*Statistica Applicata - Italian Journal of Applied Statistics Vol. 35 (2)*
doi.org/10.26398/IJAS.0035-008

# EVALUATION OF OFF-THE-BALL ACTIONS IN SOCCER

**Wu, Lucas, Swartz, Tim** [1]
*Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby BC, Canada V5A1S6*

***Abstract*** *Whereas there is no shortage of statistics that have been proposed and reported for invasion sports, almost all of the widely reported statistics are based on actions involving the ball. Yet, in football (soccer), it is well-known that players typically possess the ball for less than three minutes during a 90-minute match. In this paper, we develop automatic methods that analyze the activities of players that are "off-the-ball" in soccer. Specifically, a metric is introduced which measures defensive anticipation in soccer. The approach is conceptually straightforward: Using roughly four million spatio-temporal instances, we utilize machine learning techniques to predict the velocity (two-dimensional directional vector and speed) of a defensive player in a given situation. A metric is then developed which compares the player's actual velocity with the predicted velocity of a typical player in this situation. The interpretation of the defensive anticipation metric is based on the tenet that fast is better than slow. The analysis is facilitated through the availability of player tracking data which records the position of players at frequent and regular intervals throughout matches. The metric is calculated for players based on a season of soccer data, where validity and reliability are demonstrated. The metric also conforms to common sense where it is expected and observed that there is a reduction in defensive anticipation as players tire. The proposed approach is applicable and can be tailored to all invasion sports where player tracking data are available.*

***Keywords:*** *OR in sports, big data, machine learning, model validation, player tracking data.*

**ACKNOWLEDGMENT:**

Swartz has been partially supported by the Natural Sciences and Engineering Research Council of Canada. The work was carried out with support from the CANSSI (Canadian Statistical Sciences Institute) Collaborative Research Team (CRT) in Sports Analytics. The authors thank Daniel Stenz, former Technical Director of Shandong Luneng Taishan FC who provided the data used in this paper. We also thank three anonymous reviewers whose comments helped improve the paper.

[1]Corresponding Author: Tim Swartz, tswartz@sfu.ca

## 1. INTRODUCTION

In the sport of football (soccer), it has been estimated that on average, throughout a 90-minute match, individual players have possession of the ball for less than two minutes (Link and Hoernig 2017). Therefore, traditional "on-the-ball" statistics such as goals, tackles, assists, shots and pass completion percentages examine only a snapshot of overall player performance. Encouraged by the "moneyball" phenomena (Lewis 2013), player evaluation via statistical analysis has become widespread across sports (Albert et al. 2017). In particular, there have been significant contributions to the sport of football as described and reviewed by Cefis (2022). For example, Cefis and Carpita (2022) utilize 29 key performance indicators to create composite indicators of performance quality. This paper considers a particular aspect of player evaluation in the context of "off-the-ball" activity in soccer.

This paper introduces novel methods and a metric that evaluates a fundamental defensive objective in soccer, namely defensive anticipation. When a defender anticipates quickly, the defender denies the offensive team both time and space, and this contributes to winning. Defensive awareness is important and is not always recognized. For example, by moving quickly, the defensive player may prevent a valuable pass which is never realized and hence, never recorded. We apply our methods to an actual dataset, where the validity and reliability of the metric are demonstrated.

There are currently no automatic methods (i.e. computer code) that produce metrics for defensive anticipation. For an analyst (e.g. coach) to assess the defensive anticipation of a player, there are two overriding difficulties. First, the analyst would need to monitor the player for the entire 90 minutes of a match, and repeat this over many matches. This is both time consuming and expensive. Second, the analyst would need to objectively evaluate the player's actions, sometimes in contexts where it is not obvious what the player ought to do. The purpose of this paper is to develop automatic methods which objectively evaluate defensive anticipation. With these methods, information on defensive anticipation could be made available for players from various leagues across the world. Therefore, we believe that our methods may be beneficial when teams are seeking a replacement player.

Our investigation is made possible by the availability of player tracking data. Player tracking data in soccer consists of the Cartesian coordinates of the ball and the 22 players on the pitch recorded at regular and frequent time intervals. With player tracking data, we know the locations of all players at all times during

a match, and this facilitates off-the-ball evaluation. Gudmundsson and Horton (2017) provide a review paper on spatio-temporal analyses used in invasion sports (including soccer) where player tracking data are available. The visualization of team formations in soccer is a problem that has received particular attention (Wu et al. 2019). The analysis of player tracking data has also been prominent in the sport of basketball; see for example, Miller et al. (2014).

The study of off-the-ball activity is a new research area of great potential. Historically, a limiting factor for such research has been the availability of tracking data. Tracking data are necessary because we need to know what all players are doing at all times - this is the basis for off-the-ball studies. There have been some off-the-ball analyses in basketball and soccer that are based on the concept of "ghosting" (Lowe 2013, Le et al. 2016, Le et al. 2017 and Seidl et al. 2018). The rationale behind ghosting is that there are optimal and expected paths for defensive players. In the ghosting work (which is proprietary), a main contribution is the claim that if defensive players can replicate the optimal ghosting paths, then outcomes would improve for the defensive team in terms of lower expected points/goals by the offensive team. Also, coaches may be able to assess what-if scenarios. That is, if a given play is drawn up, the expected ghost paths may indicate how the defensive team ought to respond. In the ghosting approach, actual match sequences are studied from a given frame where observed defensive positions are established. Then time frames are allowed to advance where the offensive players continue on their observed path and the ghosts react to the offensive movement. A limitation is that in real matches, offensive players move and react according to the defence. Therefore, the offensive movements that were observed cannot be utilized as responses to the ghosting paths. Spearman (2018) also used tracking data to investigate off-the-ball activity through positioning. Goal scoring probabilities were estimated at player locations using expected goal (xG) considerations and the probabilities of making successful passes to the player locations. This interesting line of research is instructive in identifying optimal positioning from an offensive perspective.

A major challenge in off-the-ball research is the evaluation of actions. Our approach is conceptually straightforward: Using roughly four million spatio-temporal instances, we utilize machine learning techniques to predict the velocity (two-dimensional directional vector and speed) of a defensive player in a given situation. A defensive anticipation metric is then developed which compares the player's actual velocity with the predicted velocity of a typical player in this situation. The interpretation of the defensive anticipation metric is based on the tenet

that fast is better than slow (Blank 2012). Of course, "playing fast is better than slow" is a general principle that may not apply absolutely in every situation. Players that excel in this trait may be thought of as energetic and quick-thinking, and they provide a particular benefit to teams. Importantly, this type of analysis is amenable to other invasion sports for which tracking data are available.

In Section 2, we describe the dataset. In Section 3, we develop the methods used to evaluate defensive anticipation. The work is highly computational and we describe our approach which is based on the use of a tree-based boosting algorithm. In Section 4, the methods are then applied to an analysis of players from the Chinese Super League where validity and reliability of the approach are demonstrated. We conclude with a short discussion in Section 5.

## 2. DATA

Our data consists of matches from the 2019 season of the Chinese Super League (CSL). The league involved 16 teams where each team played every opponent twice, once at home and once away. From these potential 240 matches, we have three missing matches.

From these 237 matches, event data and tracking data were collected independently where event data consists of occurrences such as tackles and passes, and these were recorded along with auxiliary information whenever an "event" takes place. The events were manually recorded by technicians who view film. Both event data and tracking data have timestamps so that the two files can be compared for consistency. For example, the times for the event data are recorded to two decimal points whereas the tracking data times are recored to one decimal point. We dealt with this by rounding the times for the event data. In the CSL dataset, tracking data were obtained from video and the use of optical recognition software. The tracking data consists of roughly one million rows per match measured on 7 variables where the data are recorded every 1/10th of a second. The 7 variables are the time, the $x$ coordinate, the $y$ coordinate, a player identifier, the player jersey number, an indicator variable for the ball, and the half of play. The data were collected by Stats Perform which is a leading service provider of tracking data and operates the Opta data platform. The Stats Perform cameras are high resolution and are placed in various locations for the extrapolation of 2-d pictures to 3-d images. The accuracy and reliability of optical tracking data is high (Mara et al. 2017). Wu and Swartz (2022) have investigated the accuracy of the Stats Perform data in the context of player velocities in soccer. Each row of tracking data corresponds to a particular player at a given instant in time. Therefore, we

have a big data problem where both event data and player tracking data are available based on 237 regular season matches. Although the inferences gained via our analyses are specific to the CSL, we suggest that the methods are applicable to any soccer league which collects tracking data.

## 3. METHODS

### 3.1. Rationale of the Approach

Consider a defender at a particular instant in time during a match. Our approach begins with the prediction of a velocity vector $(\hat{y}_1, \hat{y}_2)$ for the defender. It is important to emphasize that the two-dimensional velocity vector contains both a directional component and magnitude (i.e. speed). The prediction is facilitated through the availability of tracking data associated with the 2019 season of the CSL. With this massive dataset, there exist "similar" circumstances in a spatio-temporal sense to the spatio-temporal state of the defender whose velocity that we are attempting to predict. For example, the defenders in each circumstance may have an opponent with the ball directly in front of them and who is stationary. Therefore, the prediction represents the velocity (i.e. speed and direction) of a typical player in the situation of interest. Of course, the observed velocity $(y_1, y_2)$ of the defender will not be exactly the same as the predicted velocity $(\hat{y}_1, \hat{y}_2)$. The observed velocity is calculated from the tracking data as the change in location given by the positional coordinates by the change in time over a short time period (Wu and Swartz 2022). The change in location is calculated as Euclidean distance. We posit that the defender will have performed above average if they move quicker than predicted in the predicted direction. The quantification of performance is formalized in Section 3.4. The desirability of moving quickly is a tenet of many sports, including soccer, and is discussed in Chapter 1 of Blank (2012).

### 3.2. Prediction of Velocities

Given a snapshot of the spatio-temporal state of the match which includes player locations, player velocities, possession and the location of the ball, it is possible for subject matter experts to predict where players ought to move. However, such assessments are subjective. Alternatively, formulating a parametric predictive model is a formidable task due to the complexity of spatio-temporal configurations.

A rationale for machine learning methods in prediction is that complex phenomena are often difficult to model explicitly. We may have a response variable $y$ and a high-dimensional explanatory vector $x = (x_1, x_2, \ldots, x_k)$ where we have

little apriori knowledge about the relationship between $y$ and $x$. For example, the relationship may only involve a subset of the variables $x$, the components of $x$ may be correlated, and most importantly, the relationship $y \approx f(x)$ involves an unknown and possibly complex function $f$. In addition, the stochastic aspect of the relationship is typically unknown and big data sets may introduce computational difficulties.

In our problem, we face all of the challenges mentioned above. The response variable $y$ is the velocity (speed and direction) that a player moves in a specific off-the-ball situation. We emphasize that $y$ is a two-dimensional response. The explanatory variable $x$ is the state of the match as described by the player tracking data. The idea is that the state of the match $x$ is predictive of movement $y$. For each observation $(x,y)$, the covariate $x$ and the response $y$ are each measured at a specific point in time $t$.

A restriction that we introduce is that we consider off-the-ball actions only for defensive players. Whereas offensive reactions are also important, we find this to be a more challenging prediction problem since offensive players may choose from multiple potential paths.

A first step in the data analysis is the determination of ball possession which then defines the defensive and offensive teams. In addition to player tracking data, we are also provided with tagged event data that provides the timing of passes, dribbles, shots, etc. A possession is retained if the same team maintains the control of the ball by either passing, dribbling or attempting a shot, and the possession ends when the opponent gains control of the ball, a penalty occurs, the ball goes out of bounds, etc. For a pass, the team that made the pass is deemed to be in possession until the ball is intercepted.

To make the prediction problem more tractable, we introduce two data reductions. First, we analyse match states every $\varepsilon = 1$ seconds. This is a tremendous data reduction (reduction by a factor of 10) since tracking data are recorded every 1/10th of a second. However, over a 90-minute match this still leaves us with 5,400 potential observations per player per match. With 11 defensive players on the pitch and the 237 regular season matches, this provides us with over 14 million records. We view $\varepsilon > 0$ as a tuning parameter which we can increase or decrease to adjust the total number of observations. The data reduction is advantageous in the sense that player actions are essentially independent for larger values of $\varepsilon$. In soccer, a player's objectives at a given point in time are different and independent from his objectives $\varepsilon$ seconds later for sufficiently large $\varepsilon$. Our intuition is that player options change considerably over states separated by $\varepsilon \geq 1$ second.

Another data reduction decision involves the covariate vector $x$ provided by the tracking data. Based on our soccer knowledge, we posit that a player's actions are mostly dependent on the spatio-temporal characteristics of the ball and the players within their immediate vicinity. Of course, there are long passes in soccer, but we exclude these considerations as they are the exception rather than the rule. We therefore introduce the following covariates for a given defensive player in a particular state:

- $x_1$ - location of the player (2-dim)
- $x_2$ - player velocity at time $t - \Delta$ (2-dim)
- $x_3$ - location of the ball (2-dim)
- $x_4$ - distance of the player to the ball (1-dim)
- $x_5$ - distance of the player to offensive goal (1-dim)
- $x_6$ - angle of the player to offensive goal (1-dim)
- $x_7$ - location of the goalkeeper (2-dim)
- $x_8$ - distance of the player to goalkeeper (1-dim)
- $x_9$ - indicator for player on offensive or defensive side of the field (1-dim)
- $x_{10}$ - indicator for player belonging to the home or away team (1-dim)
- $x_{11}$ - seconds remaining in the half of the game (1-dim)
- $x_{12}$ - seconds remaining in the full game (1-dim)
- for each of the player's three nearest teammates:
  - $x_{13}$ - location of the teammate (2-dim)
  - $x_{14}$ - velocity of the teammate (2-dim)
  - $x_{15}$ - distance of the player to teammate (1-dim)
  - $x_{16}$ - distance of the ball to teammate (1-dim)
  - $x_{17}$ - relative angle of the teammate to the player of interest (1-dim)
- for each of the player's three nearest opponents:

- $x_{18}$ - location of the opponent (2-dim)
- $x_{19}$ - velocity of the opponent (2-dim)
- $x_{20}$ - distance of the player to opponent (1-dim)
- $x_{21}$ - distance of the ball to opponent (1-dim)
- $x_{22}$ - expected possession value $EPV$ of opponent (1-dim)
- $x_{23}$ - relative angle of the opponent to the player of interest (1-dim)

Therefore, even though we have dramatically reduced the dimensionality of the tracking data, we have retained a 61-dimensional covariate which we hope captures the main drivers of how a player responds in a given situation. We note that the covariates contain a great amount of information which is related to $y$ in complex ways. For example, if a player is close to goal, they may behave differently than if they are near midfield. Also, the movements and space of nearby players naturally impact decisions. We experimented with different numbers of nearest teammates and opponents (i.e. covariates $x_{13}$ through $x_{23}$). However, we found little improvement in prediction beyond using the three nearest teammates and opponents.

The variable $x_2$ and the associated tuning parameter $\Delta \geq 0$ require additional discussion. We cannot include $x_2$ as a covariate with $\Delta = 0$ as this would render $y = x_2$ at all times $t$, and consequently, any fitting algorithm would yield the useless prediction $\hat{y} = y$. That is, our predicted velocity would not be a typical velocity given the circumstances, but instead, the observed velocity of the player of interest. However, the observed velocity $y$ of the player of interest at time $t$ depends on his movement prior to time $t$. For example, if a player is moving forward at speed $s$, it is easier for him to quickly transition to speed $s+\delta$ moving forward than speed $s+\delta$ moving backward. In summary, we ought to know about a player's movement before time $t$ as this impacts movement at time $t$. In Section 3.3, we investigate the selection of $\Delta$.

The expected possession value $EPV$ (feature $x_{22}$) was made publicly available by Shaw (2019). Given the spatial state of a match, $EPV$ provides a measure of the attacking value of each location on the field. We modify the $EPV$ covariate of a player by setting it equal to zero if the offensive player is offside. This is an important covariate in our analysis since defenders should be cautious of balls being played to high $EPV$ positions.

There is some redundancy in our covariates. For example, if we know the Cartesian coordinates of two objects, then the distance between these two objects is a function of their positions. However, to assist the machine learning algorithm

of Section 3.3, we provide some of these derived covariates. We have limited the covariates to the three nearest teammates and three nearest opponents. In most cases, these are the players who most influence the movement of the player of interest. The player of interest cannot intervene in locations that are too distant.

### 3.3. Computational Overview

Recall that our fundamental problem is the development of a metric for defensive anticipation. This metric requires the prediction of the velocity $y$ corresponding to the features $x$ which describe the temporal-spatio state of the match. We constructed a design matrix where we stepped through each $\varepsilon = 1$ seconds of time over all matches to determine team possession. If the time $t$ is part of a possession sequence, then one row of the design matrix is generated for each defender. The columns consisted of all of the features $x$ for a defender as described in Section 3.2. The procedure resulted in a design matrix with 3,770,289 rows and required just under 100 hours of computation on a laptop computer. Note that the construction of the design matrix consists of tasks that can be divided according to matches. Therefore, this data management component is amenable to parallel processing.

For the prediction problem, we used a fast and efficient gradient boosting model, LightGBM, which is based on tree-based learning algorithms (Ke et al. 2017). We trained two LightGBM models, one for predicting the horizontal velocity component and one for predicting the vertical velocity component based on the field orientation. We partitioned the 20-week data into training and test datasets, where the training data included all the even weeks (eg. weeks $2,4,\ldots,20$) and the test data included all the odd weeks (eg. weeks $1,3,\ldots,19$). For model training, we used leave-one-week-out cross validation to select the best tuning parameters which minimized the mean absolute error of the response variables. The training procedure using LightGBM required approximately 1.5 hours of running time on a laptop computer. The LightGBM procedure assumes independent data which should approximately be the case with player velocities measured at one second intervals.

Recall that we are interested in setting the parameter $\Delta \geq 0$ which provides the velocity covariate $x_2$ of the player of interest at time $t - \Delta$. We want to set $\Delta$ so that it assists prediction of the velocity of a typical player at time $t$. Again, $\Delta$ cannot equal zero; otherwise we simply obtain predictions that are the actual velocities of the players of interest. On the other hand, if we choose large $\Delta$, then the velocity covariate $x_2$ is too distant in time from the current time $t$ to facilitate

the prediction of velocity at time $t$. In Figure 1, we plot the correlation of the predicted speed at time $t$ and the actual speed at time $t - \Delta$. For $\Delta = 0$ seconds, the correlation is perfect (i.e. $r = 1$) as expected. We wish to choose a time lag $\Delta$ so that the model provides a good but not a perfect predictor. From Figure 1, we choose the tuning parameter $\Delta = 0.5$ seconds where the correlation $r \approx 0.9$. We experimented with other choices of $\Delta$ up to $\Delta = 1$ second, and found that our results were robust to the choice. The fitted model from LightGBM provides a mean absolute error of 0.319 m/sec in the x-coordinate velocity and 0.398 m/sec in the y-coordinate velocity.

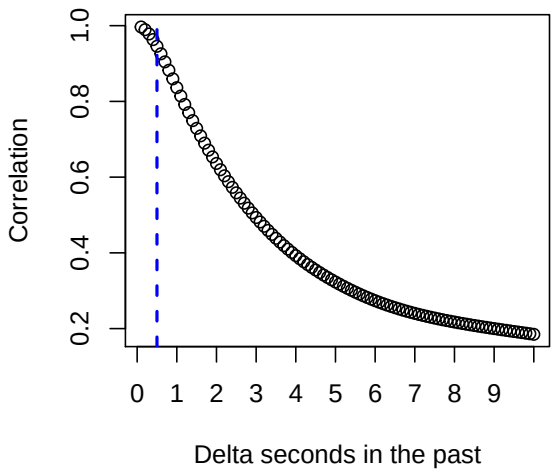


**Figure 1: Correlation of predicted speed at time $t$ and actual speed at time $t - \Delta$ where time is measured in seconds. The dashed line corresponds to the selected value $\Delta = 0.5$ seconds.**

### 3.4. Derivation of a Metric for Defensive Anticipation

We return to the motivation for off-the-ball player evaluation. Recall that the core idea from Chapter 1 of Blank (2012) is that doing things quickly in soccer is better. For example, one could imagine a defender moving towards a forward who is about to receive a pass. In this case, getting there early increases the chance of intercepting the pass or preventing the forward from creating a goal

scoring opportunity. Now, there are many instances during a match where moving quickly makes no sense (e.g. the ball is at the opposite end of the field). In cases where the predicted response is to move slowly, we will not use these cases for the purpose of player evaluation.

Consider time $t$ where the match state $x$ is recorded and the predicted velocity for a defensive player is $(\hat{y}_1, \hat{y}_2)$. Recall that velocity is two-dimensional as it involves both speed and direction in the plane. Again, we only consider observations where the predicted speed exceeds a specified threshold speed. The predicted velocity $(\hat{y}_1, \hat{y}_2)$ is obtained by the machine learning prediction methods of Section 3.2. We let $(y_{\text{obs1}}, y_{\text{obs2}})$ denote the corresponding observed velocity of the player under evaluation. Then, we define the player's off-the-ball performance at time $t$ by

$$p = \begin{cases} \left( \sqrt{v_1^2 + v_2^2} - \sqrt{\hat{y}_1^2 + \hat{y}_2^2} \right) / \sqrt{\hat{y}_1^2 + \hat{y}_2^2} & v_1 \hat{y}_1 \geq 0 \\ \left( -\sqrt{v_1^2 + v_2^2} - \sqrt{\hat{y}_1^2 + \hat{y}_2^2} \right) / \sqrt{\hat{y}_1^2 + \hat{y}_2^2} & v_1 \hat{y}_1 < 0 \end{cases} . \tag{1}$$

A geometric interpretation of $p$ is provided in Figure 2. The statistic $p$ in (1) is based on the projection $(v_1, v_2)$ of the observed performance $(y_{\text{obs1}}, y_{\text{obs2}})$ onto the velocity line defined by the predicted velocity $(\hat{y}_1, \hat{y}_2)$. The line $k(\hat{y}_1, \hat{y}_2)$ for $k > 0$ emanates from the origin and is given by $y_2 = (\hat{y}_2/\hat{y}_1)y_1$ and the projection is calculated by $(v_1, v_2) = ((\hat{y}_1^2 y_{\text{obs1}} + \hat{y}_1 \hat{y}_2 y_{\text{obs2}})/(\hat{y}_1^2 + \hat{y}_2^2),\ (\hat{y}_1^2 \hat{y}_2 y_{\text{obs1}} + \hat{y}_1 \hat{y}_2^2 y_{\text{obs2}})/(\hat{y}_1^3 + \hat{y}_1 \hat{y}_2^2))$. Therefore, "good" performance according to (1) takes into account moving quicker in the predicted direction. Longer projections on the yellow velocity line in Figure 2 are better in terms of player performance, and lead to larger values of $p$. Values of $p > 0$ are interpreted as above average performance and values of $p < 0$ are interpreted as below average performance.

The player's season long performance is then given by the defensive anticipation metric

$$P = \left( \frac{1}{N} \sum_{i=1}^{N} p_i \right) 100\% \tag{2}$$

where the summation is taken over all instances where the predicted velocity exceeds the threshold speed and the index $i = 1, \ldots, N$ corresponds to the cases involving the player during the season. In (2), it is sensible to calculate an average since the observations can be regarded as independent due to the time spacing provided by the tuning parameter $\varepsilon = 1$ seconds introduced in Section 3.2.

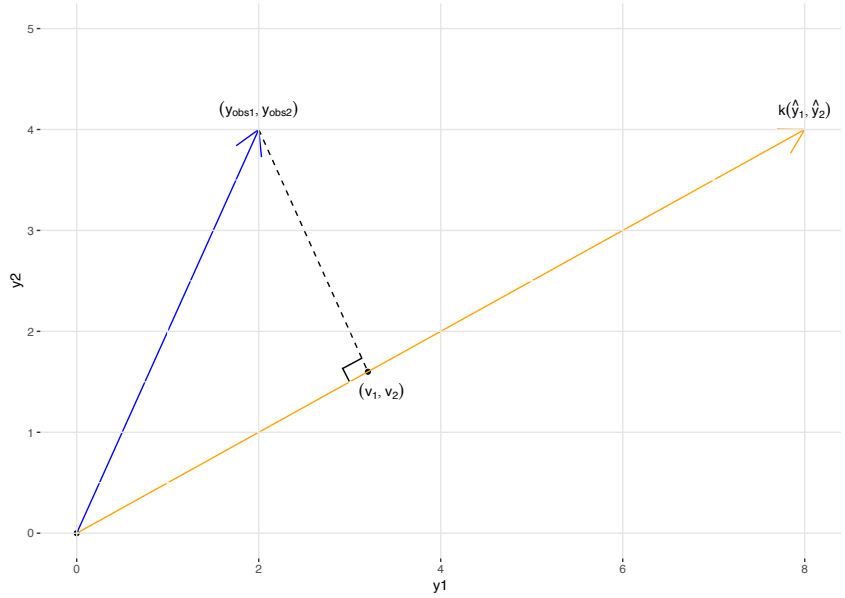


**Figure 2: Geometric diagram which illustrates the components of the statistic $p$ in equation (1). Imagine a player who is located at the origin $(0,0)$. The observed velocity of the player is shown by the blue vector pointing towards $(2,4)$. The predicted velocity of an average player is shown by the yellow vector pointing towards $(8,4)$. The perpendicular line indicates the projection of the observed velocity vector on the predicted velocity vector. Using equation (1), the defensive anticipation value, $p$, is equal to $-0.6$, which can be interpreted as a 60% reduction compared to the average player.**

We can think of the metric (2) as a measure of defensive anticipation. It also possibly assesses aspects of player energy and quickness of thought. The multiplicative factor 100% in (2) permits a nice interpretation; a $P$-score of $+x$ describes a player whose defensive anticipation is $x\%$ above the average player whereas a $P$-score of $-x$ describes a player whose defensive anticipation is $x\%$ below the average player. The reported scores for players (Table 2) are relative to players from the 2019 season of the CSL. Although the metric $P$ is unbounded by it's construction, it appears that reported values are surely contained in the interval $(-10,10)$.

## 4. RESULTS AND ASSESSMENT

Of course, with new metrics such as defensive anticipation, there is no truth against which results can be compared. For example, we simply don't know which players are best at defensive anticipation. In this section, we look at the defensive anticipation metric from various angles with an attempt to establish validity and reliability.

First, to get a sense of the prediction results, Figure 3 provides a plot of the predicted velocities and the observed velocities for all 20 players on the field (not including the keepers) at a given instant in time. In most cases, the predicted and observed velocity vectors tend to point in roughly the same direction. For illustration, consider defensive player #16. His movement is directly towards the ball. However, the model predicts that he ought to move a little bit more towards his own goal at roughly the same speed. The predicted movement may be viewed as cautious and preferable since offensive player #35 (who is in possession) is moving downfield and may pose a risk. We observe that for some players, the velocity vectors are short, and this suggests that little is happening in their immediate surroundings. For example, defensive player #7 is barely moving, and this appears sensible as there are no threatening offensive players in the vicinity. For the evaluation of the defensive anticipation metric (2), we removed observations for which the predicted speed $\sqrt{\hat{y}_1^2 + \hat{y}_2^2}$ is less than the threshold speed of 0.20 m/sec which corresponds to 0.72 km/hour. In the test dataset, 1.8% of the observations were removed due to the threshold constraint.

Based on the examination of many frames such as given in Figure 3, we did not find predicted velocities that contradicted our soccer intuition. This provides an indication that in a given situation, the predicted velocity of a typical player is sensible. This may be expected because the predicted velocity is based on the fitting of a massive dataset, where on average, professional athletes make good decisions. We note that the model was assisted by the inclusion of covariate $x_{22}$ (previously discussed). The recognition of players in offside positions improved prediction.

### 4.1. Reliability

With respect to a metric, *reliability* refers to the consistency of the measure. In other words, reliability addresses reproducibility. For example, it would be undesirable if our defensive anticipation metric (2) identified a player as having great defensive anticipation for half of the matches and terrible defensive anticipation in the other matches. Since we expect some consistency in professional athletes,

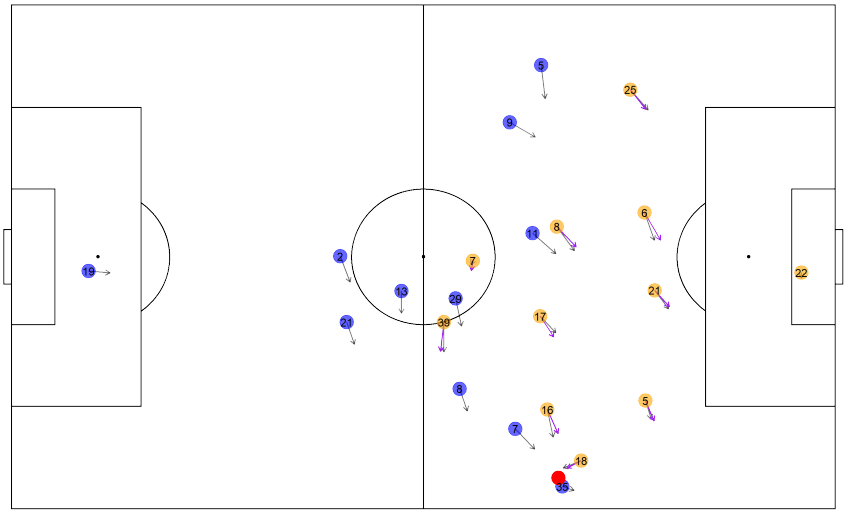


**Figure 3: Plot of predicted velocities (purple arrows) and observed velocities (black arrows) at a given instant in time. The blue team is in possession, the yellow team is defending and the red dot corresponds to the ball.**

this would suggest that there is little value in the metric.

To investigate this, we divided the 2019 CSL season into even and odd weeks. The premise is that the metric (2) measures an aspect of playing style, and that style should not differ greatly between the two sets of weeks. In Table 1, we provide results for the 10 players on Shandong Luneng for whom the number of instances $N > 10,000$ in (2) for both sets of weeks. Shandong Luneng is an interesting CSL team as two of the international players (Fellaini and Pelle) are well known to those who follow the English Premier League. We observe that there is consistency in the player metrics across the two sets of weeks. In fact, the ranks of the 10 players are the same across the two weeks. This suggests that the defensive anticipation metric (2) is reliable and is capturing an aspect of playing style. The standard errors for $P_{\text{even}}$ and $P_{\text{odd}}$ are small for all players, lying between 0.31 and 0.55.

| Player | $N_{\text{even}}$ | $N_{\text{odd}}$ | $P_{\text{even}}$ (rank) | $P_{\text{odd}}$ (rank) |
|---|---|---|---|---|
| Marouane Fellaini | 17,146 | 17,340 | 2.8 (1) | 2.4 (1) |
| Zhang Chi | 16,647 | 17,235 | 2.4 (2) | 2.0 (2) |
| Liu Yang | 19,556 | 19,845 | 1.6 (3) | 1.8 (3) |
| Wang Tong | 13,955 | 20,034 | 0.4 (4) | 0.2 (4) |
| Hao Junmin | 16,050 | 16,696 | -0.3 (5) | -1.4 (5) |
| Zheng Zheng | 14,582 | 10,849 | -1.6 (6) | -2.6 (6) |
| Dai Lin | 14,030 | 18,423 | -2.1 (7) | -3.1 (7) |
| Graziano Pelle | 19,337 | 18,302 | -3.7 (8) | -4.1 (8) |
| Gil | 10,159 | 13,306 | -4.1 (9) | -5.1 (9) |
| Roger Guedes | 14,067 | 16,737 | -5.5 (10) | -5.7 (10) |

**Table 1: The defensive anticipation metric $P$ calculated during even and odd weeks for players on Shandong Luneng during the 2019 season.**

### 4.2. Validity

With respect to a metric, *validity* refers to the accuracy of measure. In our investigation, we are interested whether the metric $P$ in (2) really measures defensive anticipation.

To investigate validity, we first consider the defensive anticipation metric (2) for all 438 outfield players in the CSL dataset. The players are categorized according to the five broad playing positions as follows: wide midfielder ($n = 79$) wide defender ($n = 77$), and forward ($n = 86$), central midfielder ($n = 110$) and central defender ($n = 86$). Density plots of (2) corresponding to each of the playing positions are shown in Figure 4. We observe that there is little difference in (2) across the playing positions. We note that central midfielders have slightly larger values of (2) than other players on average (as might be expected). This may be related to the defensive aggressiveness required at that position. We also observe that there is more variability in (2) amongst the forwards than the other playing positions.

Recall that a difficulty in assessing the validity of the proposed metric (2) is that there is no gold standard for the truth. We do not know with certainty which players play with more and less defensive anticipation (combination of energy and quick-thinking). Therefore, we took the same players from Shandong Luneng as in Table 1, and ranked these players according to their $P$-scores (2) from the entire 2019 season. The results are provided in Table 2. In Table 2, we made compar-

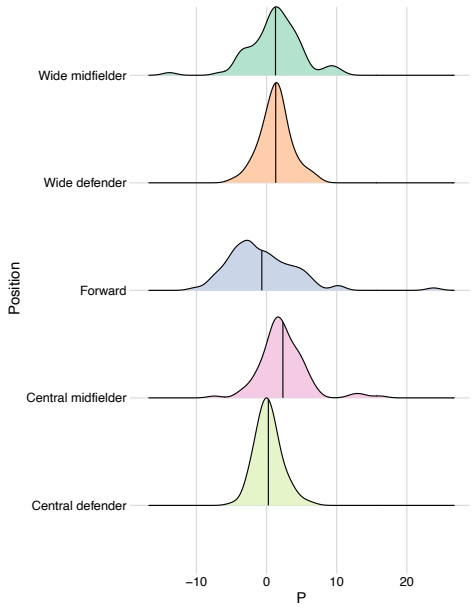


**Figure 4: Density plots of (2) based on playing position. For each player, the defensive anticipation metric (2) was calculated for all matches in the 2019 CSL season. We observe that central midfielders have slightly larger defensive anticipation values than other players on average, and there is more variability amongst the forwards than the other playing positions.**

isons with various measures of aggression. We provide season long data on fouls, successful tackles and interceptions. We excluded card accumulation as cards are relatively rare events. We observe that the aggressiveness inherent in fouls, successful tackles and interceptions correlates with our defensive anticipation metric. For example, the correlation coefficients between $P$ and these three statistics are 0.58, 0.65 and 0.74, respectively. The corresponding 95% confidence intervals are $(-0.08, 0.88)$, $(0.03, 0.91)$ and $(0.21, 0.93)$, respectively.

In Table 2, we explored the relationship between $P$ with player interceptions and tackles in the context of Shandong Luneng. We expanded this investigation by considering all players in the CSL who had played at least 500 minutes during the 2019 season. Figure 5 provides scatterplots relating $P$ to interceptions and tackles. Of course, we do not expect extraordinarily high correlations between $P$, interceptions and tackles since these measurements consider different aspects of play. What these statistics have in common is an element of aggression. We

| Player | $P$ (rank) | Fouls (rank) | Tackles (rank) | Interceptions (rank) |
|---|---|---|---|---|
| Marouane Fellaini | 2.64 (1) | 46 (1) | 21 (5.5) | 23 (4) |
| Zhang Chi | 2.20 (2) | 32 (2.5) | 21 (5.5) | 29 (2) |
| Liu Yang | 1.71 (3) | 26 (4.5) | 33 (1) | 6 (8) |
| Wang Tong | 0.26 (4) | 15 (9) | 19 (7) | 27 (3) |
| Hao Junmin | -0.85 (5) | 25 (6) | 23 (4) | 22 (5) |
| Zheng Zheng | -1.99 (6) | 17 (8) | 29 (2) | 12 (7) |
| Dai Lin | -2.67 (7) | 32 (2.5) | 24 (3) | 33 (1) |
| Graziano Pelle | -3.91 (8) | 26 (4.5) | 6 (10) | 2 (9.5) |
| Gil | -4.65 (9) | 6 (10) | 13 (8) | 13 (6) |
| Roger Guedes | -5.63 (10) | 21 (7) | 7 (9) | 2 (9.5) |

**Table 2: The defensive anticipation metric $P$ given by (2) for 10 players on Shandong Luneng who received the most playing time during the 2019 CSL season. We also provide comparison metrics involving aggression during the 2019 season, namely the total number of fouls committed, tackles made and the number of interceptions.**

observe that interceptions and tackles correlate positively with $P$ leaguewise.

We investigated the validity of our metric further by calculating the average $P$-score for all CSL players where we divided matches into 10-minute intervals. The plot is provided in Figure 6. We observe that $P$ decreases as the match progresses. Since players tire as the game proceeds (both physically and mentally), it makes sense that our metric (2) decreases. There appears to be a big drop after the 70-th minute of the match.

It is interesting that amongst CSL players with regular minutes, the two players with the highest $P$-scores are Chang Feiya of Wuhan Zall ($P = 5.71$) and Yang Shiyuan of Shanghai SIPG ($P = 5.33$). Feiya is primarily a midfielder and does not have remarkable statistics; he scored only one goal in the 2019 season. Interestingly, the website https://www.allfamousbirthday.com/chang-feiya/ describes Feiya as one of the most popular Chinese football players. Shiyuan is a midfielder who also does not have remarkable statistics; he did not score during the 2019 season. Interestingly, the website https://www.whoscored.com/Players/143864/Show/Yang-Shiyuan describes Shiyuan as a player who likes to tackle and commits fouls often.

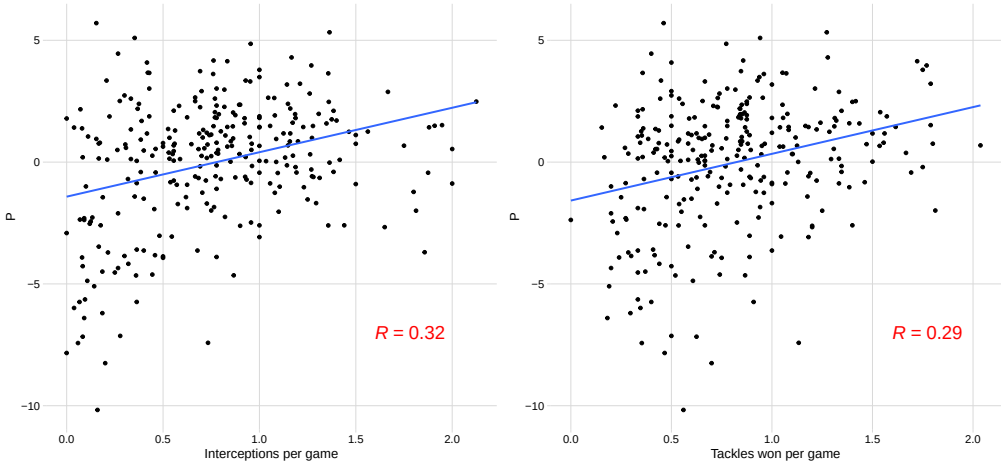


**Figure 5: Scatterplots of the defensive anticipation metric (2) plotted against player interceptions and tackles made during the 2019 CSL season.**

## 5. DISCUSSION

We have introduced an important and seminal area of research where automatic and objective methods have been developed to assess a particular defensive characteristic of off-the-ball behaviour. We have referred to the proposed metric (2) as defensive anticipation. The methods can be adapted to any invasion sport where tracking data are available.

The evaluation of off-the-ball performance is viewed in a narrow context where fast is considered better than slow. Even if speed is not the ultimate metric in off-the-ball evaluation, the metric (2) developed here may uncover insights into aspects of play. Perhaps players with high evaluations may be thought of as "high motor" players whose skills are useful to teams. An important aspect of the research is that our metric measures aspects of industry, laziness, anticipation and quick-thinking; these are characteristics that have not been previously quantified.

Some other notable aspects of our work include the following: the proposed metric is seen as reliable in the sense that it truly captures intrinsic player tendencies (Table 1), the metric adheres to expected results such as the positive correlation between the metric and other statistics related to aggression (Figure 5), and decreasing defensive anticipation as players tire (Figure 6).

A possible application involves the evaluation of $P$ on a game by game basis. Managers would like to know how players have performed in terms of defensive

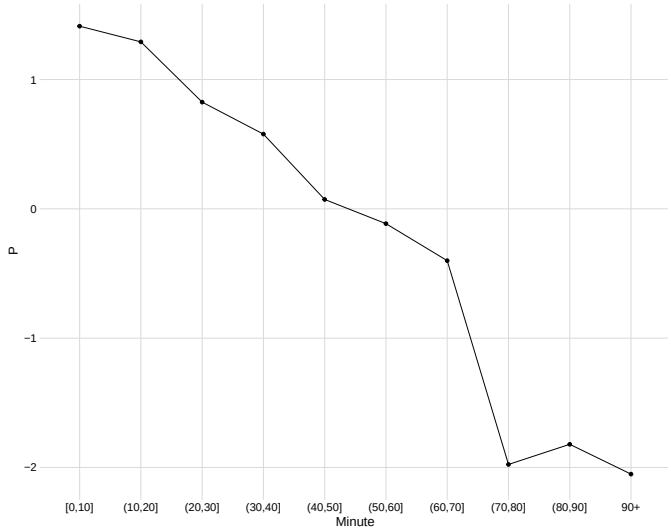


**Figure 6: Plot of the defensive anticipation metric (2) averaged over all CSL players during 10-minute intervals.**

anticipation. For example, perhaps *P*-scores may form a reason for future inclusion in the lineup. Also, it may be possible to detect the effects of illness.

### 5.1. Connections to Existing Literature

Whereas there does not seem to be any previous work on defensive anticipation in soccer, there are various recent papers that attempt to assess off-the-ball performance. Of course, this is a relatively new research topic since tracking data has only recently become available. We discuss two substantive contributions.

Dick and Brefeld (2019) use a reinforcement learning approach to evaluate player positioning. Unlike our investigation that has a defensive focus, Dick and Brefeld (2019) are interested in offensive configurations that are more likely to lead to goals. The approach takes into account current positioning and movement vectors that enable the consideration of future formations. A scoring function is learned from past data that maps game states to values that assess the benefit to the attacking team. Also, in contrast to our work, the proposed measures correspond to the team level rather than the performance of individual players.

In his PhD thesis, Fernández (2022) develops a framework for the investigation of various problems in soccer. The comprehensive approach is predicated on the development of an expected possession value model that decomposes the sport

into components. For example, the action space is composed of passes, drives and shots where each component has its own set of estimation procedures. Applications in the thesis which are related to our work concern off-the-ball performance. For example, in Chapter 7, Fernández (2022) provides insights as to how teams can defend against buildup play, and how to calculate a player's optimal offensive positioning.

### 5.2. Future Research

There are at least three avenues for future research. First, there are many alternative predictive models that could be investigated. Given that we are predicting average player movement, the determination of which model provides better predictions is not straightforward. Second, the approach only considers off-the-ball actions for players when the opponent has possession. Naturally, player movement for players on the team in possession is also important. This is a more difficult prediction problem since there appears to be more viable options for offensive players. Third, we hope to gain access to tracking data from other leagues. Assessing the proposed metric and evaluating a wider pool of players with respect to defensive anticipation are topics of great interest.

## 6. REFERENCES


Albert, J.A., Glickman, M.E., Swartz, T.B. and Koning, R.H., Editors (2017). *Handbook of Statistical Methods and Analyses in Sports*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, Boca Raton.

Blank, D. (2012). *Soccer IQ*, www.soccerpoet.com

Cefis, M. (2022). Football analytics: A bibliometric study about the last decade contributions. *Electronic Journal of Applied Statistics*, 15(1), 232-248.

Cefis, M. and Carpita, M. (2022). The higher-order PLS-SEM confirmatory approach for composite indicators of football performance quality. *Computational Statistics*, https://doi.org/10.1007/s00180-022-01295-4

Dick, U. and Brefeld, U. (2019). Learning to rate player positioning in soccer. *Big Data*, 7, 71-82.

Fernández, J. (2022). A framework for the analytical and visual interpretation of complex spatiotemporal dynamics in soccer. Department of Computer Science, Polytechnic University of Catalonia. Accessed March 16, 2022 at https://upcommons.upc.edu/handle/2117/363073

Fernández, J., Bornn, L. and Cervone, D. (2019). Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer. *13-th MIT Sloan Sports Analytics Conference*, Accessed September 21, 2020 at http://www.lukebornn.com/papers/fernandez_sloan_2019.pdf

Gudmundsson, J. and Horton, M. (2017). Spatio-temporal analysis of team sports. *ACM Computing Surveys*, 50(2), Article 22.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146¿3154.

Le, H.M., Carr, P., Yue, Y. and Lucey, P. (2016). Data-driven ghosting using deep imitation learning. *10-th MIT Sloan Sports Analytics Conference*, Accessed November 30, 2020 at https://global-uploads.webflow.com/5f1af76ed86d6771ad48324b/5fee0a8b9838792227ec7fa5_Data-Driven

Le, H.M., Yue, Y., Carr, P. and Lucey, P. (2017). Coordinated multi-agent imitation learning. *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia.

Lewis, M. (2013). *Moneyball: The Art of Winning an Unfair Game*, WW Norton, New York.

Link, D. and Hoernig, M. (2017). Individual ball possession in soccer, *PLoS ONE*, 12(7): e0179953. https://doi.org/10.1371/journal.pone.0179953

Lowe, Z. (2013). Lights, cameras, revolution. *Grantland*, Accessed August 25, 2020 at https://grantland.com/features/the-toronto-raptors-sportvu-cameras-nba-analytical-revolution/

Mara, J., Morgan, S., Pumpa, K. and Thompson, K. (2017). The accuracy and reliability of a new optical player tracking system for measuring displacement of soccer players. *International Journal of Computer Science in Sport*, 16(3), 175-184.

Miller, A., Bornn, L., Adams, R.P. and Goldsberry, K. (2014). Factorized point process intensities: A spatial analysis of professional basketball. In *Proceedings of the 31st International Conference on Machine Learning* - Volume 32, JMLR.org, Beijing, 235-243.

Seidl, T., Cherukumudi, A., Hartnett, A., Carr, P. and Lucey, P. (2018). Bhostgusters: Realtime interactive play sketching with synthesized NBA defenses. *12-th MIT Sloan Sports Analytics Conference*, Accessed August 25, 2020 at http://www.sloansportsconference.com/wp-content/uploads/2018/02/1006.pdf

Shaw, L. (2019). Friends-of-Tracking-Data-FoTD/LaurieOnTracking Accessed November 20, 2021 at https://github.com/Friends-of-Tracking-Data-FoTD/LaurieOnTracking

Spearman, W. (2018). Beyond expected goals. *12-th MIT Sloan Sports Analytics Conference*, Accessed September 21, 2020 at http://www.sloansportsconference.com/wp-content/uploads/2018/02/2002.pdf

Wu, L.Y. and Swartz, T.B. (2022). The calculation of player speed from tracking data. *International Journal of Sports Science and Coaching*, https://doi.org/10.1177/17479541221124036

Wu, Y., Xie, X., Wang, J., Deng, D., Liang, H., Zhang, H., Cheng, S. and Chen, W. (2019). ForVizor: Visualizing spatio-temporal team formations in soccer, *IEEE Transactions on Visualization and Computer Graphics*, 25(1), 65-75.