A Causal Investigation of Pace of Play in Soccer

Nirodha Epasinghege Dona and Tim B. Swartz [*]

Abstract

This paper provides a comprehensive investigation of playing with pace in soccer. The investigation begins by introducing two quantitative definitions of pace whose calculations are facilitated through the availability of player tracking data. In the study, the primary scientific question concerns whether playing with pace is an advantageous strategy in terms of playing style. This is a question that has not been adequately resolved in either soccer or hockey. Here, we use methods of causal inference to investigate the relationship between pace in soccer and shots. It is determined that playing with more pace than the opponent throughout a match confers an advantage of approximately two additional shots per game. As a byproduct of our analysis, other soccer insights related to pace are obtained.

Keywords : big data, causal inference, player tracking data, spatio-temporal analyses.

[*] N. Epasinghege Dona is a PhD candidate and T. Swartz is Professor, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6. Swartz has been partially supported by the Natural Sciences and Engineering Research Council of Canada. The work has been 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 and Associate Editor and three anonymous Reviewers whose comments have improved the paper.

1 INTRODUCTION

In team sports, playing style is a much discussed topic and an important component of success. For example, in soccer, we hear about the gegenpress (Tweedale 2022), total football (McLellan 2010) and parking the bus (Guan, Cao and Swartz 2022). However, playing style is notoriously difficult to quantify in soccer. It is difficult to quantify since playing style is a team concept, which relies on the actions of multiple players whose movements are fluid in both time and space.

However, the landscape for studying playing style has changed in recent years with the advent of player tracking data. With player tracking data, the location coordinates for every player on the field are recorded frequently (e.g. 10 times per second in soccer). With such detailed data, the opportunity to explore novel questions in sport has never been greater. The massive datasets associated with player tracking also introduce data management issues and the need to develop modern data science methods beyond traditional statistical analyses. Gudmundsson and Horton (2017) provide a review of spatio-temporal analyses that have been used in invasion sports where player tracking data are available.

This paper is concerned with "pace of play" in soccer, a relatively underexplored topic. In some sports, pace is readily defined. For example, in basketball, team pace may be defined as the average number of possessions per game. In the NBA, this is a well-studied statistic which is available from various websites including https://www.nba.com/stats/teams/advanced/

In American football, although there is a clear notion of pace of play, there is no commonly reported statistic that directly measures pace. In the National Football League (NFL), the number of plays per game is available for each team from standard box scores. Although this statistic is related to pace, it is obvious that poor offensive teams who rarely make first downs have fewer plays per game. Therefore, in football, the average number of plays per game for a team is confounded with offensive strength, and consequently, the number of plays is not a pure measure of pace. Pace in football can be increased for a team by using a "hurry-up offense" which affords more plays in a given period of time provided that the team continues to make first downs. Furthermore, teams that frequently pass the ball (as opposed to run the ball) typically use up less of the clock and have more plays from scrimmage.

In ice hockey, the definition of pace is even less clear. See, for example, Silva, Davis and Swartz (2018) where various definitions of pace are considered. Yu et al. (2018) revisit the hockey problem and suggest an alternative definition of pace.

The sport of soccer shares some of the same challenges as hockey with respect to the definition of pace. For example, how is possession determined? How do successful passes contribute to pace and should pace calculations involving a pass be counted differently than when dribbling? Shen, Santo and Akande (2022) builds on the aforementioned hockey papers and uses event data to investigate pace in soccer.

This paper differs from Shen, Santo and Akande (2022) in a number of key directions. First, this paper uses tracking data rather than event data to study pace. Second, alternative definitions of pace are provided. In particular, we define *attacking pace* which is related to "direct play", a much discussed tactic in soccer. Third, we provide various sporting implications associated with pace. Finally, our primary goal addresses the key question of whether playing with pace is strategicly sound. Many soccer experts believe that moving the ball quickly is advantageous. When you move the ball quickly, the logic is that it affords the defensive team less time to transition to solid defensive formations. However, to our knowledge, this basic tenet of soccer has never been tested. Is it better to play with pace? We address this question by using methods of causal inference (Pearl 2009). Obviously, decisions that are made on the field are often instantaneous. Therefore, it is impossible to use traditional randomized trials to determine the cause-and-effect relationship between playing with pace and success. With match data, we have "studies" as opposed to "experiments". Fortunately, the methods of causal inference allow us to address causality in studies provided that confounding variables can be identified and measured. With tracking data and our subject knowledge of soccer, confounding variables are accessible.

Related to our investigation of pace, they have been many investigations of determinants of success in soccer. A sample of recent papers include Lepschy, Wäsche and Woll (2021), Merlin et al. (2020), and in the women's game, de Jong et al. (2020).

In soccer, the most investigated aspect of playing style concerns formations. For example, the book "Inverting the Pyramid" (Wilson 2013) considers the history of soccer tactics throughout the world with an emphasis on positional play and player roles. It is also now common during television broadcasts to provide graphical statistics that depict the average location of each player during a match. Such information is useful in determining match strategy as it can point out features such as gaps in player alignment. There have also been many technical papers written on player formation. For example, Shaw and Glickman (2019) use tracking data and clustering methods to determine a team's offensive and defensive formations. This is useful as the fluidity of the sport and changing tactics sometimes makes it difficult to distinguish between formations (e.g. 4-4-2 versus 3-5-2). Goes et al. (2021) also identify formations using tracking data and relate attacking success to formations.

The association between style and results in soccer has been well investigated. For example, in their Table 1, Kempe et al. (2014) list various ball possession and passing metrics which have been explored in the literature. Kempe et al. (2014) also propose aggregate metrics and relate these to success. However, a distinguishing feature of our work is that we consider a causal approach rather than one of association. This is made possible by the availability of player tracking data.

In Section 2, we introduce and motivate two definitions of pace. We contrast these definitions with alternative definitions that have been presented in the literature. In Section 3, we describe the player tracking dataset and discuss the challenges involved in pace calculations. One of the challenges is the determination of possession. In Section 4, we provide exploratory data analyses which provides various sporting insights on pace. The sporting insights are highlighted with the letters A-E. This section is also useful in identifying confounding variables that are related to pace. In Section 5, we present a causal analysis concerning the benefit of playing with pace. This involves the fitting of a MANOVA model which is the foundation for the determination of propensity scores and matching. The main result of this section is that playing with pace is a beneficial team strategy in soccer in terms of generating more shots. We conclude with a short discussion in Section 6.

2 DEFINITIONS OF PACE IN SOCCER

Dan Blank's paperback on soccer (Blank 2002) provides 54 chapters on different tactics and advice on playing the game well. The first chapter which is titled the "Holy Grail" provides an inspiration for our investigation. In this chapter, Blank claims that playing fast is better than playing slow. In other words, Blank argues that teams should play with pace.

Although the heuristic may be appealing, it does not seem that the belief has ever been corroborated against data. If the belief is true, then a measurable and sensible definition of pace may lead to important soccer insights.

First, we review some of the previous definitions of pace. In the original investigation of pace in hockey, Silva and Swartz (2018) were limited to the analysis of event data. With event data, a finite number of event types are recorded along with a timestamp. A shortcoming of the analysis is that the skating paths between events (which are relevant to pace) are unknown. Consequently, Silva and Swartz (2018) only measured horizontal distances (i.e. down the length of the rink) during which possession was maintained. Furthermore, Silva and Swartz (2018) only evaluated pace for a game and did not differentiate pace of play between the two teams. Yu et al. (2019) used more extensive event data with events recorded approximately every second on average. With this data, they were able to define pace in various directions and considered zonal, league, team-level and player-level analyses. The pace metric defined by Yu et al. (2019) appears to be an average of velocities over event intervals and therefore differs conceptually from the Silva and Swartz (2018) definition which is based on total distance travelled. In soccer, Shen, Santo and Akande (2022) also used velocity as a pace measurement but restricted analyses to sequences where possession is retained over three or more events.

A commonality amongst all of the above pace analyses is that they were based on event data. With event data, distance calculations between events assume that the ball/puck travels in a straight line. Shen, Santo and Akande (2022) described the assumption as a major limitation. In this paper, the more detailed tracking data allows us to consider the actual paths where the ball travelled.

We begin with an analogy related to our definition of pace. We suggest that a painter is painting quickly (i.e. with pace) if they are able to apply a lot of paint on a canvas in a short period of time. In soccer, we view the brush strokes as the paths where players carry the ball and the paths where a ball is successfully passed. If a team is able to move the ball quickly, then they are playing with pace. The concept of possession is important; if a team is simply punting the ball downfield, in our view, they are not playing with pace. To operationalize these ideas, we consider the non-contiguous time intervals $(t_1, t_2), \ldots, (t_n, t_{n+1})$ in a match where a team has possession. During the possession interval i, the team moves distance d_i, $i = 1, \ldots, n$. Then, following the painting analogy, we refer to the team's *general pace* in the match as

$$GP = \frac{\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n}(t_{i+1} - t_i)} \ . \tag{1}$$

Similarly, there is a corresponding pace formula (1) for the opponent. Note that the two teams will differ in the amount of time possession during the match. Therefore, the pace measure (1) is reflective of their style of play while in possession, and is insensitive to their total time of possession. Although the general pace metric (1) is defined in terms of a match, it can also be calculated for shorter periods of time (e.g. a half) or even for a single possession.

We contrast the general pace metric (1) with the quantity

$$P_{\text{SSA}} = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i}{t_{i+1} - t_i} \tag{2}$$

which is related to the velocity concept of pace utilized by Shen, Santo and Akande (2022); note, however that Shen, Santo and Akande (2022) used medians rather than means. When comparing (1) with (2), we observe that (2) is sensitive to and is inflated by very fast passes (i.e. typically large d_i that occur over moderate time intervals $t_{i+1} - t_i$). We believe that (1) better reflects pace as the totality of distance covered with respect to the cumulative time of possession.

We now introduce a variation of the general pace metric GP defined in (1). We note that there are differences in scoring intent based on the type of passes and dribbling. For example, the "tiki-taka" approach adopted by the Spanish National team in 2006 relied on many consecutive short passes that emphasized possession. Based on the metric GP, the tiki-taka approach would be characterized as a pacey style since passes typically have larger pace contributions than dribbling. But is tiki-taka pacey?

The aforementioned tiki-taka style allows one to reflect on stylistic differences between hockey and soccer. In hockey, the playing surface is smaller and players skate at great speeds. Therefore, it is more difficult to retain possession in hockey. Consequently, possession sequences tend to be of shorter duration than in soccer. To investigate pace in soccer with an emphasis on direct play, we modify GP and introduce *attacking pace* AP where the distances d_i now correspond to displacements down the field in the direction of the opposing goal. For example, passes back to the keeper (which have no attacking intent) do not positively contribute to attacking pace AP. Large positive contributions to the statistic AP will involve transitions such as the counter-attack.

To define AP, we refer to Figure 1 where an AP contribution is illustrated. In the plot, the "most attacking" pass that could possibly be made from point A would be to the middle of the opponent's goal line C. This potential pass has associated distance d_{AC}. Instead, the pass was made from A to B, and the attacking distance from this new point B to C is denoted d_{BC}. Therefore, the contribution (in terms of attacking) from A to B is given by the residual distance

$$d = \begin{cases} d_{AC} - d_{BC} & d_{AC} \geq d_{BC} \\ 0 & d_{AC} < d_{BC} \end{cases} . \tag{3}$$

In (3), $d_{AC} - d_{BC}$ represents the reduction of the greatest attacking distance that was made due to the pass from A to B. Therefore, the new statistic AP has the same form as GP in equation (1) where the d in equation (3) assumes a subscript i corresponding to the ith possession. We also note that the same type of calculation is carried out whether a possession involves passes or dribbles. It is important to note that the tracking data allows us to deal with path curvature when dribbling by breaking up dribbling sequences into small time intervals. If event data had been used (in contrast to tracking data), only the starting point and ending point of a dribbling sequence would be known.

A feature of the construction of the metric AP is illustrated through a possession sequence where the ball travels in a forward direction from A to B to C and where we denote the center of the goal by G. Using obvious notation for distances, the total attacking distance (3) is given by $d = (d_{AG} - d_{BG}) + (d_{BG} - d_{CG}) = d_{AG} - d_{CG}$ which demonstrates that the metric is additive over the possession path.

Whereas the metrics GP and AP describe style of play while a team is in possession, insights may also be provided by considering the extent to which teams play a given style. For example, if a team is rarely in possession, then they are rarely executing their style. Therefore, we could also introduce the metric GP^* which is similar to GP except that we omit the denominator in (1). Therefore, GP^* may be thought of as total distance travelled by the team. Therefore, while GP is a statistic that describes pace during possession, GP^* takes possession into account such that teams with little possession are not playing with pace. Similarly, we could introduce the metric AP^* which is the total attacking distance during the match. However, for the remainder of our investigation, we only focus on the general pace statistic GP and the attacking pace statistic AP. Note that all of the proposed definitions of pace are properties of the possessing team.

3 DATA

For this investigation, we have a big data problem where both event data and player tracking data are available for 237 regular season matches (three matches missing) from the 2019 season of the Chinese Super League (CSL). The schedule is balanced where each of the 16 teams plays every opponent twice, once at home and once on the road.

Event data and tracking data were 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 recorded by technicians who view film. 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. 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. Each row corresponds to a particular player 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 soccer league which collects tracking data.

3.1 Possession

A possession is defined as a period where a team has control of the ball. The event data is used to identify the possession sequences of a team. The event data contains all the events that occurred during a match and therefore tells us when possession sequences began and ended. Events where neither team is determined to have possession include injuries, cards, out-of-bounds, preliminary time to the beginning of set pieces (eg corners, throw-ins, free kicks, penalties, etc). Also, when determining possession sequences, we exclude time beyond 90 minutes since different matches have different amounts of added time. Among all the matches, there is an average of 373 possessions per match.

In Figure 2, we provide histograms (GP and AP) of the length of the possession sequences in metres. The histogram is right skewed. We observe a mean length of 46.6 (21.0) metres, minimum length 0.1 (0.0) metres, and maximum length 456.8 (82.2) metres corresponding to GP and AP, respectively.

3.2 The Pace Datasets

We pre-processed the CSL tracking and event data. Originally, the data were provided in xml files and we extracted the content using the `read_xml` function from the `XML` package using R software. The resulting tracking and event data were written into csv file format.

Ultimately, we constructed a pace dataframe for each match. This is a comprehensive dataset that allows us to investigate various questions of interest. The pace dataset is a matrix where the rows correspond to pace contributions made by an individual player during a possession. The columns consist of the following variables: start time of pace contribution, end time of pace contribution, the displacement d_i (both Euclidean distance and attacking distance (3)), match score at the beginning of the pace contribution, match score at the end of the pace contribution, the player who contributed to the pace contribution, the team of the player who made the pace contribution, whether the contributing player plays for the home or road team, and the number of playing minutes during the match for the player. We note that Yu et al. (2018) shared the pace contribution equally between the player who made the pass and the player who received the pass. In our construction, we assign credit only to the player who made the pass.

To create the pace dataframe, we looped frame by frame through the tracking data, where we matched events and time using the event data. This permitted the calculation of the relevant distances during each possession. The process required approximately 15 minutes of computation for all 237 matches. Another challenge related to the calculation of AP involved slight differences in pitch size where the coordinates of point C in Figure 1 varied across pitches.

4 EXPLORATORY DATA ANALYSES

A main objective of exploratory data analysis (EDA) is to reveal insights that can be more thoroughly investigated via modelling and inferential techniques. In this section, we use EDA to gain insights related to the pace statistics GP and AP together with other variables of interest. Below, EDA reveals five insights, labelled A-E.

In Figure 3, we produce scatterplots of GP and AP related to the home and road teams for all of the 237 available matches during the 2019 season of the CSL. We obtain a mean value of 0.66 (0.66) metres/sec, minimum value 0.42 (0.48) metres/sec and maximum value 0.78 (0.82) for the home and road team, respectively, using GP. We obtain a mean value of 0.25 (0.26) metres/sec, minimum value 0.14 (0.15) metres/sec and maximum value 0.38 (0.53) for the home and road team, respectively, using AP. Therefore, we observe that the pace statistics differentiating home and road teams are minor.

Initially, we were unsure whether pace was a property of the match (e.g. both teams play at high pace due to the particular style of the game) or whether each team has control of their respective pace. We observe that the sample correlation coefficients for GP and AP are 0.16 and 0.06, respectively. It is possible to carry out a test of correlation $H_0 : \rho = 0$ in the two cases. The p-values are given by 0.014 and 0.358, for GP and AP, respectively. Although the first correlation is statistically significant, it is not strong in magnitude. The lack of strong correlations lead to the following insight.

Insight A: In a given match, each team has control of whether they play a pacey style. The pace of one team is not dictated by the pace of its opposition.

Next, we are interested in whether pace is a characteristic that can be attributed to teams. In Figure 4, we produce boxplots of the pace (GP and AP) for each of the 16 teams in the CSL where a single datapoint refers to the pace calculation in a match. We observe that there are only minor differences in the pace distributions across teams. Using a one-way ANOVA design for testing differences across teams, we obtain p-values of 0.0278 and 0.0106, for GP and AP, respectively. This leads to the following insight.

Insight B: Although some teams may play at slightly different average pace than other teams, such differences are small (particularly with GP. Pace is primarily a property of how a team plays in a particular match rather than a general property of the team.

Next, we are interested in whether pace depends on playing position. In Figure 5, we produce boxplots of the pace (GP and AP) for defenders, midfielders and forwards in the CSL where a single datapoint refers to the pace calculation in a match. We observe differences across the three positions. Due to the constraints of the field and positioning, it is logical that defenders have more open space in front of them than midfielders, and that midfielders have more open space in front of them than forwards. Therefore, it coincides with our intuition that pace should decrease according to defenders, midfielders and forwards, respectively. Using a one-way ANOVA design for testing differences across positions, we obtain highly significant test results with p-values of $4.13e^{-10}$ and $5.18e^{-6}$, for GP and AP, respectively. This leads to the following broad insight.

Insight C: Defenders play at higher pace levels than midfielders who in turn play at higher pace levels than forwards.

Next, we are interested in whether pace is related to the time of the match. In Figure 6, we produce boxplots of the total pace by both teams (GP and AP) according to the time of the match broken into 15-minute intervals from 0 to 90 minutes. Although pace changes throughout the match, we observe different patterns according to GP and AP. With general pace GP, when a match begins, we expect that teams are alert and maintain defensive discipline. As the match continues, players tire, and they discontinue running with the same pace as before. At halftime, there is a rest period where teams recover slightly, and then they continue to tire during the second half.

With attacking pace AP, the interplay between exhaustion and defensive discipline is expressed differently. As the match continues, players tire and this allows for more open space and the opportunity to seek gaps downfield. This causes a gradual increase in attacking pace with more pronounced increases in the latter stages. Using a one-way ANOVA design for testing differences across time intervals, we obtain highly significant test results with p-values of $3.66e^{-8}$ and $4.56e^{-5}$, for GP and AP, respectively. This leads to the following insight.

Insight D: Teams plays at higher attacking pace as the match progresses.

Next, we are interested in whether pace is related to goal differential. In Figure 7, we produce boxplots of GP and AP corresponding to five goal differential categories as explained in the caption. The calculation of pace is taken over five-minute intervals for all teams and matches during the season. When a goal is scored during a five-minute interval, then the pace observation for that interval is excluded since the goal differential during the interval is not constant. With respect to the home team, we observe an interesting pattern with a slight increase in the median value of AP from $GD = -2$ to $GD = -1$, from $GD = -1$ to $GD = 0$, from $GD = 0$ to $GD = 1$, followed by a drop in pace at $GD = 2$. Note that due to the home team advantage, $GD = 2$ is a more common situation than $GD = -2$. Our nuanced intuition corresponding to these observations begins with the case $GD = -2$ where the home team is losing badly. In this case, we expect that the road team is playing defensively as argued by Guan, Cao and Swartz (2002). The home team is therefore dominant in their offensive zone (i.e. near the road team's goal). On average, there is little room downfield for the home team, and consequently, they will be unable to make significant positive contributions to AP, and the AP measurement (as observed), will be low. As GD changes from -2 through 1, we would expect the road team to play less defensively, and as previously argued, AP will increase (as observed). However, a different behavioural mechanism occurs when $GD = 2$. In this case, the home team has a dominant lead. Their lead is so great, that they have little fear of losing. Hence, when $GD = 2$, the home team is not playing ultra-defensive (i.e. largely contained in their own zone, with predominantly long passes having a high AP contribution). Rather, when $GD = 2$, the home team is playing free, and this causes a reduction in AP from $GD = 1$ to $GD = 2$. Using a one-way ANOVA design for testing differences in pace across goal differentials, we obtain significant test results with p-values of 0.041 and 0.028 for GP and AP, respectively. This leads to the following insight.

Insight E: Teams play at different pace levels depending on the goal differential.

5 CAUSAL ANALYSIS

In this section, we return to our primary question whether it is advantageous to play with pace. With a sensible definition of pace and the availability of tracking data, the issue can be addressed.

Recall that questions of cause and effect are traditionally addressed using randomization in experimental contexts. For our problem, this would require the random assignment of pace to the two teams. Of course, matches are not experiments, but rather observational studies where randomization does not occur. Therefore, we address cause and effect through methods of causal inference (Pearl 2009). Although causal inference has received great attention, the methods are often difficult to implement due to the necessity of specifying and measuring relevant confounding variables. Fortunately, sport is much simpler in its objectives than many other scientific domains, and via the spatio-temporal tracking data and the EDA investigations of Section 2, confounding variables are accessible. Therefore, together with some novel ideas, and referring to the approach introduced in Wu et al. (2021), we are able to address cause and effect associated with pace.

5.1 Propensity Scores

Using causal terminology, we think of pace as the treatment which we denote X_h and X_r, corresponding to the home and road teams, respectively. We denote W as the vector of confounding variables which we believe are predictive of the pace $X = (X_h, X_r)'$. With this structure, we wish to specify propensity scores $\text{Prob}(X_h - X_r > 0 \mid W)$ that describe the probability that the home team plays at greater pace than the road team given the relevant circumstances of the match. With insights gained from the EDA of Section 2, we specify a statistical model that leads to propensity scores. For reference, we define all of our relevant variables below.

$$\begin{array}{cl} t & \equiv \text{time of the match in minutes, } \ t \in (0, 90) \\ X(t) & \equiv \text{pace vector for home and road teams at time } t\text{; either } GP \text{ or } AP \\ GD(t) & \equiv \text{goal differential in favour of the home team at time } t \\ O & \equiv \text{pre-match betting odds corresponding to the home team} \\ Y((t_1, t_2)) & \equiv \text{excess shots by home team compared to road team during } (t_1, t_2) \end{array} \qquad (4)$$

Looking ahead, our interest is in determining a cause-effect relationship regarding the impact of pace X on success Y. A natural success variable would be goals. However, in soccer, goals are rare events with less than three goals per game on average in top professional leagues. We therefore use the surrogate variable shots as defined in (4) to assess success. Of course, not all shots lead to goals but shots are an indication of success. However, let's return to the first step of the causal investigation which involves the construction of a propensity score model.

We first bin the data to define levels for each of the three confounding variables $W(t) = (t, GD(t), O)'$. We segment the time t into 18 five-minute intervals: $(0,5)$, $(5,10)$, $\ldots$, $(85,90)$. We do not include added time beyond 90 minutes since the amount of added time differs across matches.

For the second variable, we restrict $GD(t)$ to five states with goal differentials -2, -1, 0, 1 and 2 corresponding to the home team at time t. Note that $GD(t) = -2$ corresponds to the home team losing by two or more goals and that $GD(t) = 2$ corresponds to the home team winning by two or more goals. For a given match, we consider each of the 18 time intervals, and if the goal differential is constant throughout the interval (either -2, -1, 0, 1 or 2), then an observation is recorded.

For the third variable O, 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, we define four bins for the decimal odds of the home 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 across bins. Note that the bettings odds O do not depend on the time t.

The variable O 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 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 difference in odds in these two situations is typically minor.

For the response variable in the propensity score model, we calculate $X_h(t)$ and $X_r(t)$ during the time interval which is intended to convey the style of pace over the time period.

We use attacking pace AP for the pace calculation, as it is more definitive and perhaps more interesting that general pace GP. We also emphasize that the response variable $X(t) = (X_h(t), X_r(t))'$ is bivariate which makes the causal investigation nonstandard.

To illustrate the variables in the propensity score model, consider a match where the score is 2-0 just prior to the 70-th minute. Following conventional notation where the first team in the scoreline is the home team, this implies that the home team is leading by two goals. Assume further that the home team is the favoured team with pre-match decimal betting odds $o = 1.5$. In this match, suppose that neither team scores during the time interval $(65, 70)$ minutes, and that the AP statistics for the home and road teams during this period are 2.34 and 2.07, respectively. Then, for this time interval, we have the observed response $X = (2.34, 2.07)$ and covariates $W = (14, 2, 1)$ where $t \in (65, 70)$ corresponds to the 14th time category, $GD = 2$ is the goal differential (categorical), and odds $o = 1.5 \in (1.3, 1.7)$ corresponds to the first category.

Based on the above considerations, we have 3679 observations recorded across $18 \times 5 \times 4 = 360$ cells. For linear models based on categorical data, it is prudent to have adequate numbers of observations in each cell. For this reason, we consider a reduction in the number of cells by instead defining six time categories, $(0, 15), (15, 30), \ldots, (75, 90)$ minutes. In this case, we have 944 observations recorded across $6 \times 5 \times 4 = 120$ cells. The cell counts are provided in Table 1. In most cells, we have the recommended minimum number of five counts per cell; exceptions tend to occur with large goal differentials (i.e. $GD = -2$ and $GD = 2$), especially early in matches.

Our propensity score model is a multivariate analysis of variance (MANOVA) model where the response variable X is two-dimensional and the covariate $W = (t, GD, O)$ has $6 \times 5 \times 4 = 120$ cells (as described above). The MANOVA model is preferred to two separate ANOVA models for X_h and X_r since the MANOVA model permits a covariance structure between X_h and X_r. Details on MANOVA models are given by Smith, Gnanadesikan and Hughes (1962).

We used MANOVA software using the `manova` function in the `Stats` R package. One of the assumptions of MANOVA concerns the normality of observations. A quantile plot of the residuals does not suggest any serious departures from normality. In Table 2, we present the results of fitting the MANOVA model where we have allowed for the possibility of first-order interaction terms. The main takeaway is that the time of the match t, the

	$GD=-2$				$GD=-1$				$GD=0$				$GD=1$				$GD=2$			
O	[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)	[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)
$t \in (00,15)$	0	0	0	2	0	0	0	3	48	51	33	37	6	2	1	2	0	0	0	0
$t \in (15,30)$	0	0	0	3	6	3	2	8	32	32	19	24	14	6	5	7	2	1	0	0
$t \in (30,45)$	1	0	0	3	6	3	4	5	24	24	17	20	17	9	7	10	6	2	1	0
$t \in (45,60)$	0	0	2	8	5	5	6	7	18	18	11	16	18	14	5	8	12	1	2	2
$t \in (60,75)$	1	1	3	13	3	4	3	12	15	15	7	7	12	16	8	5	14	5	4	3
$t \in (75,90)$	1	3	4	7	3	4	4	11	9	16	6	8	13	9	6	4	16	3	3	1

Table 1: Cell counts for the $6 \times 5 \times 4$ covariate categories where the categories correspond to the time t, the goal differential GD and the betting odds O.

goal differential in favour of the home team GD and the relative strength of the home team O are strongly associated with attacking pace X. There is also mild evidence of some first-order interactions involving t, GD and O.

Variable	Df	Pillai	approx F	num Df	den Df	Pr($> F$)
t	5	0.062411	5.7077	10	1772	1.801e-08 ***
GD	4	0.075761	8.7208	8	1772	9.575e-12 ***
O	3	0.126651	19.9665	6	1772	<2.2e-16 ***
$t{*}GD$	18	0.050637	1.2786	36	1772	0.12531
$t{*}O$	12	0.054758	2.0784	24	1772	0.00164 **
$GD{*}O$	15	0.035338	1.0624	30	1772	0.37509
Error	886					

Table 2: Results from the MANOVA which relates pace X to the covariates $W = (t, GD, O)$.

Finally, we need to induce the required probability $\text{Prob}(X_h - X_r > 0 \mid W)$ from the fitted MANOVA model. The calculation is based on a simple result from mathematical statistics using properties of the normal distribution. For example, for a given match situation W, suppose that the MANOVA model yields $X \sim \text{Normal}_2(\mu, \Sigma)$ where $\mu = (\mu_1, \mu_2)'$ and $\Sigma = (\sigma_{ij})$. Then $\text{Prob}(X_h - X_r > 0 \mid W) = \Phi((\mu_1 - \mu_2)/\sqrt{\sigma_{11} + \sigma_{22} - 2\sigma_{12}})$ where Φ is the cumulative distribution function of the standard normal distribution. Note that the bivariate normal parameters are estimated through the fitting of the MANOVA model. For example, the estimated values of σ_{11}, σ_{22} and σ_{12} are 0.0051, 0.0061 and 0.0003, respectively. Therefore, the estimated correlation between home and road attacking pace is $\sigma_{12}/(\sqrt{\sigma_{11}\sigma_{22}}) = 0.054$ which indicates that the MANOVA formulation (which takes into account the relationship between home and road pace) provides only a slight improvement over ANOVA.

5.2 Matching and Results

In the most basic randomized experiment, an experimenter randomly assigns M subjects from a population to receive a treatment and M subjects from the population to receive the control. The hope is that through random assignment, the treatment group will on average be similar to the control group, and that differences in the response between the two groups can be attributed to the treatment.

The use of propensity scores and matching (Austin 2011, Imbens 2004) attempts to mimic the basic randomized experiment in the context of observational studies. A propensity score for a subject in a clinical trial is the probability that the subject receives the treatment. In the pace problem, $\text{Prob}(X_h - X_r > 0 \mid W)$ is the estimated probability that the home team will play at a higher pace than the road team. Therefore, $\text{Prob}(X_h - X_r > 0 \mid W)$ serves as the relevant propensity score in the pace application.

In our problem, we have a dataset involving 944 pace observations (see Section 5.1) resulting in $M_1 = 450$ cases where the home team plays at greater pace (the treatment) and $M_2 = 494$ cases where the home team plays at lesser pace (the control). Since $M_1 < M_2$, the matching idea is that we attempt to match each of the M_1 treatment cases with a corresponding control case so that each pair has a similar estimated propensity score based on the underlying match circumstances W. Then the resulting two groups (M_1 treatments and M_2 controls) will be similar in the match characteristics, and that differences between the two groups can be attributed to the treatment (i.e. pace).

There are many ways that the matching of propensity scores can be carried out (Stuart 2010), and caution ought to be exercised in the process. In our application, we begin with the M_1 cases where the home team plays at a greater pace, and we use a nearest neighbor method for selecting the matched cases where the home team plays at a lesser pace.

Specifically, we use the *Matching* package (Sekhon 2011) in the statistical programming language R to randomly select (with replacement) control cases that fall within a specified tolerance of the propensity scores for the treatment cases. Sampling with replacement tends to increase the quality of matching when compared to sampling without replacement. Unlike deterministic matching procedures, the random aspect of the nearest neighbor procedure allows us to repeat analyses to check the sensitivity of the inferences.

Following the implementation of the matching procedure, Figure 8 displays the balance between the two groups with respect to the propensity scores. The similarity in the histograms is important as it provides confidence that the two groups are similar according to the characteristics that affect whether the home team plays at greater pace.

The inferential component of the investigation begins with a simple paired two-sample test between the two groups based on the response Y (excess shots by the home team) as described in (4). Again, we prefer to use shots rather than goals since goals are rare events. The quantity of interest is the average treatment effect $\text{ATE} = \bar{Y}(1) - \bar{Y}(0)$ where $\bar{Y}(1)$ is the excess number of resultant shots by the home team when they are playing at greater pace, and $\bar{Y}(0)$ is the excess number of resultant shots by the home team when they are playing at lesser pace. We obtain $\text{ATE} = 0.73 - 0.41 = 0.32$ with standard error 0.103. The result is significant and suggests that pace is beneficial in the sense of playing at a higher attacking pace.

To put the above result into context, suppose that the home team outpaces the road team during all six 15-minute intervals during the match. Then, we would expect the home team to have roughly $6(0.32) = 2$ more shots during the match than the road team. Note also that we have been careful to distinguish the home and road teams. If we flipped the analysis to consider the average treatment effect due to the road team playing at pace, we would obtain $\text{ATE} = -0.41 - (-0.73) = 0.32$. Therefore, the benefit of outpacing the opposition applies to either team.

In Figure 9, we present a more nuanced view of the situation. For each group (treatment and control), we smooth the variable Y with respect to the propensity score. We observe that as the propensity score increases (i.e. conditions become more favourable for the home team to play at greater pace), the excess shots for the home team increases for both groups. We also observe that the excess shots by the home team remains relatively constant across the two groups as the propensity score increases. In practice, this means that the advantage of playing at pace persists no matter the circumstances that dictate whether a team should play at pace.

Therefore, the takeaway message is that playing with pace is a good strategy. It leads to more shots for than against. This provides support to Blank's thesis - the Holy Grail of tactics (in Chapter 1 of Blank (2012)) that fast is better than slow.

6 DISCUSSION

Despite its importance, style of play is an understudied aspect in team sport. In this paper, we investigate pace of play as it relates to soccer. Although the analyses were restricted to the study of tracking data in the Chinese Super League, it is conjectured that the broad results hold true for other high-level professional soccer leagues.

In particular, we found that teams that play at higher attacking pace are more advantaged in producing shots than teams that play at lower pace. For a team that outpaces its opponent throughout a match, this translates to roughly two extra shots. The conclusion was facilitated through the adaptation of causal methods. In particular, we sought confounding variables that were important in determining propensity scores. Furthermore, the propensity scores were obtained by reducing a bivariate normal distribution to a relevant Bernoulli distribution. The EDA produced additional sporting insights (A-E) related to pace.

There are possible future investigations related to pace of play. For example, we believe that similar analyses may be carried out in other invasion sports where tracking data are available. Also, it may be interesting to analyze pace separately in terms of passing and dribbling. The ball generally moves more quickly when passing, and there may be stylistic differences between teams in terms of how much they pass relative to how much they dribble.

A limitation in our work is that the response variable Y (shots) in the causal analysis correspond to rare events and is known to be noisy. A better response variable may be expected goals (Spearman 2018, Anzer and Bauer 2021), and this could be considered in future investigations. Another limitation of our work is the restriction to matches from the CSL. It would be good to see if the results also hold in top-level European leagues where the best players from all over the world compete. Although we argue that confounding variables can be identified with tracking data, for sure, there are latent variables that we have not utilized (e.g. level of player fatigue). This is also a limitation.

7 REFERENCES

Anzer, G. and Bauer, P. (2021). A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). *Frontiers in Sport and Active Living*, 3, Article 624475.

Austin, P.C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46, 399-424.

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

de Jong, L.M.S., Gastin, P.B., Angelova, M., Bruce, L. and Dwyer, B. (2020). Technical determinants of success in professional women's soccer: A wider range of variables reveals new insights. *PLOS ONE*, 15(10), e0240992.

Goes, F.R., Brink, M.S., Elferink-Gemser, M.T., Kempe, M. and Lemmink, K.A.P.M. (2021). The tactics of successful attacks in professional association football: Large-scale spatiotemporal analysis of dynamic subgroups using position tracking data. *Journal of Sports Sciences*, 39(5), 523-532.

Guan, T., Cao, J. and Swartz, T.B. (2022). Should you park the bus? Manuscript under review. Available at https://www.sfu.ca/∼tswartz/

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

Imbens, G.W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86, 4-29.

Kempe, M., Vogelbein, M., Memmert, D. and Nopp, S. (2014). Possession vs. direct play: Evaluating tactical behavior in elite soccer. *International Journal of Sports Science* 4(6A), 35-41.

Lepschy, H., Wäsche, H. and Woll, A. (2021). Success factors in football: an analysis of the German Bundesliga. *International Journal of Performance Analysis in Sport*, 20(2), 150-164.

McLellan, I. (2010). Total football: Whatever happened to the beautiful game? *Bleacher Report: World Football*, Accessed June 2, 2022 at https://bleacherreport.com/articles/321814-beautiful-game-what-ever-happened-to-total-football

Merlin, M., Cunha, S.A., Moura, F.A., Torres, R., Gonçalves, B. and Sampaio, J. (2020). Exploring the determinants of success in different clusters of ball possession sequences in soccer. *Research in Sports Medicine*, 28(3), 339-350.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference, Second Edition.* Cambridge University Press: Cambridge.

Sekhon, J.S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software*, 42, 1-52.

Shaw, L. and Glickman, M. (2019). Dynamic analysis of team strategy in professional football. *Barça Sports Analytics Summit.*

Shen, E., Santo, S. and Akande, O. (2022). Analyzing pace-of-play in soccer using spatio-temporal event data. *Journal of Sports Analytics*, 8(2), 127-139.

Silva, R., Davis, J. and Swartz, T.B. (2018). The evaluation of pace of play in hockey. *Journal of Sports Analytics*, 4, 145-151.

Smith, H., Gnanadesikan, R. and Hughes, J.B. (1962). Multivariate analysis of variance (MANOVA). *Biometrics* 18(1), 22–41.

Spearman, W. (2018). Beyond expected goals. *Proceedings of the 2018 MIT Sloan Sports Analytics Conference*, Accessed October 6, 2022 at https://www.researchgate.net/publication/327139841_Beyond_Expected_Goals

Tweedale, A. (2022). Counter-pressing and the gegenpress: football tactics explained. *The Coaches Voice: Coaching Knowledge*, Accessed June 2, 2022 at https://www.coachesvoice.com/cv/counter-pressing-gegenpressing-football-tactics-explained-klopp-guardiola-bielsa-hasenhuttl/

Stuart, E.A. (2010). Matching methods for causal inference. A review and a look forward. *Statistical Science*, 25(1), 1-21.

Wilson, J. (2013). *Inverting the Pyramid*, Nation Books, New York.

Wu, Y., Danielson, A., Hu, J. and Swartz, T.B. (2021). A contextual analysis of crossing the ball in soccer. *Journal of Quantitative Analysis in Sports*, 17(1), 57-66.

Yu, D., Boucher, C., Bornn, L. and Javan, M. (2019). Evaluating team-level pace of play in hockey using spatio-temporal possession data. *Proceedings of the 2019 MIT Sloan Sports Analytics Conference*, accessed October 18, 2021 at https://arxiv.org/pdf/1902.02020.pdf

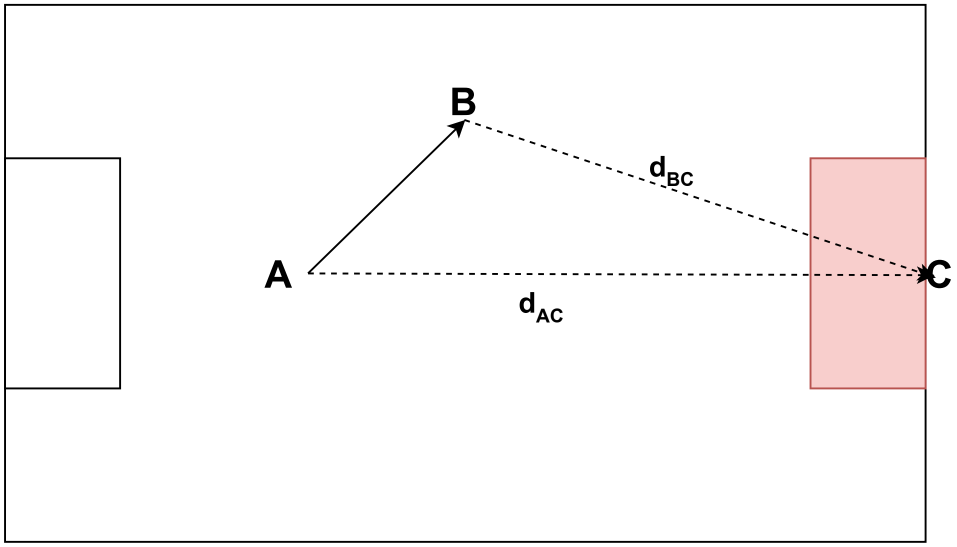

Figure 1: The plot illustrates a pass with attacking intent. A is the starting point of the pass, B is the end point of the pass and C denotes the middle of the goal line of the opponent. The values d_{BC} and d_{AC} represent the distances from B to C, and A to C, respectively. Attacking pace AP for this component of play is obtained using the distance $d_{AC} - d_{BC}$.

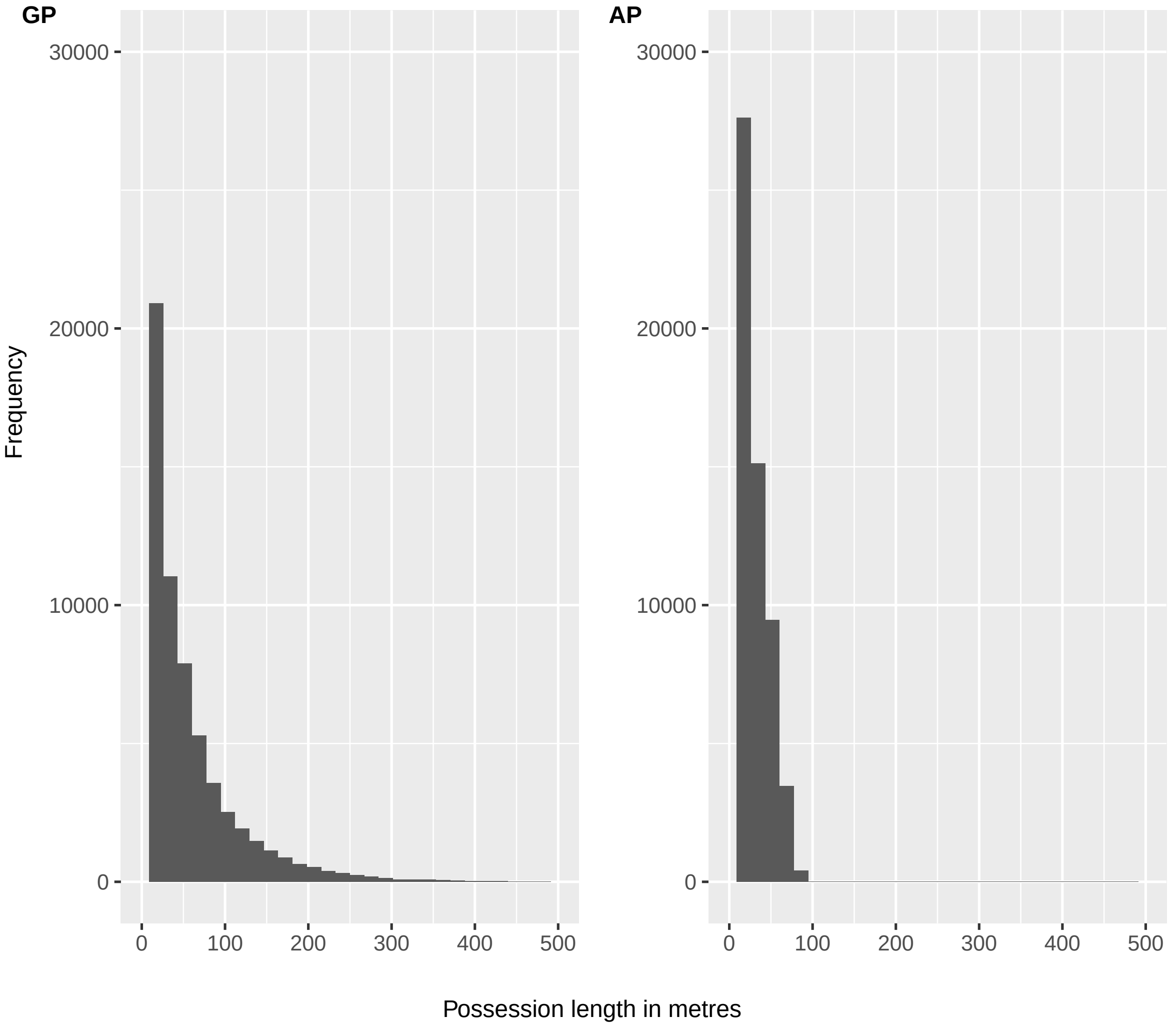

Figure 2: Histogram of the lengths of all possession sequences in metres corresponding to GP and AP, respectively.

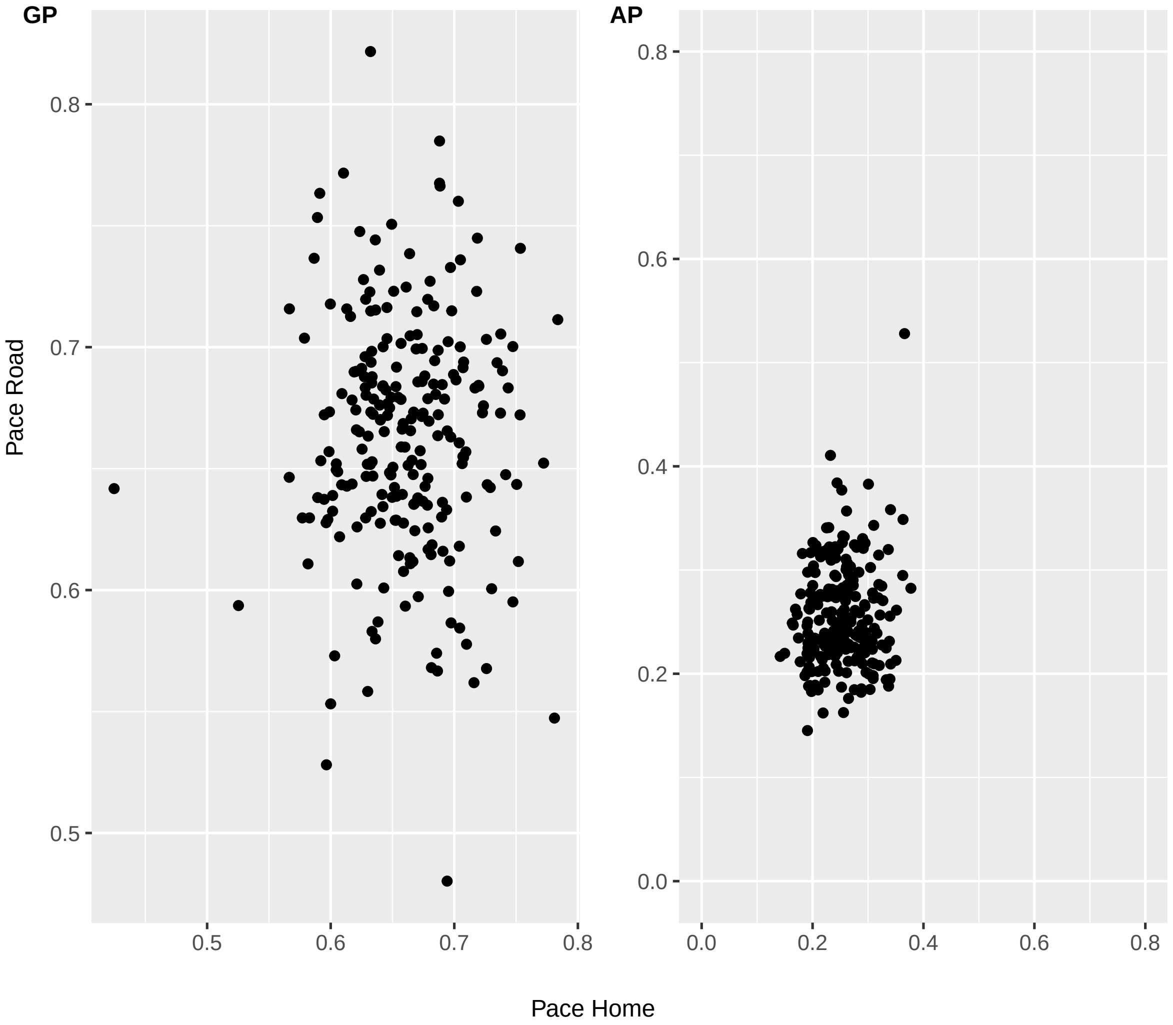

Figure 3: Scatterplots for GP and AP related to the home and road teams for each match.

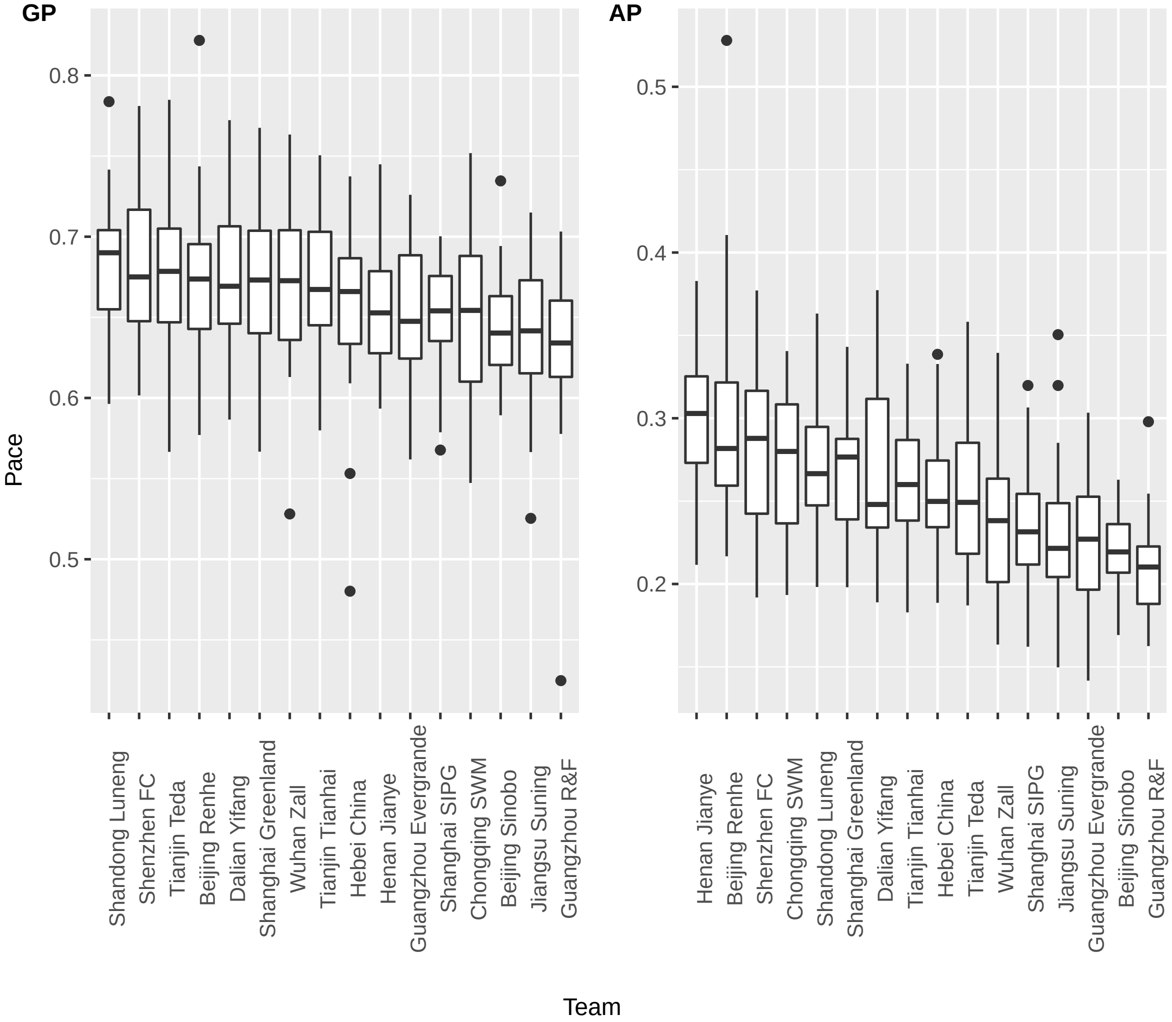

Figure 4: Boxplots of GP and AP for each of the 16 teams in the CSL based on their 30 matches.

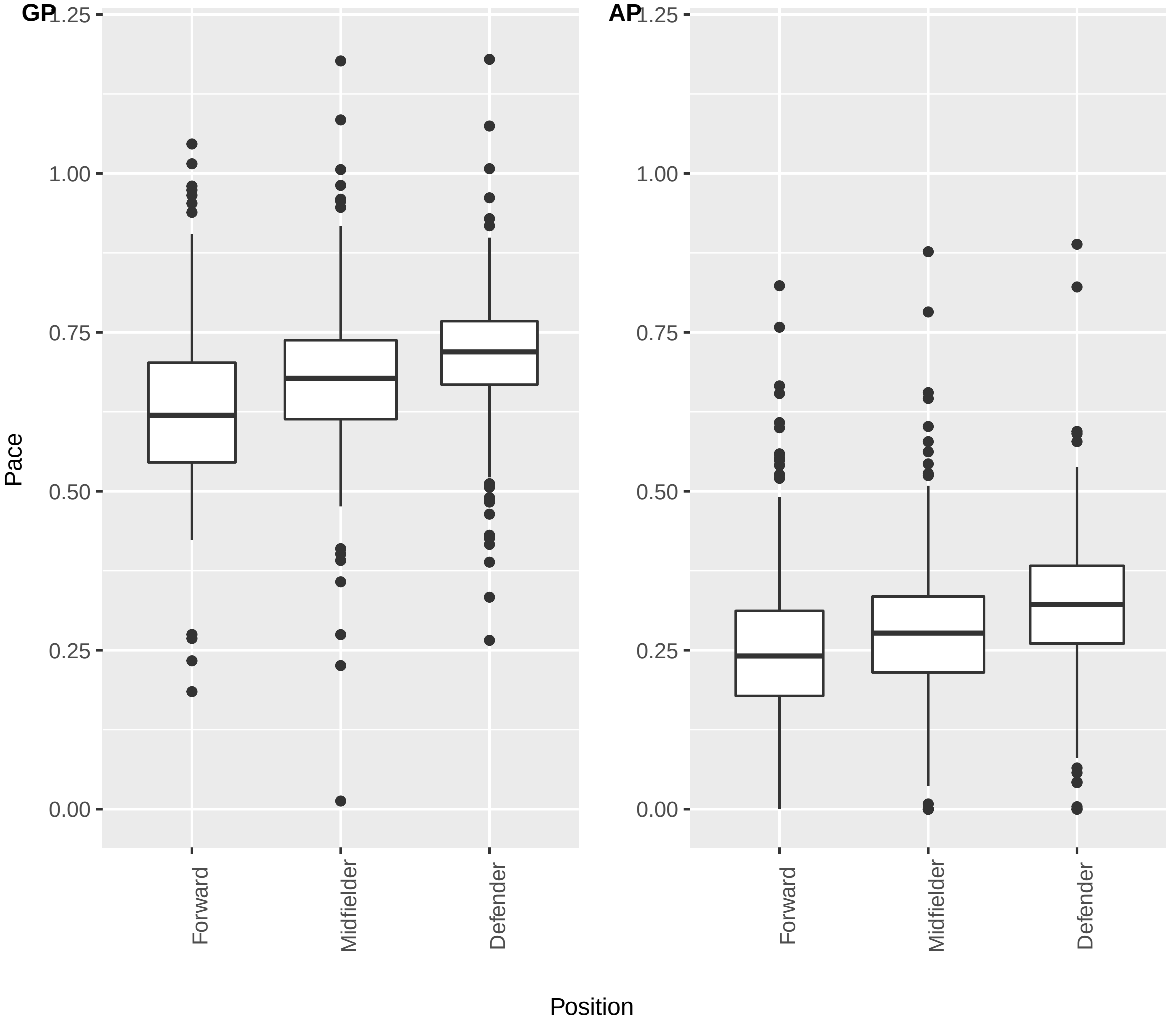

Figure 5: Boxplots of GP and AP for forwards, midfielders and defenders based on their individual match statistics.

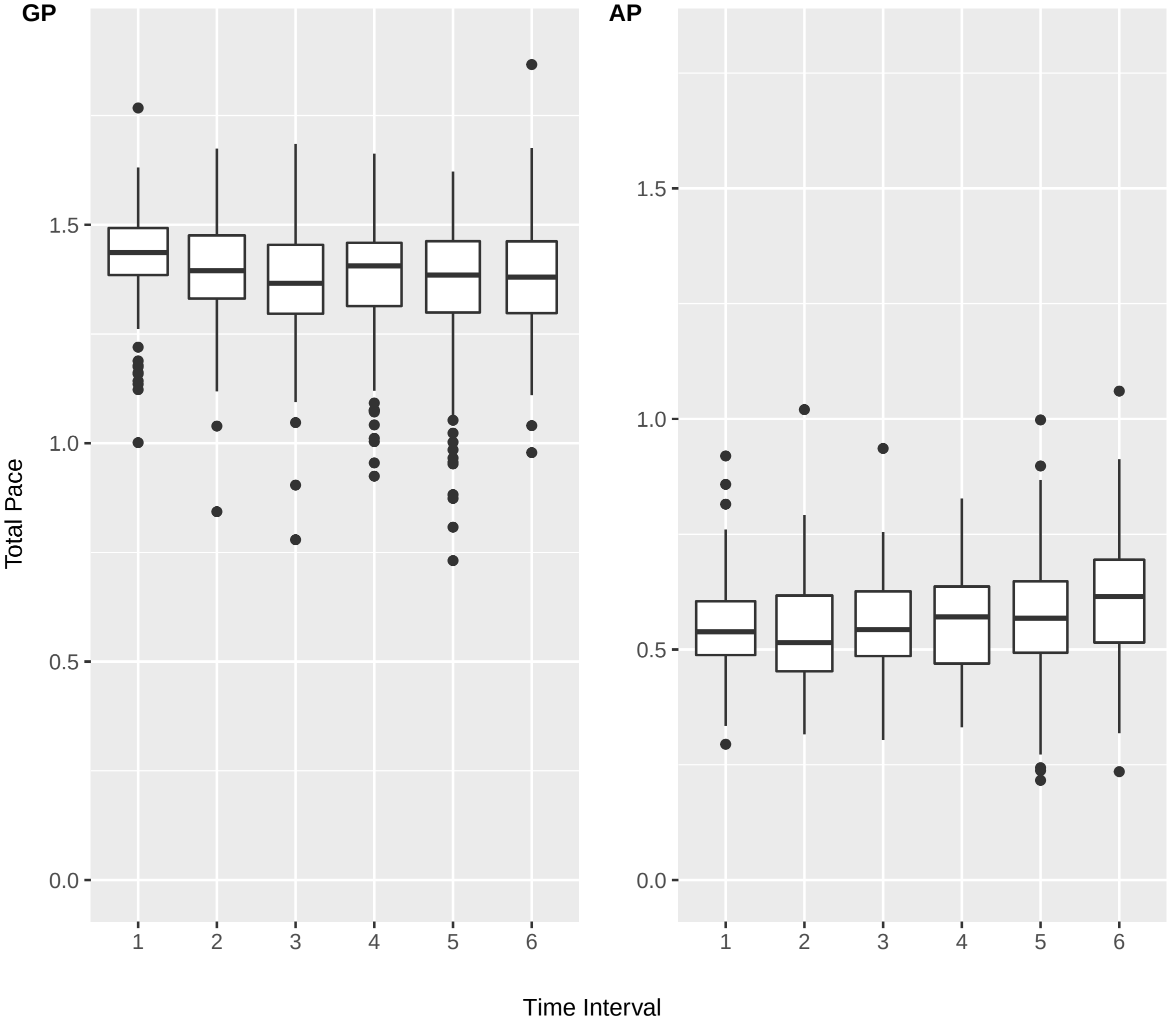

Figure 6: Boxplots of GP and AP according to the time of the match where time is divided into six 15-minute intervals from 0 to 90 minutes. The pace calculations are the total pace corresponding to both teams.

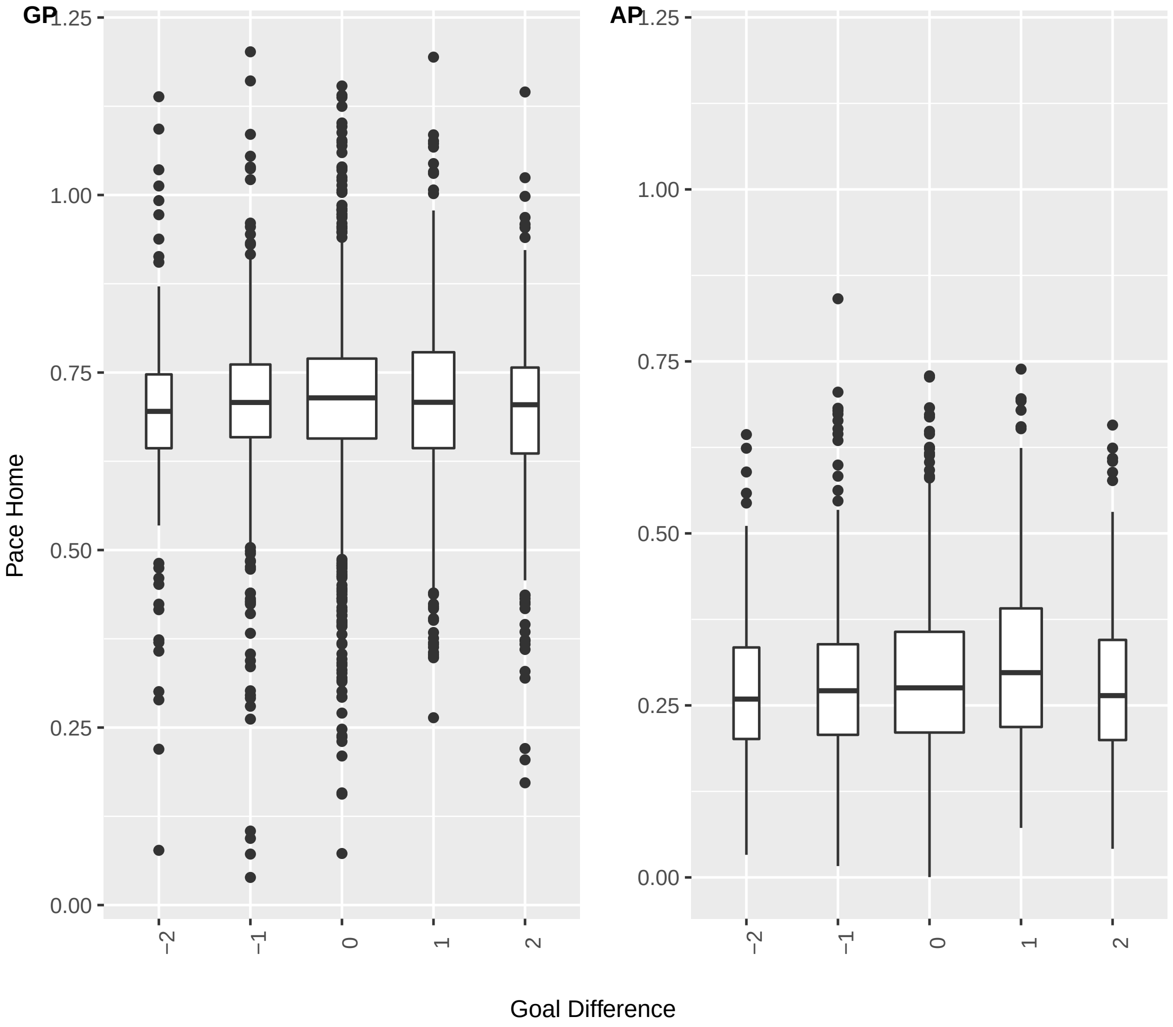

Figure 7: Boxplots of GP and AP corresponding to the goal differential (GD) taken at 5-minute intervals where -2 corresponds to the home team losing by 2 or more goals, -1 corresponds to the home team losing by 1 goal, 0 indicates a tied match, 1 corresponds to the home team winning by 1 goal and 2 corresponds to the home team winning by 2 or more goals.

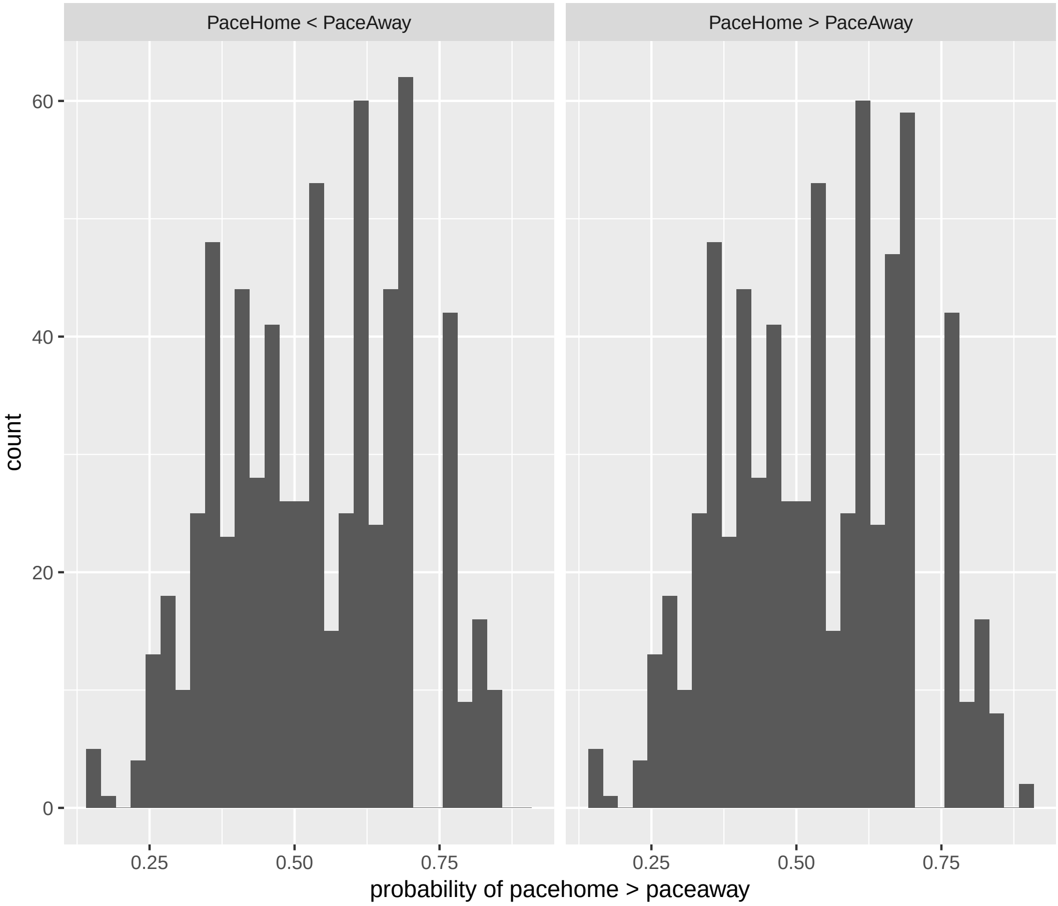

Figure 8: After matching, histograms of the two groups (treatment and control) are depicted where the horizontal variable is the propensity score.

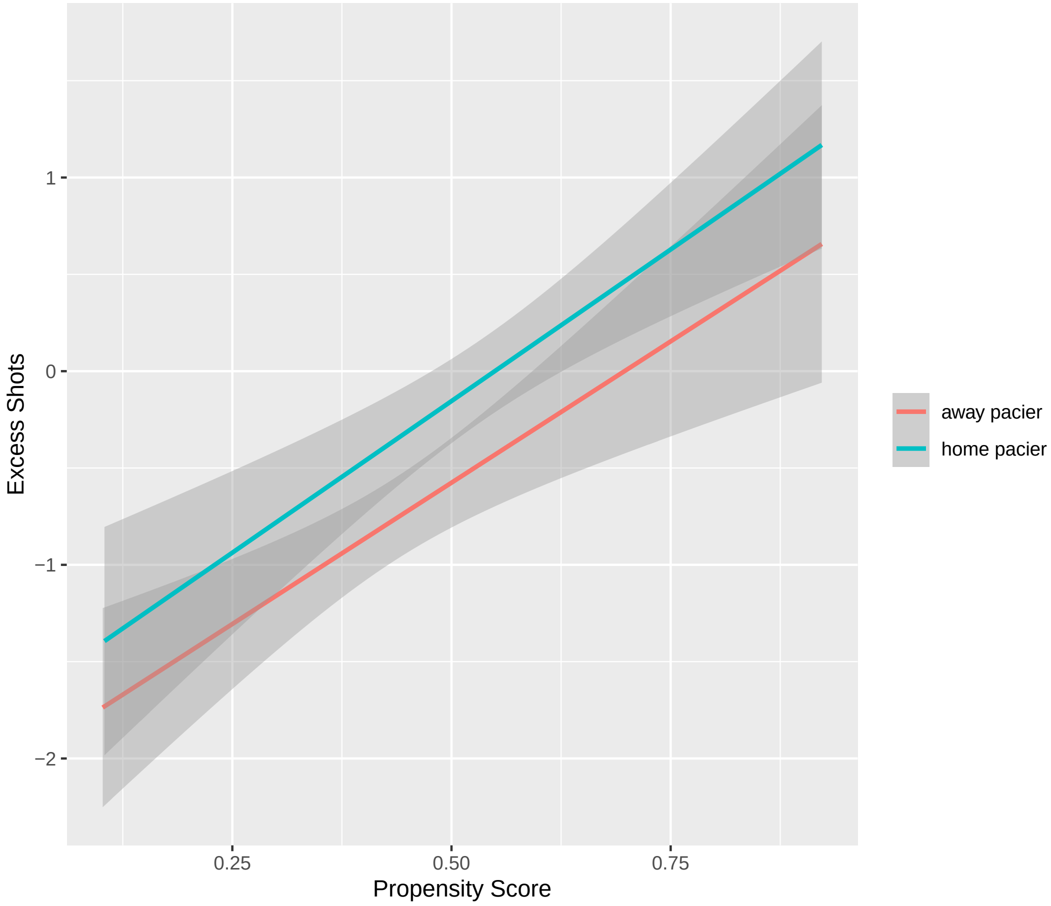

Figure 9: After matching, smoothed plots of the excess shot variable Y for the home team with respect to the propensity score under the treatment (blue) and the control (red).