MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 11 - 12, 2016 BOSTON CONVENTION AND EXHIBITION CENTER

The Pressing Game: Optimal Defensive Disruption in Soccer

Iavor Bojinov* and Luke Bornn*†
*Harvard University, Cambridge, MA, 02138,
†Simon Fraser University, Burnaby, BC, V5A 1S6
Email: bojinov@fas.harvard.edu, bornn@stat.harvard.edu

Abstract

Soccer, the most watched sport in the world, is a dynamic game where a team's success relies on both team strategy and individual player contributions. Passing is a cardinal soccer skill and a key factor in strategy development; it helps the team to keep the ball in its possession, move it across the field, and outmaneuver the opposing team in order to score a goal. From a defensive perspective, however, it is just as important to stop passes from happening, thereby disrupting the opposing team's flow of play. Our main contribution utilizes this fundamental observation to define and learn a spatial map of each team's defensive weaknesses and strengths. Moreover, as a byproduct of this approach we also obtain a team specific offensive control surface, which describes a team's ability to retain possession in different regions of the field. Our results can be used to distinguish between different defensive strategies, such as pressing high up the field or sitting back, as well as specific player contributions and the impact of a manager.

1 Introduction

The Barclays English Premier League (EPL) has more fans, is watched by more people, and generates more revenue than every other major soccer league [2, 7]. The twenty teams that participate in the EPL each year compete for the league title, qualifying for the UEFA Champions/Europe league or simply surviving relegation and being one of the seventeen teams that remain. At the end of each season pundits dissect the teams and write about their quality and style of play. Amongst other things they discuss which teams had the best offense, which teams had the best defense, who were the stars of the season, and who was the manager of the season. They often employ basic summary statistics for these analyses; for example, in the 2014/2015 season Manchester City scored the most goals, Lukasz Fabianski made the most saves, and Burnley's players ran the most miles. These, although interesting, fail to capture one of the most important aspects of the game, that is, that soccer is fundamentally a spatial game.

In this paper we remedy this shortcoming by first providing summary statistics that quantify a team's ability to retain possession when in control of the ball and to disrupt the opposing team when not. Subsequently, by expanding our model, we provide a team-specific cartography that maps out the strengths and weaknesses of its offense and defense. Coaches can use these maps to understand and correct their team's weaknesses or to exploit an opponent's vulnerabilities. This work is applied to the Prozone EPL season 2012/13, 2013/14 and 2014/15 events data.

The paper is organized as follows. In Section 2 we derive the season average disruption surface and obtain team specific disruption and control coefficients which can be used to quantify a team's performance. In Section 3 we expand the model to obtain a cartography of each team's disruption and control surfaces. In Section 4 we show how these surfaces can be used to understand the tactics employed by managers across different teams. In the final section we provide some concluding remarks and a possible extension to our model.

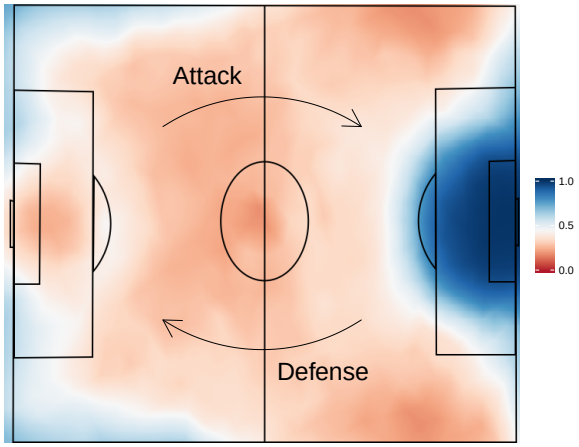

Figure 1: Average disruption surface for the English Premier League 2014/15. Blue indicates a high probability of a disruption whilst red indicates a low probability. The data is flipped such that the team with the ball is shooting from left to right and the defending team is shooting from right to left.

2 Season Average Disruption Surface

To obtain a season long average disruption surface thats accounts for the spatial variations in the data we use a generalized linear spatial regression (GLSR) model [1, 3]. A GLSR is an extension to the classical generalized linear model, where the random component consists of responses $Y_1(\mathbf{s}_1), \ldots, Y_n(\mathbf{s}_n)$ which are allowed to depend on the spatial location $\mathbf{s}$ where they are observed. The outcome of interest is a binary variable indicating if a disruption event ($Y(\mathbf{s}) = 1$) or a continual control event ($Y(\mathbf{s}) = 0$) occurred at location $\mathbf{s}$. A disruption of the attacking team (the team in possession) is defined as an action taken by the defensive team that leads to an interruption of the flow of play. Examples include interceptions, tackles, clearances and blocks. Completed passes and player touches are used as a proxy for continual control of the ball. With the above notation we define the conditional probability of a disruption at location $\mathbf{s}$, given the attacking and defending team, to be

$$P(Y_i(\mathbf{s}_i) = 1 | \mathbf{X}_i^{\text{Atk}}, \mathbf{X}_i^{\text{Def}}, \mathbf{s}_i) = \sigma\left(\alpha - \mathbf{X}_i^{\text{Atk}}\boldsymbol{\beta}^{\text{Atk}} + \mathbf{X}_i^{\text{Def}}\boldsymbol{\beta}^{\text{Def}} + Z(\mathbf{s}_i)\right), \tag{1}$$

where $\sigma(x) = e^x/(1+e^x)$. The vector $\mathbf{X}_i^{\text{Atk}} = (X_{i1}^{\text{Atk}}, \ldots, X_{i20}^{\text{Atk}})$ is one-hot encoded, and indicates which team is currently in possession of the ball, $X_{ij}^{\text{Atk}} = 1$ when the attacking team in event i is j and 0 otherwise, the vector $\mathbf{X}_i^{\text{Def}} = (X_{i1}^{\text{Def}}, \ldots, X_{i20}^{\text{Def}})$ indicates which team is defending and is defined in a similar manner. The intercept of the model is α, and the team specific offensive and defensive coefficients that describe the team's ability to control the ball, when in possession, and to disrupt the play when not are $\boldsymbol{\beta}^{\text{Atk}}$ and $\boldsymbol{\beta}^{\text{Def}}$ respectively. The final component of equation (1) is $Z(\mathbf{s})$, a two dimensional Gaussian process (GP) with Matérn covariance function, which is used to model the spatial aspect of the data; we term the posterior mean of this spatial process as the *disruption surface* (DS). The Matérn covariance function has two free hyper-parameters that have natural interpretations. The *smoothness* parameter controls the differentiability of the covariance function, and the *range* parameter controls at what distance the covariance between two points becomes small [9].

We fit the model in Equation 1, with vague Gaussian hyper-priors for the parameters of the covariance function and the regression coefficients, using INLA [8, 10]. The season long disruption surface, displayed in Figure 1, is in line with our intuition, and has picked up key futures such as the area inside the penalty box is where the defending team is most likely to disrupt their opponent. This can be used to better understand how the game is played. From a coaching perspective however, it is more important to study how each team deviates from this

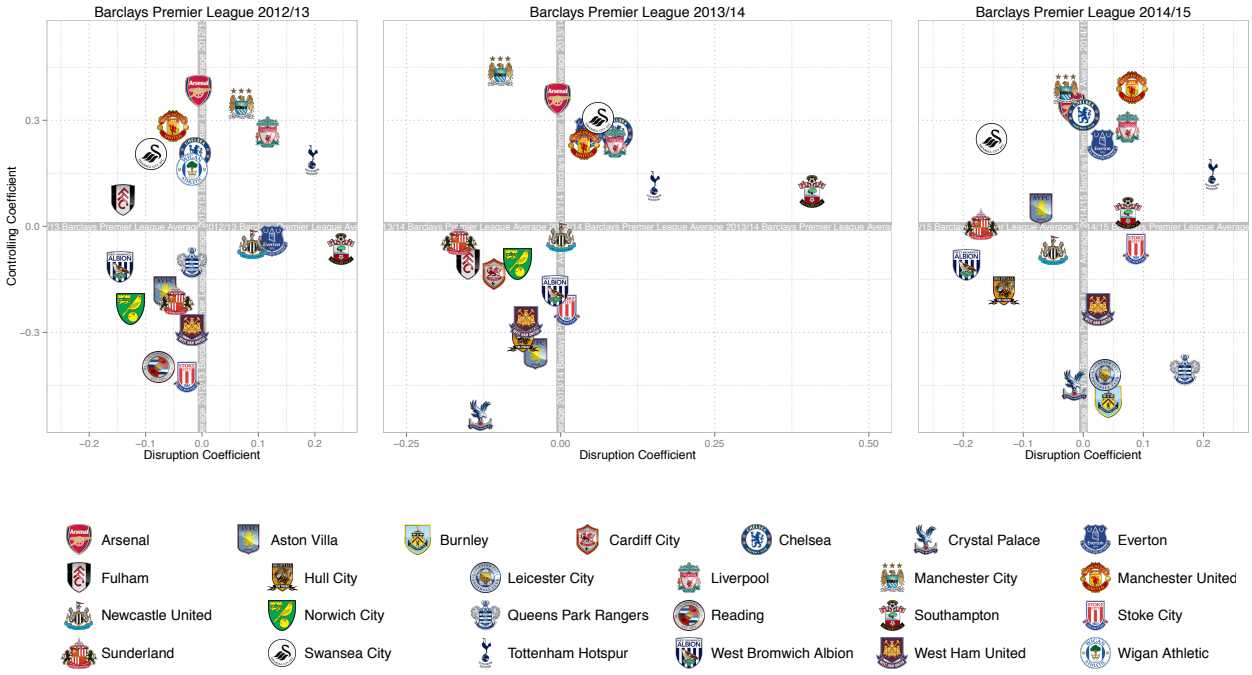

Figure 2: Controlling and disruption coefficients for the last three full English Premier League seasons. We see that Arsenal consistently ranks high in terms of offensive control of the ball, while Fulham struggled to induce turnovers before being relegated at the end of the 2013/2014 season.

season average; this is discussed in the next section.

Intuitively, the defensive disruption and offensive control coefficients shift the DS up or down depending on their sign. From Figure 2 we see that on average teams that finish in the top half of the league table have the highest controlling coefficient[1] and average disruption coefficient. To provide a better understanding of these results we compared the coefficients to the number of shots taken for and against each team; the results are shown in Figure 3. Overall there is a clear positive relationship between shots taken by a team and the value of both the control and disruption coefficient. This agrees with the general perception that if a team is able to retain possession for longer then they will also take more shots [11, Chapter 5]. This trend is reversed when we examine shots taken against a team; if a team is better able to disrupt the opposing teams natural flow of play then they also tend to have less shots taken against them.

The disruption and control coefficients can be viewed as a team metric that quantifies team attributes in a unique way. The disruption coefficient provides information regarding how pressing[2] or passive a team is, whilst the control coefficients encapsulates a team's ability to retain control once in possession. Both coefficients capture a mixture of regular team players contributions and manager tactics. The former is discussed further in the next section and the latter is the focus of Section 4.

3 Cartography of Defensive Disruption and Offensives Control Surfaces

In this section, we map out team specific defensive disruption and offensive control surfaces. To do this, we expand the model given in Equation (1) to include multiple GPs in two ways.

[1] 8 of the top half teams in 2013/14 and 2014/15 had above average control coefficients as did 7 in 2012/13.
[2] Pressing, in soccer, is the attempt to put pressure on the opposition when they have the ball.

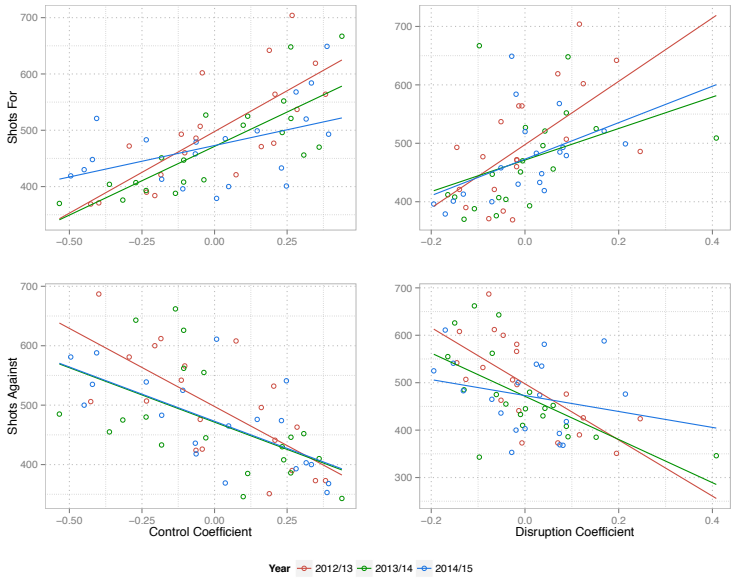

Figure 3: A comparison between shots for and against a team and the control and disruption coefficient across the last three full seasons. The lines are obtained from fitting a linear regression with least squares.

Firstly, we define the individual team disruption surfaces model as

$$P(Y_i(\mathbf{s}_i) = 1|\mathbf{X}_i^{\text{Atk}}, \mathbf{X}_i^{\text{Def}}, \mathbf{s}_i) = \sigma\left(\alpha + \sum_{j \in \text{Teams}} V_j^D(\mathbf{s}_i) X_{ij}^{\text{Def}} + Z(\mathbf{s}_i)\right), \tag{2}$$

where $\alpha, X_{ij}^{\text{Def}}$ and $Z(\mathbf{s})$ were defined in the previous section, and $V_j^D(\mathbf{s})$ is a two dimensional GP with Matérn covariance function. The Disruption Surface (DS) - a spatial surface referenced over the pitch - for team j is defined as the posterior mean of $V_j^D(\mathbf{s})$ and indicates the team's propensity to induce turnovers when the opposing team has possession of the ball. Negative values indicate a team presses less than the average, while positive values indicate that a team presses more.

Secondly, we define the offensive control surface model as

$$P(Y_i(\mathbf{s}_i) = 1|\mathbf{X}_i^{\text{Atk}}, \mathbf{X}_i^{\text{Def}}, \mathbf{s}_i) = \sigma\left(\alpha - \sum_{j \in \text{Teams}} V_j^A(\mathbf{s}_i) X_{ij}^{\text{Atk}} + Z(\mathbf{s}_i)\right), \tag{3}$$

where $V_j^A(\mathbf{s})$ is a two dimensional GP with Matérn covariance function. The Control Surface (CS) - a spatial surface referenced over the pitch - for team j is defined as the posterior mean of $V_j^A(\mathbf{s})$ and indicates the team's ability to retain possession when in control of the ball. Similar to the the DS, negative values indicate a team is less able to retain possession than the average, while positive values indicate the opposite.

The most computationally intensive part of fitting GLSR models is obtaining the posterior distribution of the hyper-parameters of the covariance function. To improve efficiency we used an approximate empirical Bayes approach, whereby we set the two hyper-parameters for each of the GPs to the mode of the posterior distribution obtained from fitting $Z(\mathbf{s})$ in Equation 1. This approximation hinges on the expectation that the smoothness and the correlation range are similar across teams. Since the posterior distribution of the hyper-parameters of $Z(\mathbf{s})$ obtained in Section 2 is tightly centered around the mode, which is roughly the same across different seasons, we can be confident that this is indeed the case.

The results for every team across all three seasons are insightful; however, to keep our exposition short we focus on the 2014/15 season. Figure 4 shows the control and disruption surfaces of three teams: Burnley, the

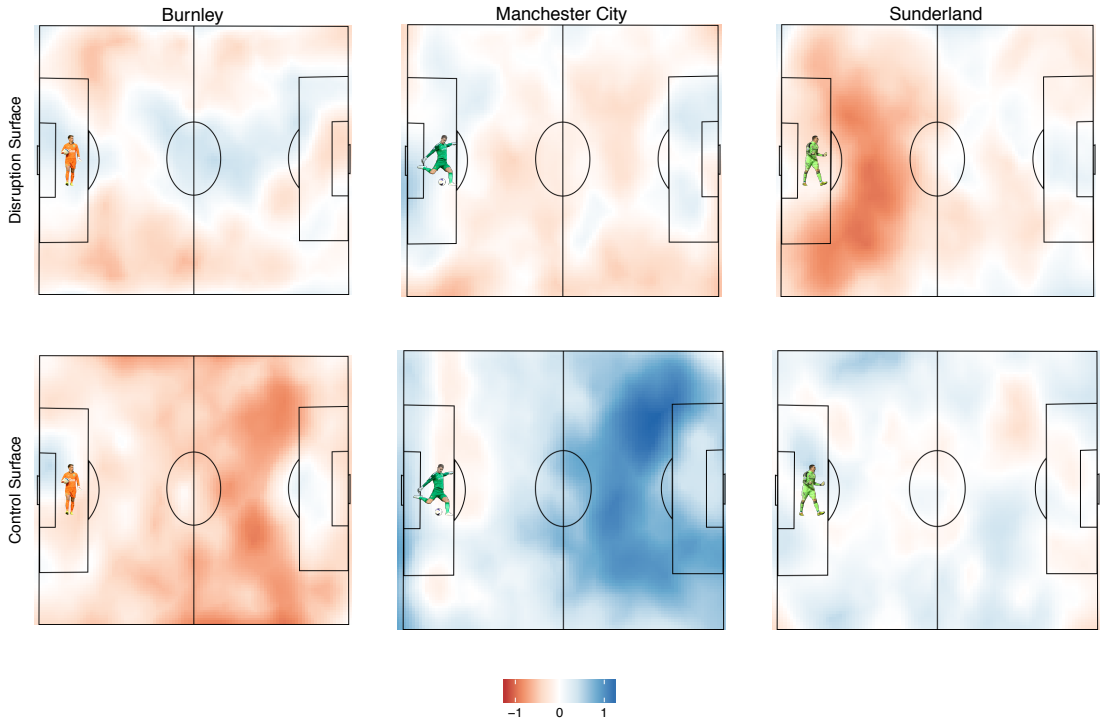

Figure 4: Average control and disruption surface for Burnley, Manchester City and Sunderland during the 2014/15 English Premier League. During the analysis the data is flipped so that the goalkeepers are protecting their goal. For both the disruption and control surfaces red indicate below average, white indicates average and blue indicated above average values.

team that scored the least number of goals; Manchester City, the team that took the most shots and scored the most goals; and Sunderland, the team that conceded the most shots.

3.1 Burnley's lack of control

Burnley's inability to score goals was without a doubt one of the main reasons why they finished 19^{th} in the 2014/15 seasons and were relegated. They only managed to net a meager 28 goals - 20 fewer than the league average - unfortunately, this number alone provides little insight into what went wrong. Burnley had the lowest control coefficient that season, by examining their control surface we can take step towards understanding this number and through it their shortcomings. Their control surface is almost entirely below the league average and in particular there is a major dip outside the opposing teams' penalty box. This means that first team regulars like midfielder David Jones and forward Danny Ings were about half as likely to retain possession in a threatening position as the league average. The disruption surface tells a slightly different story: the central midfield was able to cause more disruptions than the average, and the left wing was slightly better than the right wing, however in the end this was undermined by their inability to retain possession.

3.2 Manchester City's offensive domination

Scoring 83 goals from 649 shots, Manchester City was the 2014/15 season's most threatening team. When examining their control surface we see a stark contrast with that of Burnley. In particular their ability to retain possession in the final third was unmatched by any team, including the winners of the league Chelsea. The

team was slightly better at controlling the left wing, the side favored by their top scorer Sergio Agüero. From the disruption surface we see that tactically Manchester City employed passive defensive strategy, not heavily pressing the opposing team.

3.3 Sunderland's lucky escape

Premier league teams collectively took just under 700 shots against Sunderland - approximately 200 above the league average. Their disruption surface is almost entirely below average, particularly in their own zone, meaning that their defense was causing little trouble to the opposition. Although left back Patrick van Aanholt was more likely to cause a disruption compared to the alternating right backs Billy Jones and Anthony Réveillère, he was still underperforming. Once in possession Sunderland were able to retain it as well as any other team, but this was still not enough to stop the onslaught of shots and make up for their poor defense in their own half.

3.4 Using the cartography to improve teams

Coaching staff spend hours analyzing their opponents' previous games to understand their strengths and weakness. Through the use of disruption and control surfaces we can improve this procedure. Control surfaces provide the topographies for the areas in which a team dominates possession and the areas in which they lose it. Disruption surfaces map where a team presses and where they are more passive. This information can be used by coaches to expose the opposing team's vulnerabilities and to strengthen their own weaknesses. For example, when playing against Manchester City, using a high pressing tactic will disrupt their dominating control surface. Stoke City used this tactic at the start of the 2014/15 season to register their first ever win at the City of Manchester (Etihad) Stadium [6].

We restricted our focus to an entire season; however, it is possible to fit the model using a subset of the data to answer more specific questions. Questions such as "how does playing Chelsea's Branislav Ivanović as central defender instead of his usual right back alter the team's DS?" can easily be answered by fitting the model only on games that he played in. The use of our approximate empirical Bayes approach ensures that the resulting surfaces mimic the smoothness and correlation structure obtained from the full season's data.

4 The Pochettino Effect

Over the past three seasons teams that have had the same manager (such as Arsenal, Liverpool and West Ham United) had a consistent value for their disruption and control coefficients. On the other hand, teams that have gone through several managers (such as Aston Villa, Manchester United, and Southampton) have seen dramatic shifts in their coefficients. The one exception is Stoke City, which under Mark Hughes' management, have transformed from a team that avoids being in possession to a team that presses and outplays their opponents [5, 11]. The transition is reflected in the way their controlling and disruption coefficients have changed over the seasons. These examples suggest that Figure 2 contains information about the manager as well as the players.

Mauricio Pochettino, the former manager of Southampton who moved to Tottenham Hotspur at the end of the 2013/14 season, is known for a very high-pressing attacking style of football [4]. This can be seen in Figure 2, where the team that he manages always has the highest disruption coefficient.

To better understand Pochettino's effect we focus our analysis on the last two full seasons, Figure 5 shows the disruption surface for Southampton and Tottenham Hotspur. It is clear that Southampton's 2013/14 and Tottenham Hotspurs 2014/15 disruption surfaces are well above the season average and substantially higher than the year that Pochettino was not managing them. Moreover, Pochettino managed to make these drastic changes without substantially altering the team [3].

Team disruption and control surfaces allow us to understand how a manager impacts a team's playing style. Often, a change in management is not instantly followed by a drastic change in first team players, therefore any major changes in the team's control and disruption surfaces are due to tactical changes made by the manager.

[3] Pochettino purchased 6 players before the start of the 2014/15 campaign, none of whom became first team regulars.

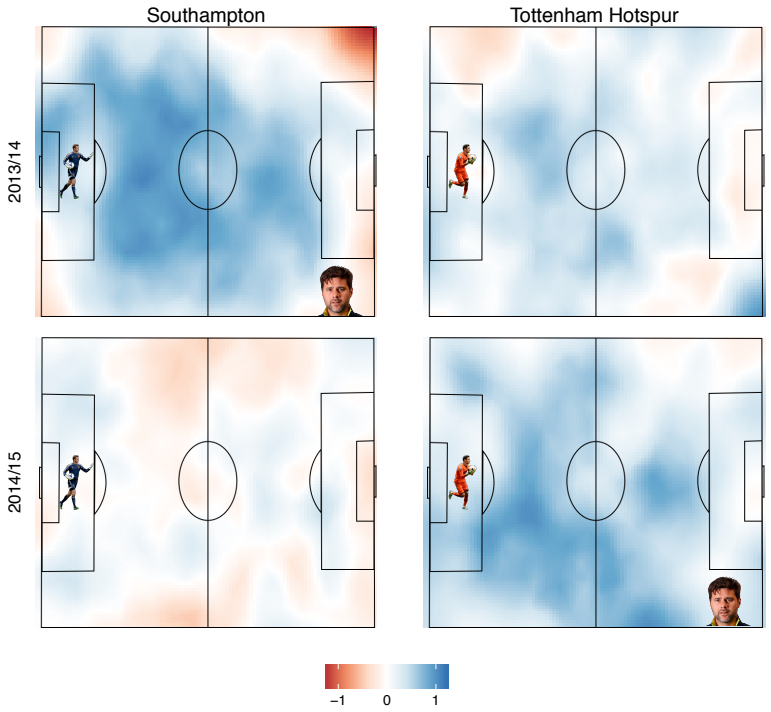

Figure 5: Average disruption surfaces for Southampton and Tottenham Hotspur during the 2013/14 and 2014/15 English Premier League. The picture of Mauricio Pochettino in the lower right corner indicates that he was managing the team. During the analysis the data is flipped so that the goalkeepers are protecting their goal. Red indicate below average, white indicates average and blue indicated above average values.

This approach allows us to follow managers across different seasons and leagues; leading to a better understanding of their preferred playing tactics.

5 Conclusions

In this work, we explore the disruptive ability of teams, learning the map of where they control the ball (when on offense) and disrupt their opponents (when on defense). With this new spatial understanding, we are able to quantify the strengths and weaknesses of a team's defense and offense in specific regions. This allows coaches to not only adjust defensive strategy to bolster weak regions, but also to build offensive strategies to exploit an opponent's spatial vulnerabilities. We showed how these can be used at both the player- and manager-specific levels. To the best of our knowledge, our quantification of a manager's cartographic offensive and defensive surfaces is the first of its kind and can be used to allow executives to select coaches that fit the team's desired style of play.

We believe that our work is a starting point for the development of models that are able to capture the spatial aspects of soccer and can lead to more informative team metrics. As a first step, our model can be expanded to include more covariates, such as whether the fixture was played at home or away, which will help gain further insights.

Acknowledgements

The computations in this paper were ran on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University.

References

[1] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. Crc Press, 2014.

[2] David Conn. Premier league top of the rich list with record income of £3.26bn. URL http://www.theguardian.com/football/2015/jun/04/premier-league-tv-income-la-liga-deloitte.

[3] Peter J Diggle, JA Tawn, and RA Moyeed. Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350, 1998.

[4] Sophie Figueira. The reign of Mauricio Pochettino: One year on. URL http://www.givemesport.com/416067-the-reign-of-mauricio-pochettino-one-year-on.

[5] Mike Keegan. Mark Hughes says he has taken Stoke City to 'next level' by changing style of football. URL http://www.dailymail.co.uk/sport/football/article-3076052/Mark-Hughes-says-taken-Stoke-City-level-changing-style-football.html.

[6] Mike Keegan. Stoke City registered their first Premier League victory in Manchester as Mame Biram Diouf's fine solo effort stunned the champions, August 2014. URL http://www.bbc.com/sport/0/football/28907276.

[7] Premier League. What we do - The world's most watched league. URL http://www.premierleague.com/content/premierleague/en-gb/about/the-worlds-most-watched-league/.

[8] Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

[9] Carl Edward Rasmussen. *Gaussian processes for machine learning*. Citeseer, 2006.

[10] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

[11] David Sally and Chris Anderson. *The numbers game: why everything you know about soccer is wrong*. Penguin, 2013.