A Contextual Analysis of Crossing the Ball in Soccer

Lucas Y. Wu, Aaron J. Danielson, X. Joan Hu and Tim B. Swartz *

Abstract

The action of crossing the ball in soccer has a long history as an effective tactic for producing goals. Lately, the benefit of crossing the ball has come under question, and alternative strategies have been suggested. This paper utilizes player tracking data to explore crossing at a deeper level. First, we investigate the spatio-temporal conditions that lead to crossing. Then we introduce an intended target model that investigates crossing success. Finally, a contextual analysis is provided that assesses the benefits of crossing in various situations. The analysis is based on causal inference techniques and suggests that crossing remains an effective tactic in particular contexts.

Keywords: Association football, Causal inference, Event data, Player tracking data.

*L. Wu is a PhD candidate, A. Danielson is a Postdoctoral Fellow, and X.J. Hu and T. Swartz are Professors, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6. Hu and Swartz 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 data used in this paper.

1 Introduction

The sport of soccer (association football) has a long history dating back to 1863 when the Laws of the Game were codified by the Football Association in England. Throughout the history of the sport, tactics have evolved with the intention of providing a competitive advantage (Wilson 2013). As a strategy, the action of crossing the ball in soccer has always been a staple of the game that has been thought to produce goals. A crossed ball occurs when a player (normally situated in a wide area of the attacking third of the pitch) kicks the ball towards the box with the intention that an attacking teammate will score.

However, in recent years, research has been carried out that casts doubt on the benefits of crossing the ball. Vecer (2014) provides a persuasive argument that the overall effect of crossing the ball has a strong negative impact on scoring. Vecer (2014) uses both aggregate crossing statistics and multilevel Poisson regression to study the impact of crossing. In the analyses, there is a suggestion that crossing (when executed properly) is valuable; however, the rate of bad crosses greatly exceeds the rate of good crosses, and this is a primary argument against crossing. Vecer (2014) also demonstrates that missed scoring opportunities due to open crossing is associated with the quality of the attacking team. In recent years, teams have become more reluctant to cross the ball. For example, Vecer (2014) states that the number of open crosses in the German Bundesliga dropped from 12.0 per match in the 2009/2010 season to 8.9 per match in the 2015/2016 season, a decrease exceeding 25%.

Sarkar (2018) investigates crosses from a game theoretic perspective. They assume the attacking team can cross the ball or not, and the defending team can utilize an offside trap or not. The vector of equilibrium strategies determines the probabilities of the possible outcomes. Somewhat surprisingly, Sarkar (2018) suggests that teams that are good at aspects of executing a cross should cross the ball less often. Sarkar (2018) and Sarkar and Chakraborty (2018) also confirm the inverse relationship between the number of crosses and the number of goals scored in a match. Other papers that have provided nuanced views on the negative effects of crossing include Liu et al. (2015) and Oberstone (2009).

Given the longstanding history of crossing the ball in soccer, the conclusions reached by Vecer (2014) and Sarkar (2018) have been surprising to many, including the authors of this paper. We hypothesize that there are contexts in which crossing the ball in soccer is a beneficial strategy. Knowing when to the cross the ball is a step in the direction of effective

playing strategy. Our contextual 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. Player tracking data in sport are the catalysts for big data analyses and do not form part of the analyses by Vecer (2014) and Sarkar (2018). 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 analysis of player tracking data has been particularly prominent in the sport of basketball; see for example, Miller et al. (2014).

Although tactical decisions are a fundamental aspect of sport, sporting decisions are not typically based on the results of randomized designs, the bread and butter of causal inference. Clearly, in professional sport, match outcomes are important and coaches would be unwilling to implement a tactic in a random selection of games and then implement an alternative tactic in a remaining subset of games. There are many approaches that estimate causal effects with observational data (see Pearl 2009), but these methods have not received much attention in the sports analytics literature. One exception is the work of Yam and Lopez (2020) who investigate the impact of "going for it" on fourth down in the National Football League as opposed to punting or kicking a field goal. Their approach is based on matching propensity scores and covariates associated with game situations. As another example, Toumi and Lopez (2019) use propensity score matching and Bayesian additive regression trees to estimate the causal effects of zone-entry decisions in the National Hockey League.

Our work uses spatio-temporal data to investigate three aspects of the crossing problem in soccer. First, we investigate the spatio-temporal conditions that lead to crossing. Then we introduce an intended target model that investigates crossing success. Finally, a contextual analysis is provided that assesses the benefits of crossing in various situations. The analysis is based on causal inference techniques and suggests that crossing remains an effective tactic in particular contexts.

Section 2 introduces the dataset. We outline the steps involved in converting the player tracking data into features that are used in the ensuing analyses. The resultant design matrix consists of rows that correspond to crossing opportunities and columns (covariates) that are believed to related to aspects of crossing. Our analysis is based on various assumptions used

in the definition of a crossing opportunity and on the definition of outcomes arising from crossing opportunities. In cases where the rationale for the assumptions is less clear, we introduce tuning parameters so that analyses can be carried out using a range of values of the tuning parameters.

Section 3 is concerned with the spatio-temporal conditions that lead a player to cross the ball. We develop a logistic regression model which relates the attempt (or non-attempt) to cross the ball to covariates (situational variables) which are believed to be related to the crossing decision. We observe that the model makes physical sense according to our understanding of soccer. The fitted model provides evidence of the rich information embedded in the player tracking data. The logistic model is subsequently used in the causal analysis of Section 5.

Section 4 develops an intended target model. The model introduces additional covariates that are relevant to the probability of success of a cross. The analysis concerns a sender (the player contemplating the cross) and potential receivers (players to whom the cross may be intended). The intended target model provides insight to whom a cross ought to be made. Again, the fitted model aligns with our understanding of soccer. The information gleaned from the model may benefit players and coaches in terms of tactical decisions.

In Section 5, we first review concepts needed to apply basic causal inference techniques to the crossing problem. Then we use propensity score matching to assess whether crossing is beneficial. Our results are nuanced as crossing is seen to be beneficial in particular circumstances, and these circumstances are those when a player is more likely to cross. We therefore see that the intuition of soccer players involving the decision to cross corresponds to good decision making. And importantly, we dispel the notion that crossing is not a valuable tactic in soccer.

Some concluding remarks are then provided in Section 6.

2 Data Preprocessing

Statistical analyses begin with the existence of a dataset. However, with big data, the preprocessing of data has become an integral part of statistical practice that defines the types of models and analyses that can be entertained.

In this paper, we have a big data problem where both event data and player tracking data are analyzed based on the 30 regular season matches of the 2017 season for Shandong Taishan Luneng FC of the Chinese Super League. Event data and tracking data are collected independently where event data consists of occurrences such as tackles and passes, and these are manually recorded along with auxiliary information whenever an "event" takes place. Both event data and tracking data have timestamps so that the two files can be compared for internal consistency. In the Shandong Luneng dataset, tracking data are obtained from the use of optical recognition software. The Shandong Luneng 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 at a given time. Although the inferences gained via our analyses are specific to Shandong Luneng, it is plausible that some of the broad insights may hold generally to high level soccer competitions.

2.1 Defining Crossing Opportunities

Vecer (2014) suggests that there are alternative strategies to crossing that are more beneficial in terms of goal scoring. These strategies include attacking through the center of the pitch (via dribbling and passing) and shooting.

Vecer (2014) also states that when the attacking team enters the final third of the pitch, various options are more or less open. We focus on this assumption in our analysis. In particular, we utilize event and player tracking data to define a crossing opportunity. We define a crossing opportunity to be an occasion where a player has possession of the ball in a potential crossing zone and has the opportunity to either cross or not cross the ball. Also, we record covariates that describe the relevant circumstances at the time of each crossing opportunity.

Soccer is a fluid game where events frequently occur. Following Bransen, Van Haaren and van de Velden (2019), we define a possession sequence as a sequence of events involving possession of the ball by the same team. A possession sequence concludes with a change of possession or a stoppage. In our dataset, the length of a possession sequence ranges up to 19 events.

We begin by restricting our crossing analysis to occasions when the offensive team retains possession in a wide position of an attacking third of the pitch (i.e. within 13.85 metres of the

sideline). We refer to these two regions (on the opposite sides of the field) as the potential crossing zones, which are highlighted in blue in Figure 1. We are interested in the segment of possession sequences in the blue region. Only in these segments is it possible to cross the ball.

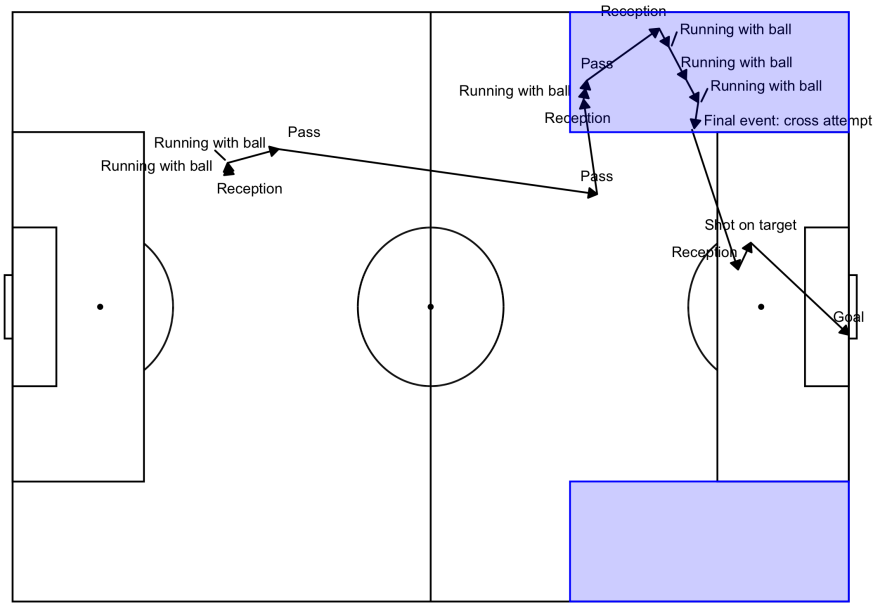
After restricting our analysis to possessions in the potential crossing zones, we identify the final event that occurred in the zone, and we record the spatio-temporal information of all players at that moment. The last event in the potential crossing zone will be either a cross or non-cross (i.e. pass or dribble).

In particular, we remove possession sequences that correspond to corner kicks and free kicks. Note that corner kicks and free kicks are not open crosses, but could possibly occur in a wide position of an attacking third of the pitch. We have $N = 2225$ final events in potential crossing zones throughout the 30 matches.

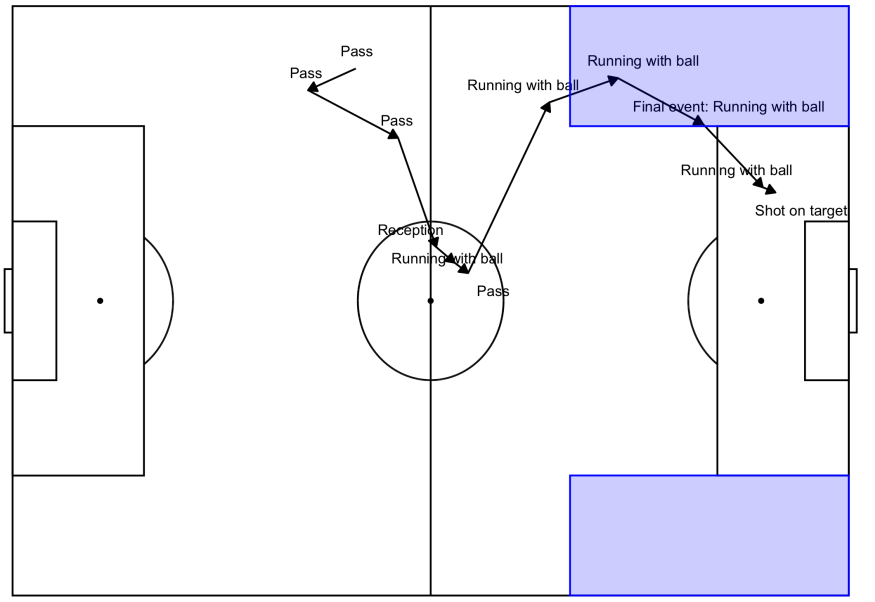
2.2 Crafting Situational Variables

Building on previous research that evaluates passing ability (Szczepanski and McHale 2016, Power et al. 2017), we propose variables specific to the context of crossing.

It is a tenet of soccer that time and space are paramount factors that lead to improved attacking outcomes. From the tracking data, it is possible to determine the location and velocity of both the ball and the player of interest. The location and velocity measurements form the basis for the situational variables presented in Table 1. Recall that the situational variables $\tau, z_1, ..., z_9$ form the columns of a design matrix Z where the rows of Z are crossing opportunities corresponding to the final event in a possession sequence occurring in potential crossing zones. Although the situational variables in Table 1 are self-explanatory, the variable z_2 (nearest defender distance) is a measure of defensive pressure on the sender. However, it does not account for the situation where multiple defenders are covering the sender. The variable z_3 indicating space controlled by the sender has been introduced using ideas from Fernandez and Bornn (2018) and Fernandez et al. (2019). Although we experimented with many other crossing variables, the variables presented in Table 1 are those that provided excellent fit for the logistic model of Section 3.



(a)



(b)

Figure 1: Examples of possession sequences with (a) a crossing attempt and (b) without a crossing attempt.

Variable	Definition of Variable
$\tau = 1\,(0)$	- the ball is crossed (not crossed)
z_1	- score differential wrt the team in possession
z_2	- distance between the sender and nearest defender
z_3	- space controlled by the sender
z_4	- distance between the sender and nearest teammate
z_5	- distance between the sender and the endline
z_6	- ratio of the number of offensive players to defensive players in the box
z_7	- indicator variable corresponding to whether the sender is a defender
z_8	- indicator variable corresponding to whether the sender is a midfielder
z_9	- indicator variable for last 10 minutes of a half

Table 1: A subset of situational variables relevant to crossing which form the columns of the design matrix Z. All distances are measured in metres.

2.3 Outcome Variable

We require a response variable that allows us to assess whether crossing is beneficial. The obvious candidate is the variable $Y_1 = 1(0)$ according to whether a crossing opportunity led (did not lead) to a goal. Although scoring and preventing goals is the primary objective of soccer teams, goal scoring is a rare event with only 2.5-3.0 goals scored per game on average in top European soccer leagues. Therefore, it is difficult to tease out subtle inferences when goal scoring is used as the dependent variable.

Alternative indicator variables that we have considered for a response variable are whether a crossing opportunity led to a shot on goal Y_2 and whether a crossing opportunity led to a shot Y_3. The variable Y_2 is more common than Y_1 and Y_3 is more common than Y_2. For this reason, we prefer the response variable $Y = Y_3$. We note that shot statistics (as opposed to goal statistics) are prevalent in the hockey analytics literature and are referred to as Fenwick and Corsi (Vollman, Awad and Fyffe 2016).

Clearly, shots do not necessarily occur immediately after a cross. Therefore, we introduce a tuning parameter k where a success (shot attempt) is defined as having occurred within the next k events. If the team maintains possession after the ball exits the potential crossing zone and a shot attempt occurs within the next k events, then $Y = 1$, otherwise $Y = 0$. In this application, we set $k = 5$. The idea to let the play "unfold" was used by Schuckers and

Curro (2013) in the context of player evaluation in hockey. Using the above definition for Y, we observed 274 shots arising from the $N = 2225$ crossing opportunities.

3 A Model for the Crossing Decision

We first consider how T (i.e. the variable denoting the decision to cross) depends on situational variables as expressed by Z (see Table 1). For this, we consider a logistic regression model based on the $N = 2225$ crossing opportunities where $\mathrm{T} \sim \text{Bernoulli}(p_{\mathrm{T}})$ and

$$\text{logit}(p_{\mathrm{T}}) = \lambda_0 + \lambda Z \ . \tag{1}$$

Parameter estimates and standard errors for the significant terms corresponding to model (1) are given in Table 2. To get a sense of the relative importance of the terms, the third column in Table 2 provides the parameter estimate multiplied by the mean value of its corresponding covariate. A notable observation is that given a crossing opportunity, crossing the ball is less frequent than not crossing the ball. For example, when the mean values of the covariates are substituted into the fitted equation corresponding to (1), the probability of a cross is $\text{Prob}(\mathrm{T} = 1) = 0.130$. We also note that all of the parameters in Table 2 are highly significant except for z_1 (p-value = 0.040) and z_9 (p-value = 0.051).

The coefficients in Table 2 also correspond to our soccer intuition. For example, we see that an increase in the ratio of offensive players in the box to defensive players in the box leads to an increased probability of crossing (i.e. positive coefficient of z_6). The most impactful covariate (column 3 of Table 2) is z_5 which is the distance between the sender and the endline. As the player runs towards to the end of the field, he runs out of options and therefore his crossing probability increases.

Coefficient	Est (Std Err)	Est*Mean
Intercept	-2.131 (0.260)	
z_1	-0.103 (0.050)	-0.0003
z_2	0.309 (0.037)	0.909
z_3	1.689 (0.232)	1.040
z_4	0.030 (0.011)	0.329
z_5	-0.125 (0.008)	-2.552
z_6	1.885 (0.227)	0.516
z_7	0.753 (0.169)	0.210
z_8	0.537 (0.157)	0.240
z_9	0.262 (0.134)	0.065

Table 2: Estimates and standard errors for the parameters corresponding to model (1). The third column provides the estimate multiplied by the mean value of its corresponding covariate.

4 The Intended Target Model

In order for a cross to occur, the offensive player in position to cross the ball (sender) must choose a target (receiver) amongst the offensive players in a potential crossing position. This section models the sender's success probability in crossing the ball to one of the possible receivers. Let $i = 1, \ldots, M = 692$ index the subset of crossing opportunities in the dataset where a cross was attempted. Therefore, this section provides an analysis conditional on an attempted cross.

The ith cross attempt is associated with a sender s_i, a set V_i of K_i possible receivers, and a random variable R_i indicating the successful recipient of the cross. The eligible receivers are restricted to the offensive players who are in the attacking half but not in the same potential crossing zone as the sender. For this analysis, the key situational variables are presented in Table 3. Some of these variables are common to those presented in Table 1, and some are additional variables. We use the superscripts (s) and (r) to denote variables that are relevant to the sender and the receiver, respectively.

Variable	Definition of Variable
$z_2^{(s)}$	- distance between the sender and nearest defender
$z_5^{(s)}$	- distance between the sender and the endline
$z_5^{(r)}$	- distance between the receiver and the endline
$z_{10}^{(r)}$	- distance between the receiver and the sideline
$z_{11}^{(r)}$	- speed of the receiver
z_{12}	- crossing angle between the sender and the receiver
z_{13}	- area of convex hull formed by potential receivers

Table 3: The key situational variables that are relevant to crossing success as modeled in Section 4. All distances are measured in metres, speed is measured in metres/second, angles are measured in degrees, and areas are measured in squared metres.

The random variable representing the receiver of the ith attempted cross takes values in the set $V_i \cup \{0\}$ where $R_i = 0$ indicates that the cross was unsuccessful. Therefore, there are $K_i + 1$ possible outcomes with respect to a given crossing attempt. Let $\mathbf{Z}_i$ denote the spatio-temporal features associated with the ith attempted cross as given in Table 3, where $\mathbf{z}_{ij}$ is the observed vector associated with potential receiver j during crossing opportunity i.

We again use the logistic regression framework where the probability that player j is the successful receiver of an attempted cross is given by

$$p\left(R_i = j|\mathbf{Z}_i\right) = \frac{\exp\left\{\beta^\top \mathbf{z}_{ij}\right\}}{1 + \sum_{j=1}^{K_i} \exp\left\{\beta^\top \mathbf{z}_{ij}\right\}}$$

and the probability of an unsuccessful cross is

$$p\left(R_i = 0|\mathbf{Z}_i\right) = \frac{1}{1 + \sum_{j=1}^{K_i} \exp\left\{\beta^\top \mathbf{z}_{ij}\right\}}.$$

The likelihood function over all crossing attempts is therefore

$$L(\beta|\mathbf{Z}_1, \ldots, \mathbf{Z}_M) = \prod_{i=1}^{M} \prod_{j=0}^{K_i} p(R_i = j|\mathbf{Z}_i)^{\mathbb{I}(r_i=j,)}. \tag{2}$$

It is worth noting this model's relation to other parametric models. Unlike multinomial logit models, the parameters are not indexed by the possible outcomes - the potential re-

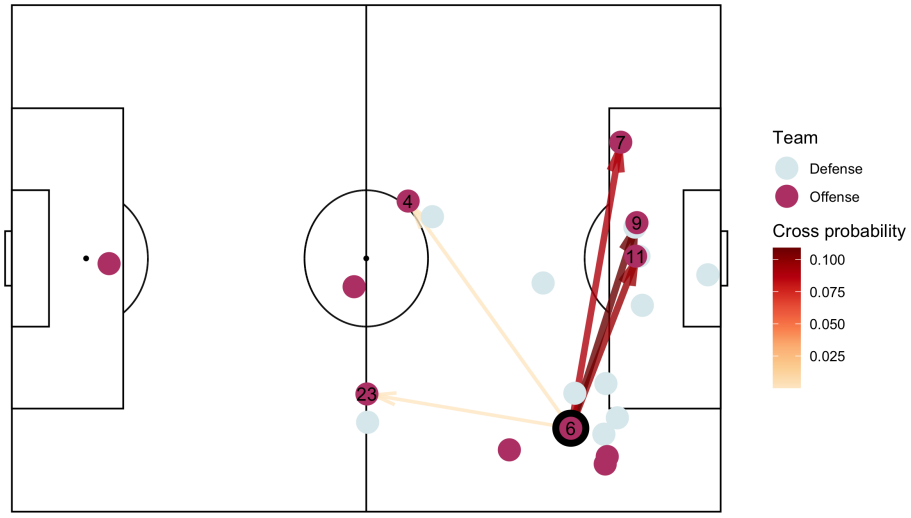
ceivers. In multinomial models, there are a fixed number of values the random variable can take. But, in our case, the number of potential receivers varies by crossing opportunity. Unlike the conditional logit model, equation (2) contains features specific to the sender that do not vary between potential receivers.

We fit the intended target model given by equation (2) using the situational variables in Table 3. Table 4 presents the maximum likelihood estimates which provide various soccer insights. We observe that as the distance between the sender and the nearest defender grows, the probability of cross completion increases. Senders located farther from the goal complete crosses at a higher rate. Receivers closer to the goal and farther from the sideline are more likely to successfully receive a cross. Wider crossing angles are associated with a higher probability of a successful reception. Faster moving receivers are more likely to receive a cross. And, more compact spatial configurations of potential receivers are associated with higher completion rates.

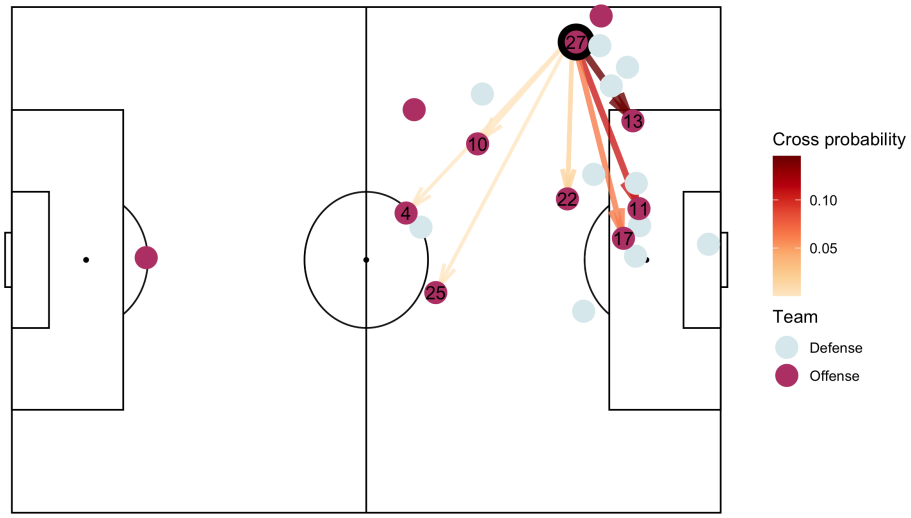
Feature	Estimate	Standard Error	P-value
$z_2^{(s)}$	0.1138	0.0410	0.0055
$z_5^{(s)}$	0.0416	0.0121	0.0006
$z_5^{(r)}$	-0.1705	0.0161	0.0000
$z_{10}^{(r)}$	-0.0176	0.0099	0.0760
$z_{11}^{(r)}$	0.0749	0.0481	0.1198
z_{12}	0.0014	0.0010	0.1540
z_{13}	-0.0012	0.0004	0.0016

Table 4: Estimates of the parameters from the intended target model. The estimates provide associations between spatio-temporal features and the successful completion of an attempted cross.

In addition to modeling the crossing process conditionally, our model provides insight as to how teams execute their offense. This can help teams recognize favorable spatial configurations for crossing. Figure 2 illustrates the success probabilities for crosses. In Figure 2(a), the sender is more open and located farther from the nearest sideline. As the distance between the nearest defender and the sender becomes larger, the probability of a completed cross increases. Also, crosses attempted by senders further from the sideline are completed with higher probability.



(a) The graph depicts the probability of cross completion to each of the potential targets during an offensive attack. In this example player 7 receives the cross.



(b) In this example, the cross is incomplete. One can see that the concentration of defenders around the most likely receiver, player 13, is much higher than the concentration of defenders around player 7 in panel (a).

Figure 2: Panels (a) and (b) present output from the intended target model. These diagrams provide a way for teams to study the spatial configurations of players and the ball during crossing opportunities.

5 Causal Inference

We begin by making an analogy between the crossing problem and retrospective medical studies. In our problem, whether a player crosses the ball (T = 1) or does not cross the ball (T = 0) is analogous to a patient receiving one of two treatments. And our response variable $Y = 1$ (a resulting shot) or $Y = 0$ (not a resulting shot) corresponds to patient outcomes where a patient may experience improved health or not. Finally, we note that a complication in both scenarios is that the treatment T and the response Y may both depend on auxiliary confounding variables Z. In the crossing problem, the variables Z are provided by the spatio-temporal tracking data.

In general, randomized experiments provide the means for investigating the cause and effect relationship of a treatment on a response. However, in the crossing problem, an experimenter can not assign (i.e. demand) that a player cross or not cross the ball during a particular crossing opportunity. For this reason, we use causal inference techniques (Pearl 2009) in this retrospective framework to investigate the cause-effect relationship of crossing on shots.

In this section we review the basic concepts used in the construction of interventional probability distributions (e.g. propensity scores). Then we review how matching estimators can be used to approximate an experiment with a randomized treatment. Finally, we compute causal effects of crossing on attempted shots and quantify the uncertainty associated with the estimates.

We expand on the setup via the directed acyclic graph presented in Figure 3 where arrows denote causal relationships. Again, the structure is analogous to a retrospective medical study in which a treatment is assigned to a patient. Whether the patient receives the treatment (T = 1) or not (T = 0) depends on the sets of covariates Z_{T} and $Z_{\mathrm{T}Y}$. The health status of the patient may be classified according to health ($Y = 1$) or sickness ($Y = 0$). The outcome variable Y depends on T as well as on the covariates $Z_{\mathrm{T}Y}$ and Z_Y. The variables Z_Y cause Y, but not T; the variables Z_{T} cause T, but not Y; and the variables $Z_{Y\mathrm{T}}$ cause both Y and T. Causal inference requires a method to address the confounding related to the common causes $Z_{Y\mathrm{T}}$.

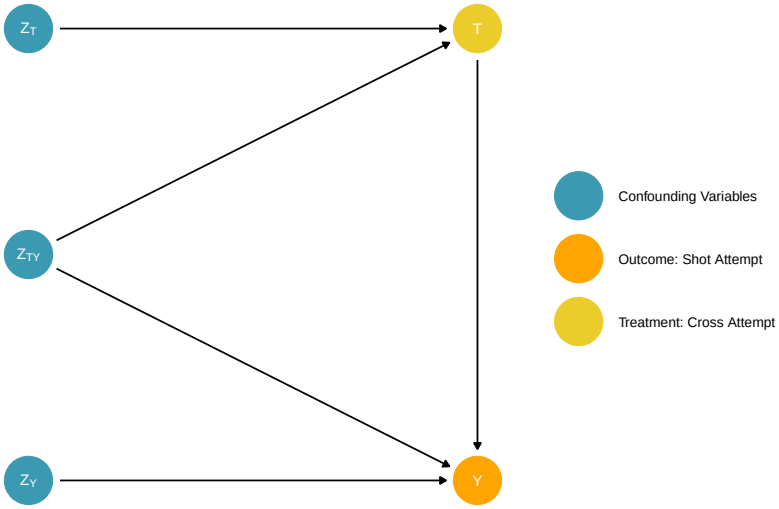


Figure 3: The directed acyclic graph describes the crossing problem. The variables Z_{T} are causes of T, but not Y. The variables Z_{TY} are common causes for T and Y. And, the variables Z_Y are causes for Y, but not T.

5.1 Propensity Score Matching

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 is the probability that the subject receives the treatment. In the crossing problem, $p_{\mathrm{T}} = \mathrm{Prob}(\mathrm{T} = 1)$ is the estimated probability that the sender will cross the ball, and is obtained via the logistic regression model (1). Therefore, p_{T} serves as the relevant propensity score in the crossing application.

In our problem, we have a dataset involving $N = 2225$ crossing opportunities resulting in $M_1 = 692$ cases where the ball is crossed (the treatment) and $M_2 = 1533$ cases where the ball is not crossed (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 propensity score. Then the resulting two groups (M_1 treatments and M_1 controls) will be similar in characteristics, and that differences between the two groups can be attributed to the treatment.

There are many ways that the matching of propensity scores can be carried out (Austin 2011, Imbens 2004), where caution ought to be exercised in the process (King and Nielsen 2019). In our application, we begin with the M_1 crossing cases and use a nearest neighbor method for selecting the matched non-crossing cases. 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 resultant inferences.

5.2 Results

Following the implementation of the matching procedure, Figure 4 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 a player decides to cross the ball.

The inferential component of our investigation begins with the simple two-sample test of proportions between the two groups based on the response Y (resultant shot) as described in Section 2.3. The quantity of interest is the average treatment effect $\text{ATE} = \bar{Y}_1 - \bar{Y}_2$ where $\bar{Y}_1$ is the mean number of resultant shots when the ball is crossed and $\bar{Y}_2$ is the mean number of resultant shots when the ball is not crossed. We obtain $\text{ATE} = 0.050$ with standard error 0.020. The result is significant and suggests that crossing is beneficial in the sense that a cross will lead to a shot 5% more often than when not crossing.

In Figure 5, 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 to crossing) the probability of a resultant shot increases for both groups. However, and most importantly,

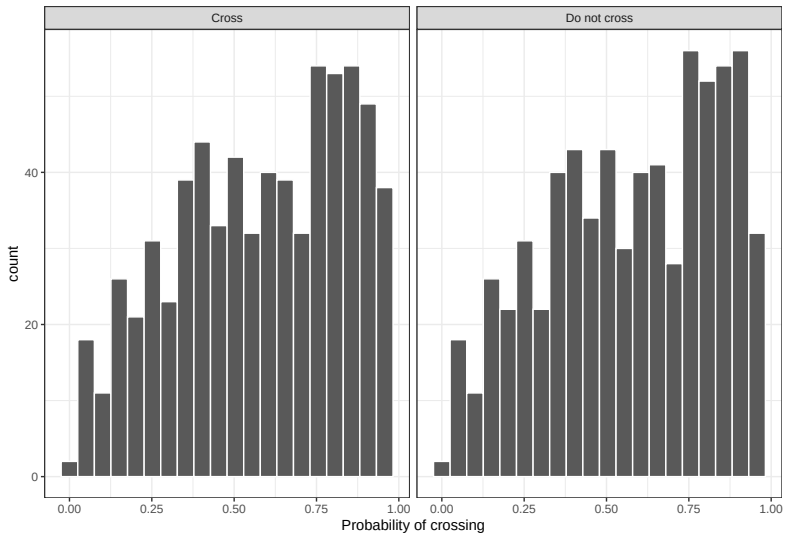


Figure 4: After matching, histograms of the two groups (treatment and control) are depicted where the horizontal variable is the propensity score.

we observe that the shot probability of the treatment group increases relative to the shot probability of the control group. In practice, what this means is that players are making correct tactical decisions. When players are more likely to cross (higher propensity scores), they will have better offensive results (higher shot probabilities) than if they did not cross.

Therefore, the takeaway message is that crossing (when done under the right circumstances) is a good thing to do. And players do cross at the right times.

6 Discussion

With access to player tracking data from the Chinese Super League, we have been able to carry out various detailed investigations of crossing in soccer. In particular, we have developed a logistic regression model that characterizes spatio-temporal conditions for crossing and an intended target model which explores crossing success.

However, the most important contribution of the paper concerns a reconciliation of the results of Vecer (2014) and Sarkar (2018) who essentially state that crossing the ball in soccer has a negative impact on scoring. The results of Vecer (2014) and Sarkar (2018) are

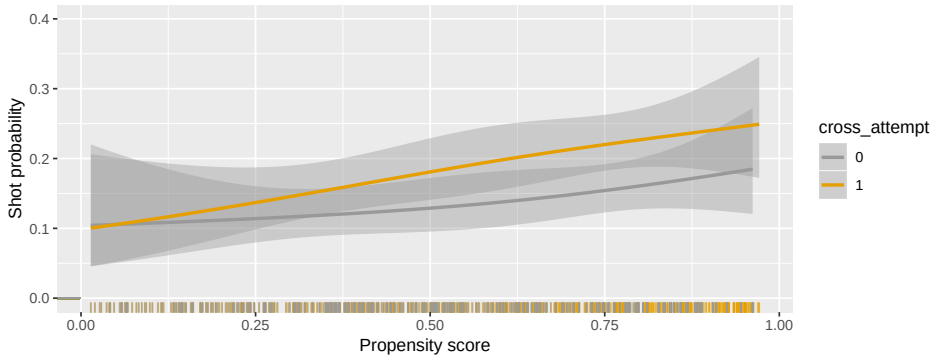


Figure 5: After matching, smoothed plots of the shot variable Y for both groups with respect to the propensity score.

puzzling since the strategy of crossing the ball in soccer has a longstanding reputation as an effective tactic. Using causal inference techniques, the message from this paper is that crossing remains a valuable tactic. Players have an intuition as to when to cross, and on average, when they cross, it is a beneficial time to cross. The beneficial crossing occasions are determined by the covariate patterns in Section 3. For example, players ought to cross when they are nearing the endline, when they have space and when there is a high ratio of offensive to defensive players in the box. These nuanced insights that we have provided are not necessarily a response to the work of Vecer (2014) and Sarkar (2018), but rather an evolution of looking at crosses using player tracking data.

A common and legitimate criticism of the use of causal matching methods based on propensity scores is that it is generally difficult to identify and measure confounding variables. However, we believe that sport offers one of the rare applications where the confounding variable problem is manageable. Sport has well defined rules with finite time spans and clear objectives. People understand sport well (say, compared to genetics), and consequently there is hope for causal analyses in sport, especially with the increasing availability of detailed player tracking data. However, even in this hopeful environment for the investigation

of causal inference, it would have been possible to introduce additional covariates. For example, we did not account for the quality of teams, the movement of players, the position of the keeper nor did we consider the EPV (expected point value) status when there is an opportunity to cross. It is our hope that this manuscript will inspire future causal studies in sport.

There are a number of potential future directions for the research presented here. One would involve modifying the binary response $Y =$ shots to the increasingly popular expected goals statistic XG (Bundesliga 2019) which is a continuous variable. Another avenue involves extensions of the intended target model in Section 4. For example, it would be possible to include the space covariate z_3 from Table 1 to provide more information about the nature of the convex hull (covariate z_{13} in Table 3).

We wish to emphasize that our data analyses were based on a single season's worth of data from Shandong Luneng FC. Although the inferences gained via our analyses are specific to Shandong Luneng, it is plausible that some of the broad insights may hold generally to high level soccer competitions. It would be interesting to see if our results hold for other teams and leagues. Our suspicion is that there is commonality across soccer competitions, and that the benefit of crossing in the specific situations mentioned in this paper would translate to other high level teams and leagues.

7 References

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.

Bransen, L., Van Haaren, J. and van de Velden, M. (2019). Measuring soccer players' contributions to chance creation by valuing their passes. *Journal of Quantitative Analysis in Sports*, 15(2), 97-116.

Bundesliga (2019). XG stats explained: The science behind Sportec Solutions' expected goals model. Accessed on July 10, 2020 at https://www.bundesliga.com/en/bundesliga/news/expected-goals-xg-model-what-is-it-and-why-is-it-useful-sportec-solutions-3177

Fernandez, J. and Bornn, L. (2018). Wide open spaces: A statistical technique for measuring space creation in professional soccer. In *12th Sloan Sports Analytics Conference*, Accessed on May 14, 2020 at http://www.sloansportsconference.com/wp-content/uploads/2018/03/1003.pdf

Fernandez, J., Bornn, L. and Cervone, D. (2019). Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer. In *13th Sloan Sports Analytics Conference*, Accessed on May 14, 2020 at http://www.sloansportsconference.com/wp-content/uploads/2019/02/Decomposing-the-Immeasurable-Sport.pdf

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.

King, G. and Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, 27, 435-454.

Liu, H., Gomez, M.A., Lago-Penas, C. and Sampaio, J. (2015). Match statistics related to winning in the group stage of 2014 Brazil FIFA World Cup. *Journal of Sports Sciences*, 33(12), 1205-1213.

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

Oberstone, J. (2009). Differentiating the top English Premier League football clubs from the rest of the pack: Identifying the keys to success. *Journal of Quantitative Analysis in Sports*, 5(3), Article 10.

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

Power, P., Ruiz, H., Wei, X. and Lucey, P. (2017). Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, 1605-1613.

Sarkar, S. (2018). Paradox of crosses in association football (soccer) - a game theoretic explanation. *Journal of Quantitative Analysis in Sports*, 14(1), 25-36.

Sarkar, S. and Chakraborty, S. (2018). Pitch actions that distinguish high scoring teams: Findings from five European football leagues in 20115-16. *Journal of Sports Analytics*, 4, 1-14.

Schuckers, M. and Curro, J. (2013). Total hockey rating (THoR): A comprehensive statistical rating of National Hockey League forwards and defensemen based upon all on-ice events. Proceeding of the 2013 MIT Sloan Sports Analytics Conference, Accessed on February 26, 2019 at http://statsportsconsulting.com/thor/

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.

Szczepanski, L. and McHale, I. (2016). Beyond completion rate: Evaluating the passing ability of footballers. *Journal of the Royal Statistical Society*, 179, Part 2, 513-533.

Toumi, A. and Lopez, M. (2019). From grapes and prunes to apples and apples: Using matched methods to estimate optimal zone entry decision-making in the National Hockey League. Accessed on May 14, 2020 at https://rpubs.com/atoumi/zone-entries-nhl

Vecer, J. (2014). Crossing in soccer has a strong negative impact on scoring: Evidence from the English Premier League, the German Bundesliga and the World Cup 2014. Accessed on February 5, 2019 at SSRN: https://ssrn.com/abstract=2225728

Vollman, R. with T. Awad and I. Fyffe (2016). *Stat Shot: The Ultimate Guide to Hockey Analytics*, ECW Press: Toronto.

Wilson, J. (2013). *Inverting the Pyramid: The History of Soccer Tactics*, Nation Books: New York.

Yam, D.R. and Lopez, M.J. (2020). What was lost? A causal estimate of fourth down behavior in the National Football League. *Journal of Quantitative Analysis in Sports*, To appear.