Research Article

Tianyu Guan, Jiguo Cao and Tim B. Swartz*

Parking the bus

https://doi.org/10.1515/jqas-2021-0059
Received July 2, 2021; accepted May 11, 2023;
published online May 30, 2023

Abstract: This paper explores defensive play in soccer. The analysis is predicated on the assumption that the area of the convex hull formed by the players on a team provides a proxy for defensive style where small areas coincide with a greater defensive focus. With the availability of tracking data, the massive dataset considered in this paper consists of areas of convex hulls, related covariates and shots taken during matches. Whereas the pre-processing of the data is an exercise in data science, the statistical analysis is carried out using linear models. The resultant messages are nuanced but the primary message suggests that an extreme defensive style (defined by a small convex hull) is negatively associated with generating shots.

Keywords: betting odds; big data; convex hulls; defensive strategies; tracking data

1 Introduction

Jose Mourinho is one of the most successful and famous managers in European soccer having won trophies at Porto, Chelsea, Inter Milan, Real Madrid and Manchester United. Self-labelled "the Special One", Mourinho has a colorful reputation and has provided the soccer world with many entertaining quotes and expressions. One such expression is the negatively perceived term "parking the bus" where he refers to a playing style that is extremely defensive and unattractive. When a team parks the bus, it is as though a bus is blocking their defending goal, where the players maintain a compact shape and demonstrate little ambition going forward. Parking the bus is a tactic that is sometimes used when a team is leading and is attempting to protect the lead.

There are indications that playing conservatively when leading is not an optimal strategy. Silva and Swartz (2016) investigate the problem of optimal substitution times in soccer, and as a by-product of their analysis, they observe that teams that are leading in a soccer match are more likely to have the next goal scored against them than if the match is tied. In the National Hockey League (NHL), Figure 2 from Beaudoin, Schulte and Swartz (2016) indicates that the probability of shots on goal by the home team increases as the goal differential in favour of the road team increases. This finding is corroborated by Thomas (2017) who demonstrates that there is an increased probability for tied matches than would be expected by independent Poisson scoring models. When a team is leading in the National Football League (NFL), it is a common occurrence for the team to allow short passes and short runs, especially near the end of a game – they are playing cautiously in the sense that they want to prevent the offense from realizing large gains. This tactic has been questioned, where former coach John Madden once stated "All a prevent defense does is prevent you from winning".

Given that professional sport is big business, and that playing cautiously is a common sporting tactic, it seems that a careful investigation of the consequences of parking the bus is a topic of widespread interest. Although the stakes are lower, the consequences of parking the bus are also relevant to amateur sport. In particular, there are various questions associated with parking the bus including:

- How can you separate the defensive tactic of parking the bus from the offensive tactic of playing aggressively?
- What are the match circumstances that lead to parking the bus?
- How is parking the bus associated with goal scoring?

This paper attempts to address these three questions in the context of soccer (i.e. association football).

Our investigation is made possible by the availability of player tracking data. Player tracking data in soccer consists of the (x, y) 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 and movement

***Corresponding author: Tim B. Swartz**, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A1S6, Canada, E-mail: tim@stat.sfu.ca
Tianyu Guan, Department of Mathematics and Statistics, Brock University, 1812 Sir Isaac Brock Way, St. Catherines, Ontario L2S3A1, Canada
Jiguo Cao, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A1S6, Canada

of all players during a match, and this facilitates the investigation of cautious playing behaviour. 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 is a problem that has received particular attention in soccer (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). For a review of statistical contributions that have been made across major sports, see the text by Albert et al. (2017).

The distinction between a team playing aggressively and its opponent playing cautiously is a primary problem in the assessment of parking the bus. When a team is playing aggressively, the players "press forward" (i.e. move down the field and challenge all passes by the opposition). Hence, the opposition may find themselves predominantly in their defensive end of the field, and it may appear that they are playing defensively. Our analysis is predicated on the assumption that the area of the convex hull formed by the players on a team provides a proxy for defensive style. It is assumed that smaller areas coincide with a greater defensive focus. Consequently, even if one team is playing aggressively and the opposition is forced into their own end, the opposition is not playing defensively if some of their players are spread out, venturing to go forward on attack when they recover the ball. In this case, the area of their convex hull is not small. The area is only small when the players are compact and sitting deep towards their own goal (i.e. parking the bus). In this case, they are playing an extremely defensive style. Convex hulls have been previously utilized in sport. For example, Metulini, Manisera and Zuccolotto (2017) form convex hulls with respect to players on the basketball court. They have used the hulls, visualization techniques and clustering to inform on player movement patterns.

In Figure 1, we illustrate the locations of all players and the convex hull for the defending team. The left plot corresponds to a match in the 65.5th minute between Guangzhou Evergrande Taobao versus Wuhan Zall (defending team) on September 22, 2019. Here, the defending team is compact (i.e. parking the bus), and the area of the convex hull is 490.8 squared metres. The right plot corresponds to a match in the 61.6th minute between Guangzhou R&F versus Guangzhou Evergrande Taobao (defending team) on July 20, 2019. Here, the defending team is more expansive (e.g. two forwards are situated near midfield). In this case, the area of the convex hull is 1159.8 squared metres. The two examples were chosen based on similar circumstances; e.g. the ball is possessed by the attacking team in similar positions on the field. The examples suggest that the defending team makes a choice as to whether they park the bus.

Under the assumption that the area of the convex hull for a team provides a measure of defensive style (i.e. cautiousness), there are various ways to investigate the relationship between cautiousness and soccer outcomes. Betting odds are introduced, and these provide a baseline measurement for the relative strength of the two teams in a match. We consider accessible analyses using linear models.

With player tracking data recorded at 10 frames per second over a 90 min match for two teams, this potentially suggests over 100,000 calculations of areas of convex hulls in a single match. This suggests a big data problem for which computational efficiency is an important consideration.

In Section 2, we describe the dataset in detail. In Section 3, simple analyses using linear models are carried out. We observe that teams are most cautious (i.e. small convex hulls) with one-goal leads, leading teams are more cautious near the end of matches and leading teams are more cautious if they are perceived as the weaker team. Most importantly, teams are more likely to be scored against when they park the bus than when they do not park the bus. We conclude with a brief discussion in Section 4.

In our work, we provide a single numerical statistic (the area of a team's convex hull) as a measure of defensive compactness. However, compactness is just one aspect of playing style. Playing style is a challenging and useful research area as it allows teams to prepare for opponents in the best tactical manner. For example, Fernandez-Navarro et al. (2016)

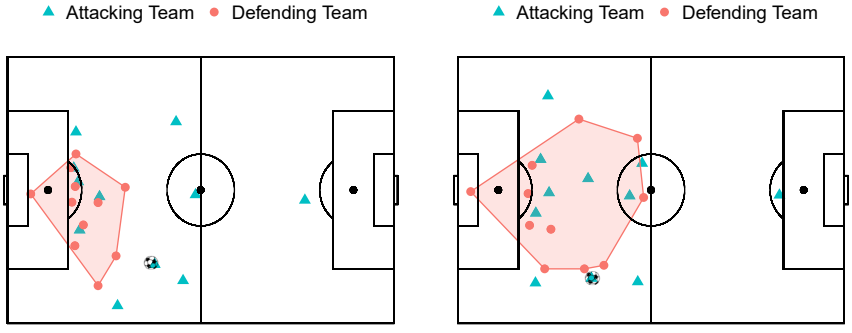

Figure 1: Player locations and the corresponding convex hull for the defending team. The left plot provides an example of a defending team that is compact (i.e. parking the bus) whereas the right plot provides an example of a defending team that is more expansive.

use factor analysis on performance indicators (match summary statistics) to characterize 12 playing styles in Spanish and English soccer. Lago-Peñas, Gómez-Ruano and Gai (2017) carry out similar analyses in the context of Chinese soccer. Fernandez-Navarro et al. (2018) and Gollan, Bellenger and Norton (2020) use various linear models to study the effects of contextual match statistics (e.g. venue, opponent, goal differential, total goals) on variables related to playing style. Ötting, Langrock, and Maruotti (2023) use hidden Markov models to analyze match states in soccer that are associated with momentum changes as characterized via possession.

Network science in soccer is a growing area with contributions that are related to playing style. For example, Buldú et al. (2018) review the various approaches used in the study of passing networks in soccer. The approaches that rely on tracking data are inherently complex due to the spatio-temporal nature of the data and the fact that passing decisions are dependent on the actions of one's opponent. In a particular application, Garrido et al. (2020) examine the consistency of a team to maintain a particular passing style. Two network contributions that do not rely on tracking data include Diquigiovanni and Scarpa (2019) and Gonçalves et al. (2017). Diquigiovanni and Scarpa (2019) develop clustering techniques applied to networks to identify tactical styles in Serie A. Gonçalves et al. (2017) analyse youth soccer where a negative relationship between match outcomes and the over-reliance on a single player in passing networks is determined.

2 Data

The 2019 regular season of the Chinese Super League (CSL) involved a balanced schedule of 240 matches where each of the 16 teams played every opponent twice, once at home and once on the road. We have access to event and tracking data for all games from the 2019 season, except for the following three matches: August 2 – Chongqing Dangdai Lifan SWM versus Dalian Yifang, August 2 – Shanghai Greenland Shenhua versus Wuhan Zall and December 1 – Henan Jianye versus Guangzhou R&F. There appears to be no systematic reason for the three missing matches.

Event data and tracking data are collected independently where event data consists of occurrences such as tackles and passes, and these are recorded along with auxiliary information whenever an "event" takes place. The events are manually tabulated by technicians who view recorded video. Both event data and tracking data have timestamps so that the two files can be compared for internal consistency. There are various ways in which tracking data are collected. One approach involves the use of RFID technology where each player and the ball have tags that allow for the accurate tracking of objects.

In the CSL dataset, tracking data are obtained from video and the use of optical recognition software. Manafifard, Ebadi and Abrishami Moghaddam (2017) provide a survey of various optical tracking systems in soccer. The CSL tracking data consists of roughly 1,000,000 rows per match measured on 7 variables where the data are recorded every 1/10th of a second. Each row corresponds to a particular player or the ball at a given instant in time. Although the inferences gained via our analyses are specific to the CSL, we suggest that the methods are applicable to any high-level soccer league which collects tracking data.

3 Analyses based on linear models

We introduce variables that may be relevant to parking the bus. The variables are only defined for time segments where the ball is in play. They are defined below for a given match:

$$\begin{aligned} t &\equiv \text{time of the match in minutes, } t \in (0, 90) \\ A(t) &\equiv \text{area of the convex hull of the leading team at time } t \\ X_1(t) &\equiv \text{goal differential in favour of the leading team at time } t \\ X_2(t) &\equiv \text{pre} - \text{match decimal betting odds corresponding to the leading team at time } t \\ Y(t) &\equiv \text{binary variable indicating a shot taken by the trailing team at time } t \end{aligned} \tag{1}$$

In our analyses, the area $A(t)$ of the convex hull evaluated at time t plays a central role. The convex hull of a set of points on a plane is the smallest convex polygon that contains all the points in the set. At a particular time in a match, we can treat the 11 players on a team as 11 points which form the convex hull. We assume that the area of the convex hull reflects the team's defensive style. Specifically, a smaller convex hull area coincides with a more cautious playing style. In our application, the convex hull and its corresponding area are calculated using the function *chull*() from the package *grDevices* and the function *Polygon*() from the package *sp* (Pebesma and Bivand 2005)

using the statistical programming language *R* (R Core Team 2021). In a single 90 min match, based on tracking data measured at 10 Hz, convex hulls and their associated areas are calculated $10(60)(90) = 54{,}000$ times. On a laptop computer with a single processor Intel Core i5 @ 2.7 GHz, and 8 GB of RAM, equipped with macOS Sierra version 10.12.6, this took 724 min of computation.

Since the keeper is not involved in attacking play, it is natural to ask whether the keeper should be included in the calculation of the convex hull. We believe that it is sensible to include the keeper in the calculation. For example, if a team is attacking and all of the players are pressing forward, the team is playing very aggressively. In this case, we want the area of the convex hull to be large in keeping with our premise that small convex hulls indicate defensive cautiousness. Fortuitously, in this instance, if we include the keeper, the area of the convex hull is large. However, if we remove the keeper from the calculation, the area of the convex hull would be much smaller. Therefore, teams that are attacking tend to have larger convex hulls.

We also comment on two dead-ball plays in soccer, corner kicks and penalty kicks. With corner kicks, the defending team has a choice regarding how they wish to defend. If they are cautious (i.e. parking the bus), their forwards will be positioned near the box. If they are expansive, the forwards will be pushed further out. Therefore, the area of the convex hull for the defending team will measure the extent to which they are parking the bus. Penalty kicks (PKs) are rare events in soccer (about one PK every six games in the English Premier League). For these, we calculate areas of convex hulls in the period of play prior to the PK.

To get a sense of how $A(t)$ varies with respect to the time t of a match, Figure 2 provides the scatterplot corresponding to the second half of play for Guangzhou Evergrande Taobao in their March 8, 2019 home match against Tianjin Teda. During the entire second half period, Guangzhou Evergrande Taobao had a one-goal lead. Since player movement is continuous, $A(t)$ is a continuous function. However, we plot $A(t)$ at a rate of 10 times per second which causes the plot to appear less smooth. The erratic up and down nature of the plot indicates how teams transition between offense and defense. In the plot, we observe a slightly decreasing trend in $A(t)$ suggesting that Guangzhou Evergrande Taobao attempted to protect their lead and played more defensively towards the end of the match.

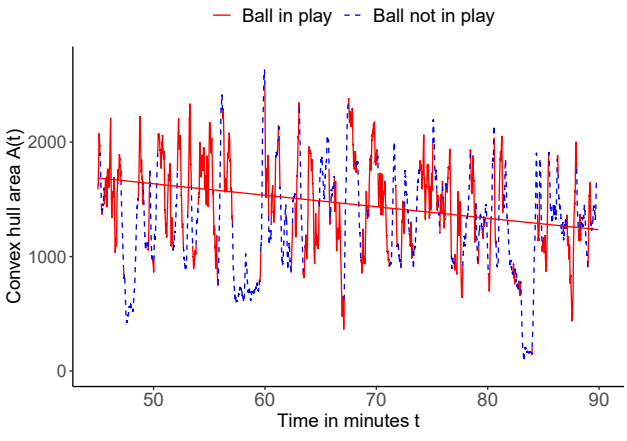

Figure 2: Plot of the area $A(t)$ of the convex hull for Guangzhou Evergrande Taobao during the second half of their March 8, 2019 home match against Tianjin Teda. $A(t)$ is calculated 10 times per second and a straight-line regression has been superimposed.

We wish to see how the areas of the convex hulls of teams interact throughout the match. In Figure 3(a), we plot the corresponding areas for the previously studied match between Guangzhou Evergrande Taobao and Tianjin Teda with respect to the time of the match. We see that there is a tendency for large values of $A(t)$ for one team to be associated with small values of $A(t)$ for the opponent. This is also evident from the associated scatterplot provided in Figure 3(b). The sample correlation coefficient corresponding to Figure 3(a) is $r = -0.349$ during the times that the ball is in play. To some extent, this is expected since a team will retreat and be less expansive if the opponent is applying continued pressure.

3.1 Three-way ANOVA

We begin by investigating how match circumstances (i.e. the covariates t, $X_1(t)$ and $X_2(t)$) relate to parking the bus as expressed by $A(t)$. We adopt the convention that when a match is tied, both teams are defined as the leading team. Therefore, when matches are tied at time t, there are two observations recorded for each of $A(t)$, $X_1(t)$, $X_2(t)$ and $Y(t)$.

To carry out a three-way ANOVA, we bin the data to define the levels for each of the three factors t, $X_1(t)$ and $X_2(t)$. We segment the time t into six intervals of interest corresponding to late-game situations where we believe that parking the bus is a prevalent tactic: (60, 65), (65, 70), (70, 75), (75, 80), (80, 85) and (85, 90). We do not include added time beyond 90 min since the amount of added time differs across matches.

For the second factor, we restrict $X_1(t)$ to three states with goal differentials 0, 1 and 2. These differentials correspond to matches that are competitive. For a given match, we consider each of the six time intervals, and if the goal differential is constant throughout the interval (either 0, 1 or 2), then an observation is recorded. We emphasize that $X_1(t)$ is defined with respect to the leading team, and we note that the leading team can change during the match. Consequently, $X_1(t)$ is nonnegative.

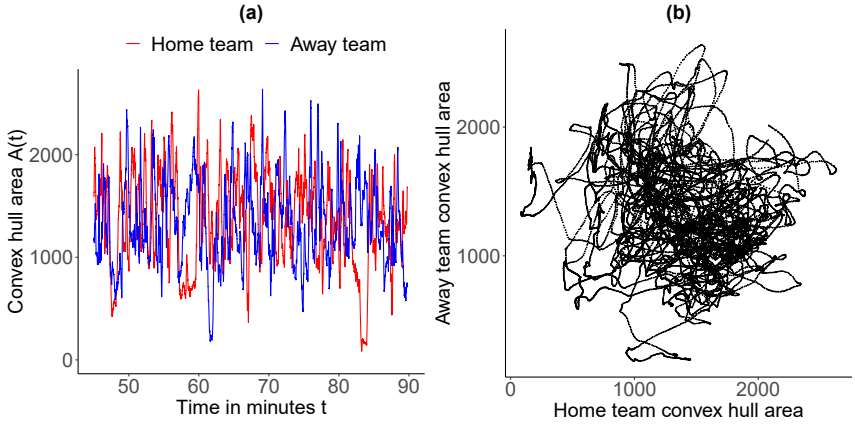

Figure 3: Plot (a) is the area $A(t)$ of the convex hull for Guangzhou Evergrande Taobao (home team) and for Tianjin Teda (road team) during the second half of their March 8, 2019 match. $A(t)$ is calculated 10 times per second. Plot (b) is the associated scatterplot of the areas for the home team and the road team.

For the third factor $X_2(t)$, we access pre-match betting odds available from the website https://www.oddsportal.com/soccer/china/super-league-2019/results/. The betting odds (reported in decimal format) provide us with the relative strength of the two teams. Ignoring the vigorish imposed by the bookmaker, the interpretation of betting odds o for a team is that the team has a pre-match probability $1/o$ of winning the match. Therefore, values of o slightly greater than 1.0 indicate a strong favourite whereas large values of o indicate an *underdog*. For a given match and a given time interval, we define four bins for the decimal odds of the leading team: [1.3,1.7), [1.7,2.3), [2.3,3.0) and [3.0,8.0). The odds are restricted so that only competitive matches are included, and the endpoints are selected to provide comparable numbers of observations in each bin.

The variable $X_2(t)$ is the pre-match betting odds of the leading team and was obtained using the standard three-way betting odds for soccer corresponding to home wins, draws and losses. Ideally, relative strength would be better measured with *moneyline* odds corresponding to wins and draws where wagers corresponding to draws are refunded. The reason why three-way betting odds are not ideal is that two matches can have identical win odds yet different draw and loss odds. However, the differences in odds in these two situations are typically minor. We have used the notation $X_2(t)$ for consistency with the other variables. However, note that the pre-match betting odds are intended to denote relative team strength, and they do not change throughout the match unless there is a change in the leading team.

For our response variable in the three-way ANOVA, we calculate the average value of $A(t)$ throughout the time interval. The average is intended to convey the general playing style (defensive vs. aggressive) over the time period.

To illustrate the terms in the ANOVA model, consider a match where the score is 2-0 just prior to the 70th minute of the match. Following conventional notation where the first team in the scoreline is the home team, the home team is the leading team who leads by two goals. Assume further that the home team is the favoured team with pre-match decimal betting odds 1.5. In this match, consider the time interval (70, 75) minutes during which neither team scores and that the average area of the convex hull formed by the home team during minutes (70, 75) is 1300 squared metres. Then, since all covariates and responses are taken with respect to the leading team, for this time interval, we have the response $A = 1300$, and covariates $t \in (70, 75)$ corresponding to the third time category, $X_1 = 2$ (goal differential), and odds $X_2 \in (1.3, 1.7)$ corresponding to the first category.

Based on the above considerations, we have $n = 1221$ observations across all matches in the $6 \times 3 \times 4$ ANOVA. The cell counts are provided in Table 1 where it is observed that we have an unbalanced design. It is apparent that there are few cases of weaker teams (i.e. $X_2 \in [3, 8)$) that lead by large goal differentials (i.e. $X_1 = 2$).

One of the assumptions of ANOVA concerns the normality of observations. A quantile plot of the residuals resulting from the three-way ANOVA does not suggest any obvious departures from normality. This is confirmed by a formal goodness-of-fit test (Anderson–Darling) where the statistic $A = 0.3368$ leads to the p-value 0.505.

In Table 2, we present the results of the three-way ANOVA where we have allowed for the possibility of first-order interaction terms. The main takeaway is that the cautiousness of the leading team is strongly associated with the time t of the match, the goal differential X_1 and the

Table 1: Cell counts for the $6 \times 3 \times 4$ ANOVA where the factors correspond to the time t, the goal differential X_1 and the relative strength X_2 corresponding to the leading team.

	$X_1 = 0$				$X_1 = 1$				$X_1 = 2$			
X_2	[1.3,1.7)	[1.7,2.3)	[2.3,3)	[3,8)	[1.3,1.7)	[1.7,2.3)	[2.3,3)	[3,8)	[1.3,1.7)	[1.7,2.3)	[2.3,3)	[3,8)
$t \in (60, 65)$	11	27	25	37	19	21	14	19	13	4	5	6
$t \in (65, 70)$	11	23	24	33	22	25	13	16	13	8	6	4
$t \in (70, 75)$	14	22	18	35	24	27	14	18	15	9	11	5
$t \in (75, 80)$	16	19	17	34	21	26	13	21	15	9	9	3
$t \in (80, 85)$	14	19	19	32	21	28	13	22	14	11	9	3
$t \in (85, 90)$	15	18	16	31	22	26	13	23	15	11	9	3

Table 2: Results from the $6 \times 3 \times 4$ ANOVA which relates cautious playing style (i.e. parking the bus via $A(t)$) to the covariates. The first-order covariates are the time t, the goal differential X_1 and the relative strength X_2 corresponding to the leading team.

Variable	Df	Sum Sq	F-statistic	p-value
t	5	1,855,123	8.066	1.71e−07***
X_1	2	6,646,271	72.242	2.00e−16***
X_2	3	3,556,860	25.774	3.76e−16***
t^*X_1	10	493,666	1.073	0.378
t^*X_2	15	193,291	0.280	0.997
$X_1{}^*X_2$	6	204,003	0.739	0.618
Error	1179	54,234,240		

relative team strength X_2. There is no evidence of first-order interactions involving t, X_1 and X_2.

We are able to drill down a little deeper on the inferences obtained from Table 2 by examining the associated interaction plots. In Figure 4(a), we examine the interaction between t and the goal differential X_1. The downward trends suggest that leading teams become more cautious as the game progresses during the second half. It is interesting that cautiousness is greatest for one-goal leads as this is the most tenuous lead. When the match is tied, it appears that teams are still trying to win until the last 5 min, at which time they appear satisfied with the draw. In Figure 4(b), we examine the interaction between t and the relative strength of the leading team X_2. We observe that weaker teams with leads are the most cautious. This is understandable as the weaker team may have less confidence that they can maintain the lead, and hence they assume an extremely defensive style. It is also possible to study the interaction between X_1 and X_2. Here, the conclusions are similar to those obtained from the previous interaction plots. We note that there is some overlap associated with the error bars in Figure 4.

It is reasonable to ask whether the choice of binning in Table 1 impacts the results. We modified our analysis with coarser and translated bins leading to a $4 \times 2 \times 2$ ANOVA based on $n = 765$ observations. The new bins for time were $t \in (55, 63.5)$, $t \in (63.5, 72)$, $t \in (72, 80.5)$ and $t \in (80.5, 90)$. The new bins for goal differential were $X_1 = 0$ and $X_1 = 1, 2$, and the new bins for team strength were $X_2 \in [1.2, 2.2)$ and $X_2 \in [2.2, 7)$. Corresponding to Table 2, the p-values for t, X_1

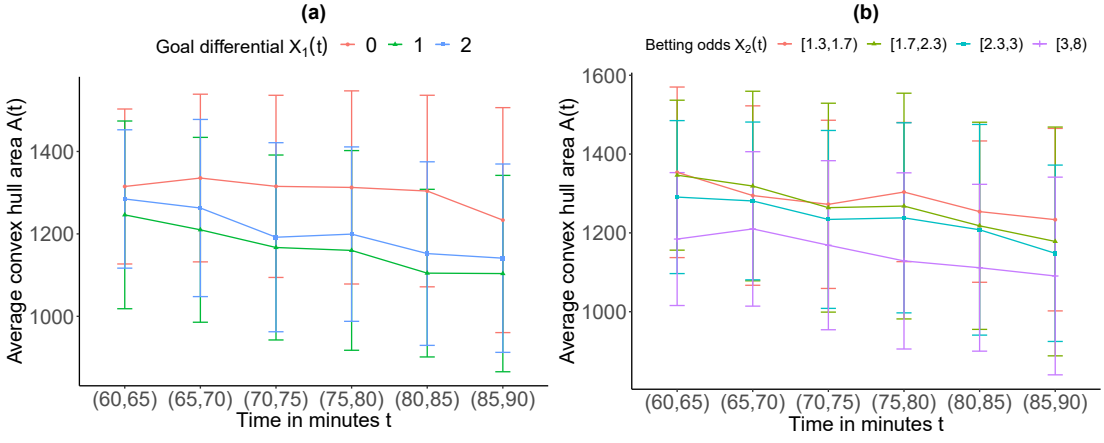

Figure 4: Plot (a) is the interaction plot between the score differential X_1 in favour of the leading team versus the time t of the match as it relates to the cautiousness of the leading team as expressed via average $A(t)$. Plot (b) is the interaction plot between the relative strength X_2 of the leading team versus the time t of the match as it relates to average $A(t)$. Error bars are included.

and X_2 remained highly significant with p-values 2.1e−05, 9.1e−27 and 9.0e−20, respectively. Under the new bin structure, the qualitative interpretations remained the same.

We now consider a variation of the previous ANOVA analysis where we include a team effect. The team effect is interesting as some teams may be more proficient at minimizing shots when leading. Of course, we keep in mind that teams often change tactics throughout a match. The new ANOVA results are presented in Table 3. Here, we see that the team effect is strongly significant. However, comparing with Table 2, we see that the p-values for the original variables do not change in meaningful ways. This suggests that there are average tendencies with respect to parking the bus across teams. Digging in further with respect to the team effect, there was no clear pattern relating the quality of the team and its propensity to play compact. For example, based on significant findings from Tukey's multiple comparisons procedure, Chongqing SWM and Wuhan Zall played the most compact styles. These teams finished 10th and 6th, respectively in the 2019 season standings (roughly middle of the table). Conversely, the three teams that played the most expansive styles were Tianjin Teda, Tianjin Quanjian and Guanzhou R&F, finishing 7th, 14th and 12th in the standings, respectively. Again, there seems to be no clear relationship between style (in terms of compactness) and overall quality.

3.2 Logistic regression

Our second investigation is primarily concerned with how the area $A(t)$ corresponding to the convex hull of the leading team relates to soccer outcomes. Naturally, the most important soccer consideration is goal scoring. However, goals in soccer are rare events (less than three goals on average per match in most professional leagues). Instead, we consider shots as the response variable since shots lead to goals.

Table 3: Variation of the ANOVA presented in Table 2 where a team effect has been introduced. Again, we relate cautious playing style (i.e. parking the bus via $A(t)$) to the covariates. The first-order covariates are the time t, the goal differential X_1 and the relative strength X_2 corresponding to the leading team.

Variable	Df	Sum Sq	F-statistic	p-value
t	5	1,774,652	8.327	9.56e−08***
X_1	2	6,988,712	81.981	2.20e−16***
X_2	3	1,808,817	14.146	4.65e−09***
Team effect	15	4,620,055	7.226	2.07e−15***
t^*X_1	10	571,876	1.342	0.203
t^*X_2	15	177,536	0.278	0.997
$X_1{}^*X_2$	6	68,415	0.268	0.952
Error	1164	49,614,185		

Recall that we adopt the convention that when a match is tied, both teams are defined as the leading team. Therefore, when matches are tied at time t, and a shot is taken, there are two observations. Although shots do not necessarily lead to goals, they do provide a measure of offensive dominance.

We first provide an exploratory plot involving the binary variable $Y(t)$ which is defined at times t when a shot is taken. We let

$$Y(t) = \begin{cases} 1 & \text{if shot at time } t \text{ is taken by the trailing team} \\ 0 & \text{if shot at time } t \text{ is taken by the leading team} \end{cases}$$

In Figure 5, $Y(t)$ is plotted against the average area of the convex hull $\bar{A}(t)$ of the leading team over the 60 s interval before the shot is taken. It is difficult to view a pattern involving a binary $Y(t)$, and therefore, the points have been smoothed. From the smoothed plot, the decreasing trend indicates that a shot is increasingly more likely to be generated by the trailing team (i.e. large smoothed $Y(t)$) as the leading team tends to park the bus (i.e. decreasing $A(t)$). From the smoothed plot, we obtain the interpretation that once the area of the leading team's convex hull is smaller than 1301.13 squared metres, they are more likely to concede a shot (i.e. $Y > 0.5$).

Using an inferential approach, we restrict attention towards the latter stages of matches given by $t \geq 60$. Then, for every shot that occurs, it is either a shot by the leading team or a shot by the trailing team. Therefore, we have two outcomes and this facilitates the conditions for logistic regression. Our regression covariates are t, $X_1(t)$ and $X_2(t)$ as defined in (1) where the values of $X_1(t)$ and $X_2(t)$ are

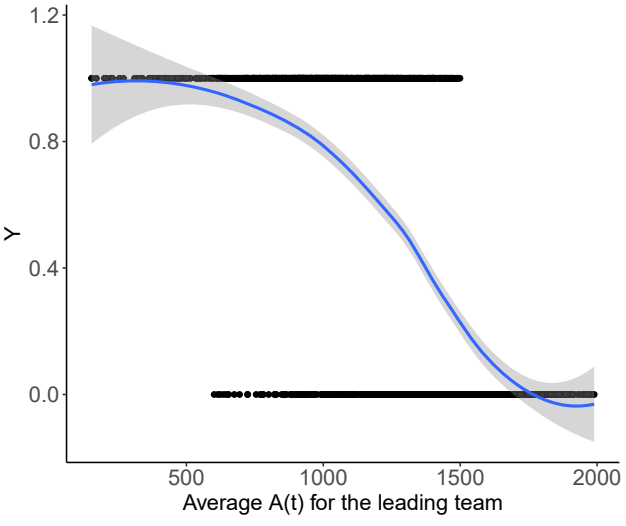

Figure 5: Observed points and the smoothed plot of the binary variable $Y(t)$ against the average area of the convex hull of the leading team during the 60 s interval prior to the shot.

determined at the time immediately prior to the shot. The final regression covariate is the average $\bar{A}(t)$ taken over the 60 s prior to the shot at time t. It is important to take this historical approach because our interest is in the cumulative effect of parking the bus. Although there are potentially other confounding variables that are related to the shot variable Y, we have selected explanatory variables that are well known to be related to Y. We note that there is some evidence that trailing teams shoot more shots and shots that are of lesser quality (Mead, O'Hare and McMenemy 2023; Younggren and Younggren 2021).

There are various ways in which the model can be formulated. For simplicity, we have chosen to code the time t, the goal differential X_1 and the relative strength X_2 corresponding to the leading team as continuous variables. We note that we obtain similar conclusions under different formulations (e.g. defining X_1 as categorical).

Table 4 provides the output for the analysis based on the corresponding logistic regression. Note that there were $n = 1792$ shots taken in this dataset. The most important result relating to our investigation is that the coefficient corresponding to $\bar{A}(t)$ is both negative and highly significant. This implies that when the leading team is playing more defensively (as suggested by having a compact convex hull), then the trailing team is more likely to take the next shot. As expected, the analysis also provides an indication that shots are related to the goal differential X_1 (i.e. teams that are trailing by greater margins tend to have more shots). Also, shots are related to the relative team strength X_2 (i.e. trailing teams that are stronger tend to have more shots). We also fit an expanded model containing first-order interactions, but none of the interactions terms were significant.

One of the assumptions of the proposed logistic regression model concerns the linear relationship between the logit function and the covariates. To investigate the linearity assumption, we fit a generalized additive model (GAM) with the shot response variable and the predictor

$$\beta_0 + f_1(t) + \beta_1 I_{(X_1(t)=1)} + \beta_2 I_{(X_1(t)=2)} + f_2(X_2(t)) + f_3(\bar{A}(t)). \quad (2)$$

In (2), we note that X_1 is categorical, and therefore, we introduced two dummy variables using the indicator function I. A supplementary document provides plots of the fitted functions f_1, f_2 and f_3. The fitted functions f_2 and f_3 appear close to linear. The fitted function for f_1 is not strictly linear, but does not show much curvature over the range $t \in (60, 90)$. By calculating the deviance statistic, it is also possible to test the adequacy of the logistic model against the full GAM model. This leads to the p-value 0.1219. We therefore conclude that the assumptions of linearity in the logistic regression model appear reasonable.

Although the linear model analyses are straightforward and interpretable, they are not without weaknesses. For example, we have greatly reduced the richness of the dataset by binning observations in Section 3.1. Further, ANOVA models assume independence of observations, and we have observations from the same matches, a weakness which is compounded in the case of tied matches.

Table 4: Results from the logistic regression which relates shots taken by the trailing team to the covariates. The covariates of interest are the time t, the goal differential X_1 corresponding to the leading team, the relative strength X_2 corresponding to the leading team, and the average area of the convex hull $\bar{A}(t)$ for the leading team 60 s prior to the shot.

Variable	Coefficient	Std error	Z-statistic	p-value
Intercept	4.9515	0.6158	8.041	8.94e−16***
t	−0.0011	0.0067	−0.172	0.8638
X_1	0.1757	0.0803	2.189	0.0286*
X_2	0.0890	0.0469	1.898	0.0577
$\bar{A}(t)$	−0.0040	0.0002	−18.485	2.00e−16***

4 Discussion

We have explored defensive playing style in soccer using a full season of tracking data from the CSL. The primary message is that an extremely cautious playing style (i.e. a small convex hull) is negatively associated with generating shots. Although a cause-effect relationship has not been established, teams that are leading may consider more expansive soccer, and to continue to seek goals. It is possible that this message is applicable to other soccer leagues and extends to other invasion sports such as ice hockey. Other interesting observations include the following:

- leading teams are most cautious with a one-goal lead
- leading teams are more cautious near the end of matches
- leading teams are more cautious if they are perceived as the weaker team

It is useful to draw connections between this work and previous work. Recall that Silva and Swartz (2016) establish that trailing teams are more likely to score the next goal. Our research establishes a plausible explanation – it is because the leading team is playing more cautious in the sense of parking the bus. When teams park the bus, the style is associated with a higher probability that the next shot is generated against them, and shots are associated with goals. Now, not all shots are of the same quality, but we already know that trailing teams are more likely to concede a goal.

In future work, we may consider differentiation between the quality of shots. This would necessitate a departure from the logistic regression approach of Section 3.2 where shots are binary (i.e. either by the trailing or the leading team), and instead, introduce a quantitative response. Such an approach could use the methods of Singh (2018), who introduces the concept of expected threat xT which assigns value to different locations on the pitch.

In this application, the area of a convex hull may be sensitive to the positioning of a rogue player (i.e. an outlier). Therefore, instead of the areas of convex hulls, we might consider alternative measures of cautiousness. For example, space ownership in soccer is a statistic that has been evaluated in various ways (Wu and Swartz 2023). Space ownership (i.e. pitch control) is defined at any position on the pitch. It may be possible to develop a composite measure of ownership in some defensive section of the pitch.

Our analyses have been based on the implementation of simple linear models. Drawbacks of the analyses involve the necessity of some strong distributional assumptions and the lack of utilization of the full dataset. In future work, we intend to expand our investigation of parking the bus by developing methods from historical functional data analysis (FDA). For example, our ANOVA analysis considers the relationship of average values of $A(t)$ taken over time intervals against covariates of interest. In FDA, the basic idea is that one considers the regression of functions. In this application, we have responses (e.g. the areas of convex hulls $A(t)$) that are time dependent. Therefore, an FDA approach would regress the entire function $A(t)$ against covariates of interest where both $A(t)$ and the covariates are functions of $t \in (0, 90)$. These methods may provide greater insight as they take the time structure of the dataset into account. FDA is an important tool in sports analytics since many quantities of interest are indexed by the time of the match. FDA has been utilized in various sporting applications including basketball (Chen and Fan 2018) and rugby league (Guan et al. 2020). For a practical introduction to FDA, see Ramsay, Hooker and Graves (2009).

Acknowledgement: All authors have been partially supported by the Natural Sciences and Engineering Research Council of Canada. The authors thank Daniel Stenz, Technical Director of Shandong Luneng Taishan FC who provided the tracking data used in this paper. The authors also thank two Reviewers, two Co-Editors and an Associate Editor whose comments improved the manuscript.

Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: None declared.

Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

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

Beaudoin, D., O. Schulte, and T. B. Swartz. 2016. "Biased Penalty Calls in the National Hockey League." *The ASA Data Science Journal* 9 (5): 365—72.

Buldú, J. M., J. Busquets, J. H. Martínez, J. L. Herrera-Diestra, I. Echegoyen, J. Galeano, and J. Luque. 2018. "Using Network Science to Analyse Football Passing Networks: Dynamics, Space, Time and the Multilayer Nature of the Game." *Frontiers in Psychology* 9: 1900.

Chen, T., and Q. Fan. 2018. "A Functional Data Approach to Model Score Difference Process in Professional Basketball Games." *Journal of Applied Statistics* 45: 112—27.

Diquigiovanni, J., and B. Scarpa. 2019. "Analysis of Association Football Playing Styles: An Innovative Method to Cluster Networks." *Statistical Modelling* 19 (1): 28—54.

Fernandez-Navarro, J., L. Fradua, A. Zubillaga, P. R. Ford, and A. P. McRobert. 2016. "Attacking and Defensive Styles of Play in Soccer: Analysis of Spanish and English Elite Teams." *Journal of Sports Sciences* 34 (24): 2195—204.

Fernandez-Navarro, J., L. Fradua, A. Zubillaga, and A. P. McRobert. 2018. "Influence of Contextual Variables on Styles of Play in Soccer." *International Journal of Performance Analysis in Sport* 18 (3): 423—36.

Garrido, D., D. R. Antequera, J. Busquets, R. L. Del Campo, R. R. Serra, S. J. Vielcazat, and J. M. Buldú. 2020. "Consistency and Identifiability of Football Teams: A Network Science Perspective." *Scientific Reports* 10: 19735.

Gollan, S., C. Bellenger, and K. Norton. 2020. "Contextual Factors Impact Styles of Play in the English Premier League." *Journal of Sports Science and Medicine* 19 (1): 78—83.

Gonçalves, B., D. Coutinho, S. Santos, C. Lago-Penas, S. Jiménez, and J. Sampaio. 2017. "Exploring Team Passing Networks and Player Movement Dynamics in Youth Association Football." *PLoS One* 12 (1): e0171156.

Guan, T., R. Nguyen, J. Cao, and T. B. Swartz. 2020. "In-Game Win Probabilities for the National Rugby League." *The Annals of Applied Statistics* 16 (1): 349—67.

Gudmundsson, J., and M. Horton. 2017. "Spatio-Temporal Analysis of Team Sports." *ACM Computing Surveys* 50 (2): 22.

Lago-Peñas, C., M. Á. Gómez-Ruano, and Y. Gai. 2017. "Styles of Play in Professional Soccer: An Approach of the Chinese Soccer Super League." *International Journal of Performance Analysis in Sport* 17 (6): 1073—84.

Manafifard, M., H. Ebadi, and H. Abrishami Moghaddam. 2017. "A Survey on Player Tracking in Soccer Videos." *Computer Vision and Image Understanding* 159: 19—46.

Metulini, R., M. Manisera, and P. Zuccolotto. 2017. "Sensor Analytics in Basketball." In *Proceedings of the 6th International Conference on Mathematics in Sport*.

Mead, J., A. O'Hare, and P. McMenemy. 2023. "Expected Goals in Football: Improving Model Performance and Demonstrating Value." *PLoS One* 18 (4): 1—29.

Miller, A., L. Bornn, R. P. Adams, and K. Goldsberry. 2014. "Factorized Point Process Intensities: A Spatial Analysis of Professional Basketball." In *Proceedings of the 31st International Conference on Machine Learning - Volume 32*, 235—43. Beijing: JMLR.org.

Ötting, M., R. Langrock, and A. Maruotti. 2023. "A Copula-Based Multivariate Hidden Markov Model for Modelling Momentum in Football." *AStA Advances in Statistical Analysis* 107: 9—27.

Pebesma, E. J., and R. S. Bivand. 2005. "Classes and Methods for Spatial Data in R." *R News* 5 (2): 9—13.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Ramsay, J. O., G. Hooker, and S. Graves. 2009. *Functional Data Analysis with R and Matlab*. New York: Springer.

Silva, R., and T. B. Swartz. 2016. "Analysis of Substitution Times in Soccer." *Journal of Quantitative Analysis in Sports* 12 (3): 113—22.

Singh, K. 2018. Introducing Expected Threat (xT). https://karun.in/blog/expected-threat.html (accessed January 25, 2022).

Thomas, A. C. 2017. "Poisson/exponential Models for Scoring in Ice Hockey." In *Handbook of Statistical Methods and Analyses in Sports*, edited by J. A. Albert, M. E. Glickman, T. B. Swartz, and R. H. Koning, 271—85. Boca Raton: Chapman & Hall/CRC.

Wu, L., and T. B. Swartz. 2023. A New Metric for Pitch Control Based on an Intuitive Motion Model. https://www.sfu.ca/tswartz/ (accessed May 5, 2023).

Wu, Y., X. Xie, J. Wang, D. Deng, H. Liang, H. Zhang, S. Cheng, and W. Chen. 2019. "ForVizor: Visualizing Spatio-Temporal Team Formations in Soccer." *IEEE Transactions on Visualization and Computer Graphics* 25 (1): 65—75.

Younggren, J., and L. Younggren. 2021. A New Expected Goals Model for Predicting Goals in the NHL. https://evolving-hockey.com/blog/a-new-expected-goals-model-for-predicting-goals-in-the-nhl/ (accessed May 5, 2023).