/sport/football/articles/cgrqd18q0rgo

[4] Garrett Fitzmaurice, Marie Davidian, and Geert Verbeke (Eds.). 2009. *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*. Chapman and Hall/CRC, Boca Raton.

[5] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis*. CRC Press.

[6] James H. Hewitt and Oktay Karakuş. 2023. A machine learning approach for player and position adjusted expected goals in football (soccer). *Frankl. Open* 4, (September 2023), 100034. https://doi.org/10.1016/j.fraope.2023.100034

[7] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning: with Applications in R*. Springer US, New York, NY. https://doi.org/10.1007/978-1-0716-1418-1

[8] Maximilian Klemp, Fabian Wunderlich, and Daniel Memmert. 2021. In-play forecasting in football using event and positional data. *Sci. Rep.* 11, 1 (December 2021), 24139. https://doi.org/10.1038/s41598-021-03157-3

[9] Matthias Kullowatz. 2015. Expected Goals 3.0 Methodology. *American Soccer Analysis*. Retrieved November 4, 2025 from https://www.americansocceranalysis.com/home/2015/4/14/expected-goals-methodology

[10] Nan M. Laird and James H. Ware. 1982. Random-Effects Models for Longitudinal Data. *Biometrics* 38, 4 (1982), 963–974. https://doi.org/10.2307/2529876

[11] KUNG-YEE LIANG and SCOTT L. ZEGER. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73, 1 (April 1986), 13–22. https://doi.org/10.1093/biomet/73.1.13

[12] James Mead, Anthony O'Hare, and Paul McMenemy. 2023. Expected goals in football: Improving model performance and demonstrating value. *PLOS ONE* 18, 4 (April 2023), e0282295. https://doi.org/10.1371/journal.pone.0282295

[13] Opta Analyst [@OptaAnalyst]. 2025. Declan Rice leads all Arsenal players this season for: Carries (97) Line-breaking passes (52) Possession Won (26) Chances created (9) Is he becoming the most complete midfielder in the Premier League? https://t.co/kkn4x3x7jv. *Twitter*. Retrieved November 5, 2025 from https://x.com/OptaAnalyst/status/1972952733616611681

[14] Opta Analyst [@OptaAnalyst]. 2025. Tottenham's xG against Chelsea of 0.1 was their second lowest on record in a Premier League game. At this rate, they would have needed to play Chelsea 10 times - 900 minutes of football, or 15 hours - just to generate 1.0 xG, enough to expect them to score one single goal. https://t.co/mrhe4B32CY. *Twitter*. Retrieved November 5, 2025 from https://x.com/OptaAnalyst/status/1985400933895868475

[15] Adan Partida, Anastasia Martinez, Cody Durrer, Oscar Gutierrez, and Filippo Posta. 2021. Modeling of Football Match Outcomes with Expected Goals Statistic. *J. Stud. Res.* 10, (March 2021). https://doi.org/10.47611/jsr.v10i1.1116

[16] Alexander Scholtes and Oktay Karakuş. 2024. Bayes-xG: player and position correction on expected goals (xG) using Bayesian hierarchical approach. *Front. Sports Act. Living* 6, (June 2024). https://doi.org/10.3389/fspor.2024.1348983

[17] Tom A B Snijders and Roel J Bosker. 2025. Multilevel Analysis. *SAGE Publications Ltd*. Retrieved November 4, 2025 from https://uk.sagepub.com/en-gb/eur/multilevel-analysis/book234191

[18] Sofascore Football [@SofascoreINT]. 2025. 🌍 | Liverpool beat Real Madrid • xG: 2.58 – 0.45 • Shots (on target): 17 (9) – 8 (2) • Big chances: 4 – 1 • Touches in opp. box: 27 – 26 • Possession: 39% – 61% Reds' November turnaround continues, with their second win (and a second clean

sheet) in as many matches! 👏👏 https://t.co/dx21rhiJs6. *Twitter*. Retrieved November 5, 2025 from https://x.com/SofascoreINT/status/1985828969053110771

[19] Tim Taha and Ahmed-Yahya Ali. 2023. Greater numbers of passes and shorter possession durations result in increased likelihood of goals in 2010 to 2018 World Cup Champions. *PLOS ONE* 18, 1 (January 2023), e0280030. https://doi.org/10.1371/journal.pone.0280030

[20] 2020. Enhancing xG models with freeze frame data. *DTAI Sports*. Retrieved November 6, 2025 from https://dtai.cs.kuleuven.be/sports/blog/enhancing-xg-models-with-freeze-frame-data/

[21] How the growth of wearable technology is transforming football - The Athletic. Retrieved November 6, 2025 from https://www.nytimes.com/athletic/4966509/2023/10/19/wearable-technology-in-football/

[22] **Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015).** *Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

[23] **Brauer, M., & Curtin, J. J. (2018).** *Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. Psychological Methods, 23*(3), 389–411. https://doi.org/10.1037/met0000159

Appendix

Figure A1

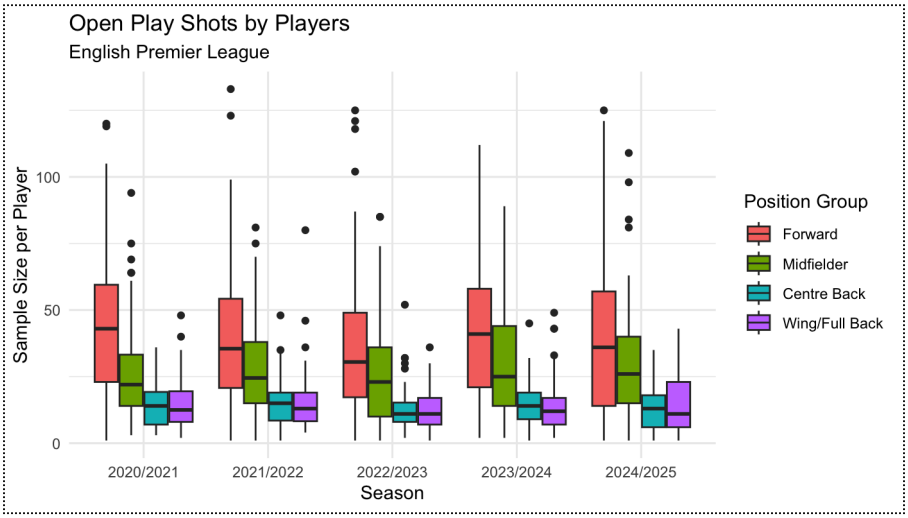
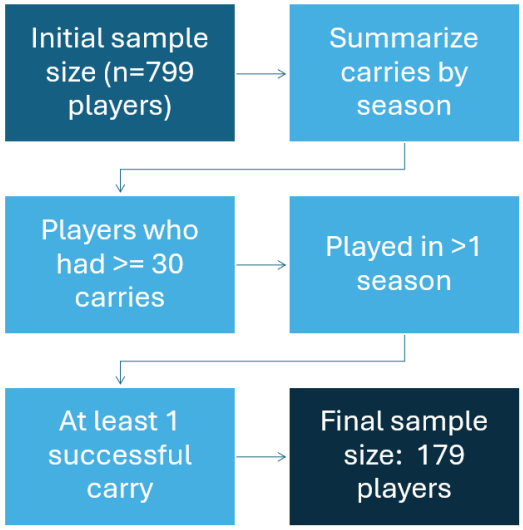
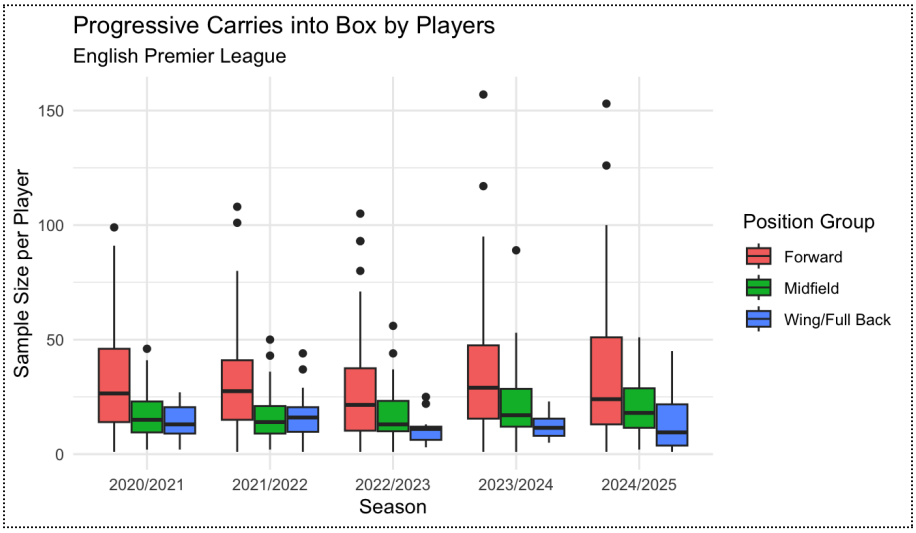
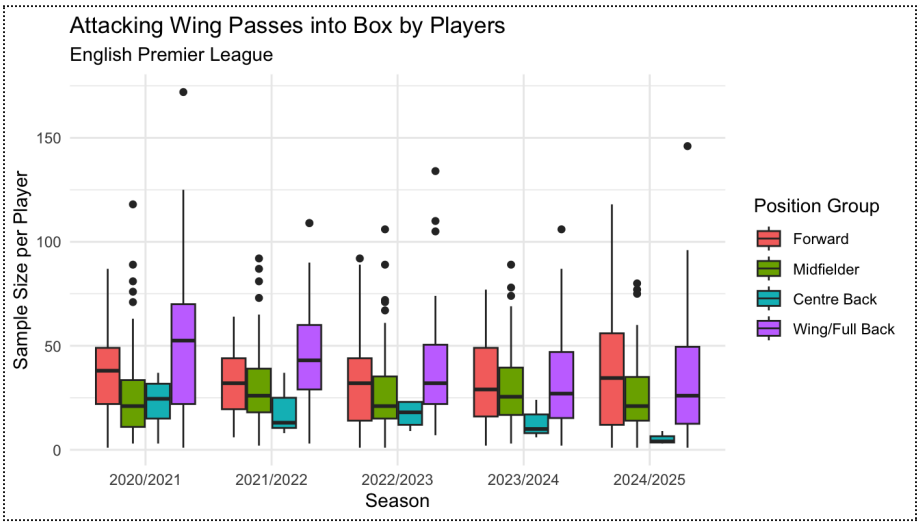
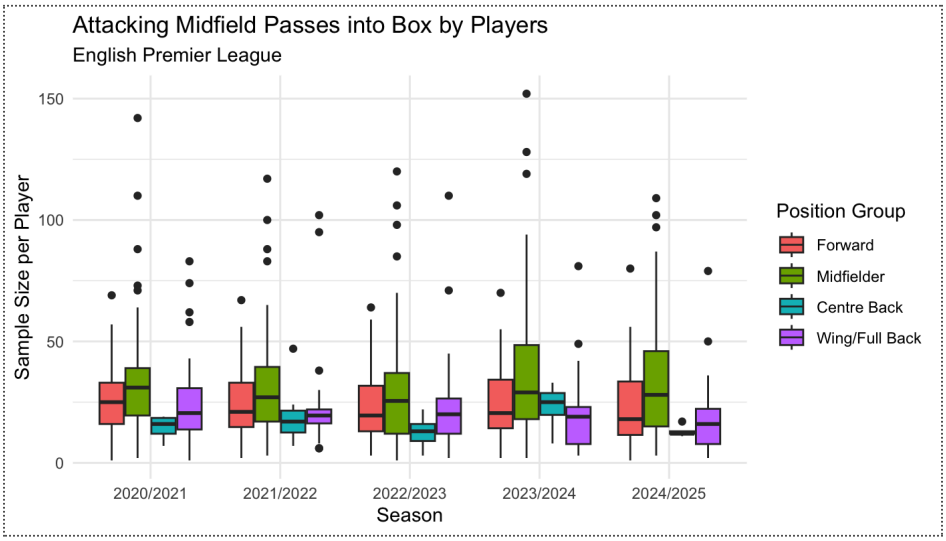


Figure A2 Figure: Over the 5 total seasons worth of data, we started with 23,232 observations from 799 players. We then restricted the dataset to players who played in more than one season, had at least 30 carries, and had at least one successful carry in total. This left us with 17,501 observations from 179 players.


Initial sample size (n=799 players)
Summarize carries by season
Players who had >= 30 carries
Played in >1 season
At least 1 successful carry
Final sample size: 179 players


Progressive Carries into Box by Players
English Premier League
Sample Size per Player
150
100
50
0
2020/2021
2021/2022
2022/2023
2023/2024
2024/2025
Season
Position Group
Forward
Midfield
Wing/Full Back


Attacking Wing Passes into Box by Players
English Premier League
Sample Size per Player
150
100
50
0
2020/2021
2021/2022
2022/2023
2023/2024
2024/2025
Season
Position Group
Forward
Midfielder
Centre Back
Wing/Full Back


Attacking Midfield Passes into Box by Players
English Premier League
Sample Size per Player
150
100
50
0
2020/2021
2021/2022
2022/2023
2023/2024
2024/2025
Season
Position Group
Forward
Midfielder
Centre Back
Wing/Full Back