A framework for the analytical and visual interpretation of complex spatiotemporal dynamics in soccer

A thesis submitted for the degree of
Doctor of Philosophy in Artificial Intelligence

Javier Fernández de la Rosa

Thesis directors:

Luke Bornn
Ricard Gavaldà

Polytechnic University of Catalonia
Department of Computer Science

November 2021

Abstract

Sports analytics is an emerging field focused on the application of advanced data analysis for assessing the performance of professional athletes and teams. In soccer, the integration of data analysis is in its initial steps, primarily due to the difficulty of making sense of soccer's complex spatiotemporal relationships and effectively translating findings to practitioners. Recently, the availability of spatiotemporal data has given rise to applying statistical approaches to address problems such as estimating passing and scoring probability, or the evaluation of players' mental pressure. However, most of these approaches focus on isolated aspects of the sport, while coaches tend to focus on the broader interplay of all 22 players on the pitch. To address the non-stop flow of questions that coaching staff deal with daily, we identify the need for a flexible analysis framework that allows us to answer these questions quickly, accurately, and in a visually-interpretable way while capturing the complex spatial and contextual factors that rule the game.

We propose developing such a comprehensive framework through the concept of the expected possession value (EPV). First introduced in basketball, EPV constitutes an instantaneous estimate of the expected points to be scored at the end of a possession. However, aside from a shared high-level goal, our focus on soccer necessitates a drastically different approach to account for the sport's nuances, such as looser notions of possession, the ability of passes to happen at any location, and space-time dependent turnover evaluation. Following this, we propose modeling EPV in soccer by addressing the question, "can we estimate the expectation of a team scoring or conceding the next goal at any time in the game?" From here, we address a series of derived interrogations, such as how should the EPV expression be structured so coaches can more easily interpret it? Can we produce calibrated and interpretable estimates for each of its components? Can we develop representative and soccer-specific features with the aid of coaches? Is it possible to learn complex features from raw level spatiotemporal data? Finally,

and most importantly, can we produce compelling practical applications?

These questions are successfully addressed in this thesis, where we present a series of contributions for both the machine learning and soccer analytics fields related to the modeling and practical interpretation of complex spatiotemporal dynamics. We propose a decomposed modeling approach where a series of foundational soccer components can be estimated separately and then merged to provide a single EPV estimation, providing flexibility to this integrated model. From a practical standpoint, we leverage several function approximation approaches to exploit complex relationships in spatiotemporal tracking data. An essential contribution of this work is the proposal of SoccerMap, a flexible deep learning architecture capable of producing accurate and visually-interpretable probability surfaces in a broad range of problems. Based on a large set of spatial and contextual features here developed, we model and provide accurate estimates for each of the components of the EPV components. The flexibility and interpretation capabilities of the proposed model allow us to produce a broad set of practical applications related to on-ball performance, off-ball performance and match analysis. Based on the proposed functional framework, future developments can easily incorporate a set of improvements for performing sophisticated analysis of unexplored soccer problems and translating this modeling approach to other team sports.

This thesis was developed under the support of the “plan de doctorados industriales del departamento de investigación y universidades de la Generalitat de Catalunya” and carried out entirely at Fútbol Club Barcelona, which promoted a close collaboration with professional coaches. As a result, a vast part of the ideas developed in this thesis is now part of the club’s daily player and team performance analysis pipeline.

Dedication

To Gaby, who always sees the best. My life partner and best friend.

To my parents, who are always there. The living example of unconditional love.

To Júlia, who teaches us to see life from a completely new angle.

Acknowledgements

I am extremely grateful to my supervisor Luke Bornn, who has taught me to explore ideas in a whole new way, and has guided me through the most exciting years of my professional career. Luke has been an invaluable mentor, and a role model.

I would like to thank my supevisor Ricard Gavaldá whose selflessness and proactive spirit was key to unblocking this work in many opportunities. I am also grateful to my tutor Marta Arias, who has been of great help in this thesis's delivery and validation process.

I would also like to thank Fútbol Club Barcelona and all the brilliant coaches and teammates, with whom I have learned and enjoyed soccer so much during all these years. In particular I would like to thank Xavier Reche, Antonio Gómez, Xavier Pavo, Javier Molina, Dídac Soler, Jordi Melero, Eduard Pons, Rafel Pol, Isaac Guerrero, Jordi Fernández and Marc Quintana, whose vision and experience were decisive for this work to be closer to the coaches' language.

I am very grateful to Daniel Medina who has always pushed to pursue excellence and trusted me to embark on the adventure that changed my life.

I would like to thank Raúl Peláez, whose trust, respect, and courage made this work possible and who opened the doors for me to collaborate with so many valuable people in all these years.

I am grateful to Ramon Canal, whose vision, support and ambition made it possible for me to carry out this work in a club with so much history.

I am also thankful to Angel Ric, with whom I enjoyed sharing discussions for hours in various stages of this thesis process, and who provided great

ideas and motivated curiosity and excellence.

I would like to thank the FC Barcelona sports science team, with whom we made the practical application of these ideas a reality, and enjoyed adding value to the sport we love so much and to the players we so admire.

I am thankful to Jesús Ruiz de la Torre, and the team of the "plan de doctorados industriales del departamento de investigación y universidades de la Generalitat de Catalunya", who always provided selfless and timely help, which was key to the development of this thesis as an industrial doctorate.

And finally, thank you to my family, all the friends, co-workers and so many other people who were always there to listen, contribute and support the ideas of this exciting work.

Contents

List of Figures

List of Tables

List of Abbreviations

2D	2-dimensional 56, 112, 115, 189
ADAM	adaptive moment estimation 46, 86, 117, 135
AI	artificial intelligence 19, 33
conv2d	2-dimensional convolutional filters 112
convnets	Convolutional neural networks 46, 112
ECE	expected calibration error 56, 118, 119, 136, 180
EPL	English Premier League 5, 11, 12, 96–98, 115, 133, 134, 152, 175, 176, 187
EPV	Expected Possession Value 3, 7, 21–25, 29–33, 42, 59–65, 67–70, 92, 95–97, 124, 125, 138, 142, 143, 146, 148, 153, 155, 172, 177–181, 183–189
EPVA	Expected Possession Value Added 8–10, 146, 151, 153–156, 169–171, 173, 189
GPU	graphical processing unit 117, 187
log-loss	logarithmic loss 6, 50, 117–119, 121, 127–129, 132
MDP	Markov Decision Process 54, 60, 61
MLB	Major League Baseball 18
MSE	mean squared error 50, 86, 129, 131, 132, 136
NBA	National Basketball Association 18

PPA	pass completion added 174, 175, 186
ReLu	Rectified Linear unit 52, 53, 112, 114, 117, 120, 130
SGD	stochastic gradient descent 45, 54
SHAP	Shapley Additive explanations 7, 29, 58, 141–143, 145, 181
xG	expected goals 5, 27, 29, 41, 62, 67, 89, 92, 93, 131, 132, 141, 169

Chapter 1

Introduction

Sports analytics is a fast-growing research field focused on the data-driven performance analysis of professional athletes and teams. In the last decade, data analysis has started to provide a competitive advantage to professional teams in a wide range of sports, particularly in basketball and baseball, where most teams in the National Basketball Association (NBA) and Major League Baseball (MLB) have a dedicated analytics department. Soccer, however, has been a late bloomer in the integration of advanced data analysis, despite being considered the world's most-watched and practiced sport in the world. One of the reasons for this is the difficulty of making sense of the complex spatiotemporal relationships of this game, which are intensified by the high number of players, the large size of the field, and, in particular, the low frequency of goals. As a reference to the difficulty of scoring goals, we find that for the 2014 World Cup, only 10% of the shots were converted from the 1236 shots that were attempted. Also, near 85% of those goals were scored from 15 meters radius from the goal location (Goldsberry, 2019).

Analytical work to date in soccer has focused on isolated aspects of the sport, while coaches tend to focus on the broader tactical interplay of all 22 players on the pitch. Although professional soccer teams are starting to incorporate new data sources and creating data analysis departments, soccer analytics still lacks a comprehensive approach that can start to address performance-related questions that are closer to the game's language. This language poses questions such as "which were the most relevant actions leading to goal-scoring chances?", "are teammates creating valuable space?", "when and how should a backward pass be taken?", "how risky is a team attacking strategy?", "what is a player's decision-making profile?",

"how should we defend against our next opponent to concede fewer spaces in the midfield?" – questions currently unanswered in the soccer analytics literature. To answer these kinds of questions and make an impact on key decision-makers within the sport, we identify two critical demands for successful soccer analytics applications:

1. To capture the dynamic spatiotemporal relationships between the twenty two players and the ball with a level of precision on par with expert practitioners.

2. To provide interpretable models that allow practitioners to conduct fine-grained analysis of game situations both visually and analytically.

Capturing soccer's complex dynamics requires approaches that can robustly approximate non-linear interactions between the different actors of the game, considering its variations in space and time. While either counting events directly or employing rule-based algorithms can provide some general statistics about a game, such as the number of goals, shots, passes, possession ending in shots, recovery balls or fouls, among others, this information is insufficient to assess player and team performance with the level of detail that a professional coach would. However, the recent availability of event and optical tracking data in soccer has provided the opportunity to develop a more sophisticated analysis of this sport's spatiotemporal dynamics.

Tracking data consists of the location of the 22 players and the ball at a frequency rate ranging from 10Hz to 25Hz, and is usually accompanied by event data, which consists of the time and location of on-ball events such as goals, shots, passes, and stop-ball events, among many others (Rein and Memmert, 2016; Stein et al., 2017). In order to approach more complex and useful concepts employing event and tracking data, such as space creation, decision-making, or team-dominance, artificial intelligence (AI) techniques and methods stand as an especially appropriate solution to approximate complex functions from observed data. Tracking and event data have been recently used for inspecting a variety of specific game situations in soccer, with particular emphasis on the use of machine learning and statistical inference approaches. Some of these applications include the quantification of pass risk and quality (Power et al., 2017; Rein et al., 2017; Spearman et al., 2017), estimating goal expectation from shots (Lucey et al., 2014; Link et al., 2016), predicting the value of individual actions (Decroos et al., 2019; Singh, 2019; Gyarmati and Stanojevic, 2016), assessing the off-ball positioning quality in shooting opportunities (Spearman, 2018), estimating the

expected value of a possession (Rudd, 2011), predicting of players' movement in time (Le et al., 2017; Dick and Brefeld, 2019), and even the quantification of mental pressure (Bransen et al., 2019).

The usual approaches found in the literature use standard machine learning algorithms and handcrafted features directly derived from data. Most of these feature sets are comprised of spatial and contextual information, with different levels of sophistication. The most used features include the action's location, the distance and angles between players, the goal and the ball, the players' velocity, and information about the event type and the time it takes place. More elaborated features have been developed from tracking data and event data such as defenders' proximity and speed of play (Lucey et al., 2014), time from regaining possession, first-time pass (Power et al., 2017) and the expected time to intercept the ball (Spearman et al., 2017). While we can intuitively recognize that these features might provide value for each of these problems, collaboration with coaches and soccer experts would allow us to identify with much more precision other aspects that might considerably enrich the data analysis process. However, these studies are rarely supported by these experts. On the other hand, only few of these approaches attempt to learn representations directly from the raw data (Le et al., 2017; Dick and Brefeld, 2019; Hubáček et al., 2018), which can lead to losing sight of relevant spatiotemporal information.

An important additional issue to existing approaches is that while most of these seek to produce a quantitative evaluation of observed events, they usually lack visual interpretation of the expected outcome of other potential decisions or the effects of each of the models' parameters. Beyond the accuracy of quantitative analysis of questions of interest, a necessity for effective communication of most of these insights is the ability to provide a visual and interpretable understanding of underlying models. A significant part of the effective communication between data analysts and coaches resides in presenting results that relate to the way the latter analyze and understand the game. The added capacity of visual interpretation of data-driven models for observed actions and unobserved potential actions during the game is an as-yet little-explored area in sports analytics.

If we think about predicting the probability of making a successful pass, for example, one way to provide visual interpretation is to produce a probability surface that shows the passing probability for each field's location. Despite the interpretive capacity that a visual representation of this type would

provide, few studies focus on providing this type of information (Spearman et al., 2017; Spearman, 2018). For example, Spearman et al. (2017) presents a physics-based model where two main elements, the time to reach the ball and the time to intercept the ball, are combined to predict pass probability surfaces. However, the disadvantage of this type of approach is that it requires tailored variables and conditioning factors to be designed for each specific problem, which slows down development and its practical implementation. There is still no approach in the literature capable of generating probability surfaces that can be flexibly adapted to different types of problems and learn from raw tracking data.

While in the current state of soccer analytics approaches to solve specific soccer-related tasks abound, there is no clear path on how we could join such models together into a more comprehensive framework of analysis.

1.1 Motivation

The difficulty of making sense of soccer's complex spatiotemporal relationships and effectively translating findings to practitioners is one of the most significant barriers for integrating data analytics within the coaching staff. To address the nonstop flow of questions that coaching staff deal with daily, we require a flexible analysis framework that allows us to answer these questions quickly, accurately, and interpretably while capturing the complex spatial and contextual factors that rule the game. Such a framework should also allow us to analyze not only actions observed in past matches but also the expected impact of other potential actions available from one situation to another.

A framework of this kind should associate the spatiotemporal characteristics of a game situation with objective indicators of success, so models built on past data can be adequately learned. In soccer, there is an unappealable indicator of success: goals. An effective way to evaluate any game situation's current state would be to answer the question: how likely is that a team scores or concede a goal in the long-term given the current situation? The first work of this kind is found in basketball, where the concept of EPV was first introduced (Cervone et al., 2016b). In the mentioned work, EPV is defined as the number of points that a team in control of the ball is expected to obtain at the end of possession, considering three possible events that can occur at any time: shots, passes, and turnovers. We can directly see that

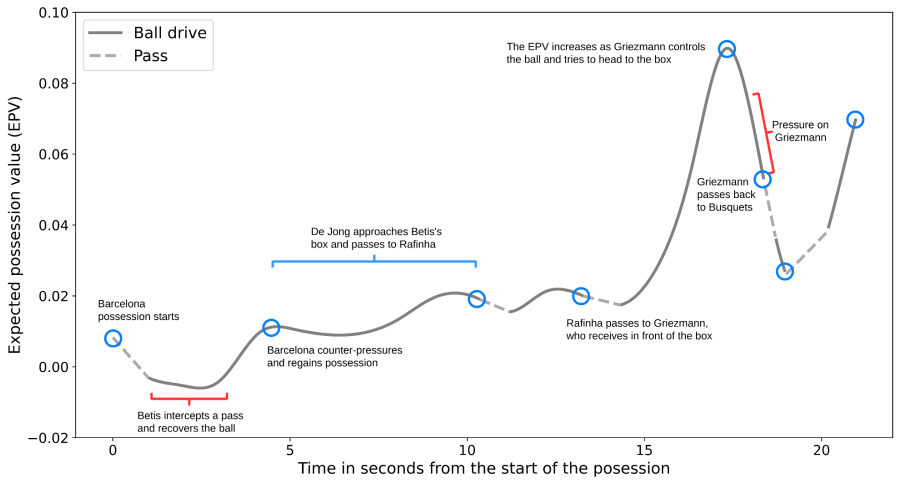


Figure 1.1: Evolution of the expected possession value (EPV) from the perspective of FC Barcelona during a match against Real Betis in La Liga season 2019/2020

a similar approach could be applied to soccer, where an EPV model would focus on estimating the likelihood of observing a goal within a possession. However, aside from a shared high-level goal, our focus on soccer necessitates a drastically different approach to account for the nuances of the sport, such as looser notions of possession, the ability of passes to happen at any location, and space-time dependent turnover evaluation. As one concrete example, in soccer, we cannot assume that passes are played directly to a player's location, as the ball can be played into open space in front of or behind the intended receiver; as such, we need to consider the full space of potential destination locations. As another example, there is no time limit for soccer possessions (aside from the 45 minute half), with complex and often blurred dynamics between offense and defense.

Modeling EPV so that we can estimate which team will score the next goal, given all the spatiotemporal information available, would present a first step towards the comprehensive analysis framework we seek. The frame-by-frame estimation of EPV constitutes a one-dimensional time series that provides an intuitive description of how the possession value changes in time, as presented in Figure 1.1. However, while this value alone can provide precise information about the impact of observed actions, it does not provide sufficient practical insight into either the factors that make it fluctuate or which other advantageous actions could be taken to boost EPV further. To

reach this granularity level, EPV should be modeled to provide fine-grained information about the impact that the game situation's spatial and contextual characteristics have on the final estimate. Moreover, EPV could be expressed in terms of a finite subset of action types that could be taken at any given time, such as passes, ball drives, and shots, which represent a wide range of potential decisions for the team in control of the ball, and whose impact would depend heavily on the spatiotemporal dynamics of the players and the ball.

To capture these dynamics in detail, a model like this would benefit from being developed on top of high-frequency tracking data and making the most of the information available at each location in the field. The use of standard machine learning algorithms could accurately approximate parts of the EPV framework if we incorporate rich spatial and contextual features developed from expert knowledge. On the other hand, from raw tracking data, we could exploit convolutional networks' ability to make sense of spatial regions to learn high-level features and generate full probability surfaces directly.

An EPV framework with the ability of both estimating the EPV itself and providing information about the impact that each of its compounding elements has on the final estimate would provide coaches with more significant analytical insights into the game, especially if this information is provided in a visually-interpretable way.

1.2 Objectives

In the introductory part of this chapter, we present the context for soccer analytics's current state. We argue that while this is a growing research field, the work developed so far addresses isolated aspects of this sport, and there is no existing approach that joins these applications together into a single analysis framework. In Section 1.1 we introduce the development of a comprehensive framework as the main motivation of this work, where spatiotemporal data is exploited to grasp the complex relationships that arise from the interaction of all the players and the ball and produce analytical and visual interpretations that can help coaches to address specific game situations better. We argue that we could obtain the desired framework by modeling the expected possession value and expressing this into a series of familiar concepts. Based on this, the main question addressed in this work

is the following:

- Can we estimate the expectation of a team scoring or conceding the next goal, at any time in the game (EPV)?

From this interrogation, a series of derived research questions arise related to the structure that this expression must-have, the ability to generate calibrated estimates, the ability to produce results that allow the interpretation of its different components, and, above all, the possibility to use the desired model in a practical way. Specifically, these are the other questions addressed in this work:

- Can we express this expectation in terms of a series of smaller components, akin to coaches' language, so they can be estimated and interpreted separately?
- Can the models built for these components produce calibrated probability estimates?
- Can soccer-specific features developed with experts contribute to the models' estimations?
- Can we develop a model capable of ingesting raw tracking data and producing probability surfaces in a way that is easily adaptable to other problems? Can this model be developed through a spatial-aware deep learning architecture?
- Can we produce practical applications from the developed models so the set of EPV components can be understood as an analysis framework?

1.3 Contributions

This section presents the main contributions of this work and indicates the published papers related to each contribution. This thesis was developed as an industrial Ph.D. under the support of the "plan de doctorados industriales del departamento de investigación y universidades de la Generalitat de Catalunya" and carried out entirely at Fútbol Club Barcelona. The work was developed in close collaboration with a series of professional coaches who provided ideas and feedback on numerous aspects of the design, development, and practical applicability of the concepts presented here. A

vast part of the ideas developed in this thesis are now integrated within the club's data analysis methodology and have been adapted to provide team and player performance information on a daily basis.

1.3.1 EPV for soccer as a decomposed model

One of this work's critical contributions is the proposal of a complete theoretical framework for modeling EPV in soccer. We describe the fundamental elements that must be considered to translate EPV effectively into soccer. Specifically, we address considerations such as the scarcity of goals, the fluid nature of soccer possessions, the necessity of considering any location as the potential destination of actions, and selecting a finite but comprehensive set of actions. These considerations constitute the foundations for implementing any EPV model in soccer.

Additionally, we address a challenging problem: how can we employ advanced machine learning algorithms to grasp the complex relationships of soccer dynamics and provide practitioners with non-scientifical backgrounds with the ability to interpret the outputs of the model. We propose a novel decomposed approach to modeling the EPV expression into a series of estimated subcomponents separately. By doing this, the decomposed approach provides a framework for modeling EPV through a series of essential building blocks that allow us to understand the EPV estimates through a set of more easily understandable pieces, which are closer to the language of the coaches. Even if any of the models are difficult to interpret themselves (e.g., black-box models or a high number of parameters), their output can still be interpreted within the context of the EPV expression. An important characteristic of this approach is that any of the components can be easily adapted or retrained when new data is available or when the model needs to be adjusted to specific practical requirements (e.g., adjust pass selection likelihood to a specific team pattern).

We also consider this approach of decomposing a single complex concept into more easily understandable components that are estimated separately and then joined to produce a single estimate contributes to the field of interpretable machine learning.

The publications related to this contribution are the following:

- Fernández J, Bornn L, Cervone D (2019) Decomposing the immeasur-

able sport: A deep learning expected possession value framework for soccer. In: the 13th MIT Sloan Sports Analytics Conference.

- Fernández J., Bornn L., Cervone D. (2021). A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. Machine Learning.

1.3.2 The SoccerMap deep learning architecture

We introduce the SoccerMap architecture, a novel application of deep convolutional neural networks that allows calculating full probability surfaces for developing fine-grained analysis of game situations in soccer. While the majority of existing research in soccer analytics has focused on using handcrafted features for solving soccer-specific problems, few have approached the use of convolutional neural networks for making sense of the full extent of spatiotemporal information. Also, most approaches lack from visual interpretability of the outcomes.

The SoccerMap architecture is a fully convolutional neural network that receives layers of spatiotemporal information and can produce a probability map covering the full extent of a soccer field. The network creates a feature hierarchy by learning convolutions at different sampling scales, allowing it to learn features capturing both global and local relationships. Predictions are produced at each of these scales and then merged into a single probability surface estimation. We show how this architecture can ingest a flexible structure of layers of spatiotemporal data and that it can be easily adapted to provide practical solutions for challenging problems such as the estimation of pass probability, pass selection likelihood, and pass expected value surfaces. We show that networks built for the problems mentioned above also produce calibrated probability estimates. In general, for adapting SoccerMap to any problem, one only has to define the layers of input information, an appropriate activation function for the prediction layer, and the loss function that is better suited to the specific problem (e.g., binary cross-entropy when dealing with pass success or mean squared error when dealing with pass expected reward).

The architecture deals successfully with a challenging learning set-up: learning from single-location labels. For most soccer problems, such as estimating the passing probability or the expected value from passes, there is no ground-truth data covering the extent of a soccer field. Instead, event

data provides only the destination location of the observed event. This poses a learning set-up where we aim to estimate, for example, a 104×68 prediction matrix, but only a single ground-truth value from that matrix is available. We show that selecting only the predicted value at the event destination location, and propagating the loss at that single-element level, is sufficient to tune the network parameters producing accurate and calibrated predictions.

The estimation of full probability surfaces provides a new dimension for soccer analytics. This approach offers a new way of providing coaches with rich information in a visual format that might be easier to be presented to players than the usual numerical statistics. We also consider this approach could also be applied directly in many other team sports, where the visual representation of complex information can bring the coach and the data analyst closer.

The publications related to this contribution are the following:

- Fernández J., Bornn L. (2021) SoccerMap: A Deep Learning Architecture for Visually-Interpretable Analysis in Soccer. In: Dong Y., Ifrim G., Mladenić D., Saunders C., Van Hoecke S. (eds) Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track. ECML PKDD 2020. Lecture Notes in Computer Science, vol 12461. Springer, Cham.
- Fernández J., Bornn L., Cervone D. (2021). A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. Machine Learning.

1.3.3 Pitch influence, pitch control, and other spatial and contextual features

For the framework components where we do not require to produce probability surfaces, we develop a broad set of soccer-specific features to be used as informed inputs for the different machine learning algorithms used. By incorporating these features, we produce point estimates for challenging and yet-unexplored problems, such as estimating the success probability and expected long-term outcome of ball drives and the probability of selecting specific actions in any given situation. We also improve the existing xG model by incorporating knowledge derived from tracking data. We developed these features in close collaboration with professional soccer coaches, which provided meaningful insights regarding the spatiotemporal elements

involved in each of these problems.

In general, we group the developed features into two main types: spatial and contextual features. The spatial features type refers to those directly derived from the location and velocity of players and the ball and considers the soccer field's full extent. Within the spatial features, we propose a new approach for calculating pitch influence and pitch control surfaces, representing a frame-by-frame estimation of the degree of ownership that either a player or a team has on any field location. The pitch influence model provides a way of quantifying the reachability of degree of influence that a player has for every location, based on its current location, its velocity, and the location of the ball. We propose to model this influence as a bivariate Gaussian distribution, where the spread and direction of each of the principal components of the distribution are conditioned to the mentioned players' motion information. From this concept, we develop a pitch control model as an aggregation of the influence that both teams' players have at any location on the field. Essentially, our formulation represents the probability of control of a given team, where a kernel-based non-parametric point process captures each team's latent surface. Specifically, the pitch control estimation provides an estimation of the degree of control that a given team has on any location, considering the 22 players, the ball, and the field's full extent. From this model, we provide the first known quantification of a popular concept among coaches and soccer experts: space creation. Here, we identify the impact of player's movement at different speeds (even standing) to create open spaces to receive the ball and further the possession towards the goal and provide an analytical approach to evaluate coordinative behavior between pairs of players. Additionally, an entire surface of space-ownership can be calculated in real-time and provide immediate visual feedback of the dynamics of creation and closure of spaces. In general, the pitch influence and pitch control features provides rich information related to ball pressure and players' density, two critical factors influencing the long-term expected outcome of a possession.

On the other hand, the contextual features are designed to weigh game situations according to the two teams' location relative to the ball. For example, two game situations where a team in the midfield controls the ball are approached very differently by coaches depending on the opponents' relative location. If all the opponent players are still in front of the ball, this situation is considered a buildup or start-of-play situation and less threatening than an alternative situation where only two opponents are left between

the ball and the goal. We expect the behavior of the players to be substantially different for each of the cases. The contextual factors are then aimed at capturing the difference in spatiotemporal dynamics that occur in two apparently similar situations according to the ball's position. Estimating players' aggregate impact on on-ball actions has also been explored by employing deep reinforcement learning to learn an action-value Q-function, based on event data Liu et al. (2020).

In this work, we propose, to our knowledge, the first approach for calculating teams' dynamic formation lines (or lines of pressure). We employ tracking data and unsupervised learning to find clusters in the x and y-axis to produce vertical and horizontal formation lines. From this information, we can identify situations such as the ball being behind the first line of pressure of an opponent, a pass attempted towards the inside of the formation block, or the lines of defenders potentially being surpassed by an action. Additionally, we incorporate knowledge from previous research and expand the concepts of "packing" and "impect" (Schaper, 2021), and create a broader concept addressing outplayed players (i.e., the number of players that are surpassed after an action is attempted). Lastly, we develop a calibrated xG model from a large set of event data to be used as a baseline model for several of the models here developed.

We present an analysis of the importance of these features within the context of the developed EPV framework, through the use of SHAP values Lundberg and Lee (2017). Additionally, we present how these can be used separately to obtain rich information for analyzing players' and teams' performance.

The publications related to this contribution are the following:

- Fernández J., Bornn L. (2018) Wide open spaces: A statistical technique for measuring space creation in professional soccer. In: the 12th MIT Sloan Sports Analytics Conference.
- Bornn, L., Cervone, D., Fernández, J. (2018). Soccer analytics: Unravelling the complexity of "the beautiful game". Significance, 15(3), 26-29.
- Fernández J., Bornn L., Cervone D. (2019) Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer. In: the 13th MIT Sloan Sports Analytics Conference.

- Fernández J., Bornn L., Cervone D. (2021). A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. Machine Learning.

1.3.4 Implementation of the EPV framework

We provide an implementation of the proposed EPV framework, where each of the ten components is estimated separately, producing an ensemble of models whose outputs can be merged to produce a single EPV estimate. Specifically, we provide functioning models providing estimates for the following problems: pass and ball drive success probability, expected value in long-term for both successful and missed passes and ball drives, joint-estimation of passes and ball drives expected value, location-wise pass selection probability, action selection probability, and the final joint-estimation of EPV. For developing such models, we follow two main learning approaches:

- Learning from soccer-specific spatial and contextual features, developed with professional soccer coaches' aid, and applying standard machine learning algorithms.
- Learning from raw spatiotemporal data to produce full probability surfaces through a novel deep learning architecture based on fully convolutional neural networks (SoccerMap).

We show that both the models built for the different components and the joint EPV model produce calibrated estimations. Additionally, the set of probability surfaces generated from the SoccerMap-based components are shown to provide visual interpretability of the expected impact of potential actions, allowing the framework to provide live question-answering dynamics with coaches.

The publications related to this contribution are the following:

- Fernández J., Bornn L. (2018) Wide open spaces: A statistical technique for measuring space creation in professional soccer. In: the 12th MIT Sloan Sports Analytics Conference.
- Bornn, L., Cervone, D., Fernández, J. (2018). Soccer analytics: Unravelling the complexity of "the beautiful game". Significance, 15(3), 26-29.

- Fernández J., Bornn L., Cervone D. (2019) Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer. In: the 13th MIT Sloan Sports Analytics Conference.
- Fernández J., Bornn L. (2021) SoccerMap: A Deep Learning Architecture for Visually-Interpretable Analysis in Soccer. In: Dong Y., Ifrim G., Mladenić D., Saunders C., Van Hoecke S. (eds) Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track. ECML PKDD 2020. Lecture Notes in Computer Science, vol 12461. Springer, Cham.
- Fernández J., Bornn L., Cervone D. (2021). A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. Machine Learning.

1.3.5 Broad set of practical applications

In this work, we present over ten different practical applications derived from this work where we exploit the soccer-specific features developed, the visual-interpretability of the probability surfaces provided by the SoccerMap-based models, and, in general, the information-rich components of the EPV framework developed. Specifically, we structure the proposed practical applications into three main groups: match and team analysis, off-ball performance, on-ball performance.

For the match and team analysis category, we present a real-time control room where the framework's different components are employed to provide dynamic inspection into actual game situations on a frame-by-frame basis. We also show how we can generate video summaries of matches and player performance by either querying possession ending with a high EPV value or by identifying abrupt changes in the EPV curve during a possession. Additionally, we show how the pass selection component can be directly adapted to grasp team-based passing tendencies, allowing for a finer inspection of a teams' playing strategy. Lastly, we show how by summarizing the on-ball and off-ball added value between pairs of players, we can identify optimal lineups boosting a team's key player performance.

Despite being one of the most attractive application areas for coaches, off-ball performance analysis is one of the least explored soccer analytics areas. We present novel applications exploiting the information-rich probability surfaces of the passing components. First, we present a defensive

analysis tool, allowing coaches to decide how to defend against a given time in specific game situations. Specifically, we focus on the tactical analysis of buildup phases and show how the pass expected value surface could be used to decide the optimal defending formation against Brendan Rodgers' 2014-2015 Liverpool team. Second, we present a methodology to calculate player's potential positioning in game situations that optimize pass reception. Lastly, we introduce the space occupation and space creation metrics to understand a player's off-ball game. A critical characteristic of these applications is that, rather than providing coaches with a single evaluation of past situations or a single best option, we present a visually interpretable tactical analysis whiteboard that empowers coaches to conduct data-informed analysis and customize their strategies.

We also present practical applications in the on-ball performance category. Since these types of applications are the most covered in soccer analytics, we focused on providing new analysis approaches. We present a different perspective on how to present passing networks by shifting the focus from the frequency of passes between players to the assessment of value added by passes. Additionally, we employ the dynamic formation lines features to analyze player's passing tendencies according to context, which would provide practitioners with more detailed information about specific players' passing tendencies. Finally, similarly to the calculation of optimal locations, we assess the optimal passing locations for specific situations. From here, we assess a player's passing skill based on the difference between the expected probability of optimal passing locations and the observed accuracy of attempted passes.

With this broad set of applications, we intend to show how a comprehensive approach for the EPV concept, like the one presented in this work, can be applied in a versatile and agile way to support coaches in their analysis work on an ongoing basis.

- Fernández J., Bornn L. (2018) Wide open spaces: A statistical technique for measuring space creation in professional soccer. In: the 12th MIT Sloan Sports Analytics Conference.
- Fernández J., Bornn L., Cervone D (2019) Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer. In: the 13th MIT Sloan Sports Analytics Conference.
- Fernández J., Bornn L. (2021) SoccerMap: A Deep Learning Architec-

ture for Visually-Interpretable Analysis in Soccer. In: Dong Y., Ifrim G., Mladenić D., Saunders C., Van Hoecke S. (eds) Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track. ECML PKDD 2020. Lecture Notes in Computer Science, vol 12461. Springer, Cham.

- Fernández J., Bornn L., Cervone D. (2021). A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. Machine Learning.

1.4 Structure of the thesis

In this section, we describe the structure of this thesis. The current Chapter presents the introduction, covering this work's context within the framework of sports analytics, the motivation, and the research questions addressed. Additionally, we describe the main contributions and publications. Chapter 2 describes the background and literature review focusing on two aspects: existing soccer analytics research and the technical background of AI methods and techniques applied in this work. Chapter 3 proposes a theoretical framework for modeling the expected possession value in soccer and introduces the decomposed EPV approach that constitutes the common thread of this work. In Chapter 4 we describe in great detail the different spatial and contextual features developed in this work in collaboration with professional soccer coaches. Additionally, we present a series of practical applications directly derived from these features and the methodology's details followed for the mentioned collaboration. Chapter 5 introduces the SoccerMap deep learning architecture that generates probability surfaces from raw spatiotemporal data. In Chapter 6 we present the technical details for learning models for all the components of the proposed EPV framework from a large tracking data set. We show the models produce calibrated probability estimations and present a finer inspection into the different features' influence on each model's predictions. Chapter 7 presents a broad series of practical applications directly derived from the presented framework and the developed soccer-specific features. These include applications in match and team analysis, off-ball performance, and on-ball performance assessment. Finally, Chapter 8 presents a discussion of the main contributions, conclusions, and limitations of this work and guidelines for future work.

1.5 Publications

In this section, we present the different publications produced during this work. First, we present the list of publications in scientific journals and conferences directly related to this work and where the author is listed as the principal author. We indicate the different chapters of this thesis in which each of the publications contributed. Then we present other research conducted during this thesis directly related to soccer analytics research. Additionally, we present a series of master in Science's thesis where the author acted as an industrial supervisor.

1.5.1 Scientific journal and conference publications

- Fernández J., Bornn L. (2018) Wide open spaces: A statistical technique for measuring space creation in professional soccer. In: the 12th MIT Sloan Sports Analytics Conference. *Awarded third place in the research papers competition.*

  – Chapter 1, 4, 6, 7, and 8

- Bornn, L., Cervone, D., Fernández, J. (2018). Soccer analytics: Unravelling the complexity of "the beautiful game". Significance, 15(3), 26-29.

  – Chapter 1, 2, and 4

- Fernández J., Bornn L., Cervone D (2019) Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer. In: the 13th MIT Sloan Sports Analytics Conference. *Awarded first place in the research paper competition, being the first soccer paper to win this competition.*

  – Chapter 1, 3, 4, 6, 7, and 8

- Fernández J., Bornn L. (2021) SoccerMap: A Deep Learning Architecture for Visually-Interpretable Analysis in Soccer. In: Dong Y., Ifrim G., Mladenić D., Saunders C., Van Hoecke S. (eds) Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track. ECML PKDD 2020. Lecture Notes in Computer Science, vol 12461. Springer, Cham.

  – Chapter 1, 2, 5, 6, 7, and 8

- Fernández J., Bornn L., Cervone D. (2021). A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. Machine Learning.

  – Chapter 1, 2, 3, 4, 6, 7, and 8

1.5.2 Other research conducted

- Arbues-Sanguesa, A., Martin, A., Fernández, J., Ballester, C., Haro, G. (2020). Using Player's Body-Orientation to Model Pass Feasibility in Soccer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 886-887).

  – Chapter 8

- Arbués-Sangüesa, A., Martín, A., Fernández, J., Rodríguez, C., Haro, G., Ballester, C. (2020, October). Always Look on the Bright Side of the Field: Merging Pose and Contextual Data to Estimate Orientation of Soccer Players. In 2020 IEEE International Conference on Image Processing (ICIP) (pp. 1506-1510). IEEE.

  – Chapter 8

- Llana, S., Madrero, P., Fernández, J. (2020). The right place at the right time: Advanced off-ball metrics for exploiting an opponent's spatial weaknesses in soccer. In: the 14th MIT Sloan Sports Analytics Conference.

- Peralta, F., Pinones, P., Sumpter, D., Fernández, J.,(2020). Seeing in to the future: using self-propelled particle models to aid player decision-making in soccer. In: the 14th MIT Sloan Sports Analytics Conference.

- Camenforte, I., Casamichana, D., Cos, F., Castellano, J., Fernández, J. (2020). Diseño y validación de una herramienta de valoración del nivel de especificidad de las situaciones simuladoras preferenciales en fútbol.[Design and validation of a Specificity level assessment tool for preferential simulation situation in football]. RICYDE. Revista Internacional de Ciencias del Deporte. doi: 10.5232/ricyde, 17(63), 69-87.

- Schelling, X., Fernández, J., Ward, P., Fernández, J., Robertson, S. (2021). Decision support system applications for scheduling in professional team sport. The team's perspective. Frontiers in Sports and Active Living, 3, 142.

1.5.3 Master thesis industrial supervision

- Daykin E., Fischer N., Ramirez S. (2018) Using Observational Event Data to Uncover Football Players' Inherent Abilities. Barcelona Graduate School of Economics.

- Soares Afonso, M. M. (2019). Learning state representations and Markov models in football analytics. Thesis for the title of Master in Intelligent Interactive Systems. Pompeu Fabra University.

- Madrero P. (2020). Creating a model for expected goals in football using qualitative player information. Thesis for the title of Master in Innovation and Research in Informatics. Polytechnic University of Catalonia.

Chapter 2

Background and literature review

2.1 Spatiotemporal data in soccer

The rise of sports analytics has been primarily due to the development of semi-automated methods for capturing spatiotemporal data in scale. We will refer as spatiotemporal data to those data sources, including information about the location of players, the ball, or actions, and the time where any of these are observed during a given match. More specifically, we will differentiate the available spatiotemporal data in soccer into two main types, event data and tracking data.

Event data consists of a series of annotated events observed during matches, which include the location and time of the start and end of the event, the name of the player attempting and receiving the action (when it applies), as well as a large set of additional game-related labels, depending on the event. Usually, event data includes on-ball actions such as goals, passes, shots, aerial duels, crosses, set-pieces, tackles, dribbles, or many other typical soccer actions. Additionally, some specialized sources might include off-ball actions, such as ball pressure, or team-related information such as lineups and formations. On the other hand, tracking data consists of all the players' locations on the field and the ball. This data is usually provided at a frequency ranging from 10Hz to 25Hz and captured using computer-vision algorithms on top of soccer match videos. This type of spatiotemporal data is typically obtained through a semi-automated process. First, ball and player locations are automatically recognized and then manually verified

and corrected in the case of misidentification. For both data types, the locations are usually normalized according to the team in possession of the ball (e.g., left to right) and are provided in 2-dimensional (or 3-dimensional) coordinate systems that can be transformed following the length and width of the field dimensions. In the following link, tracking data 2d example, we present a video example of spatiotemporal tracking data, where the yellow and blue dots represent the attacking and defending teams, respectively, and a green dot represents the ball location. The arrows represent the velocity vector of each player, based on the players' location the second before.

These two data sources present an uneven balance regarding richness in detail and availability. While tracking data provides greater detail and volume of information per match, it is considerably more challenging to gain access to it, given its high costs and low availability for its acquisition. On the other hand, event data has a lower level of detail but a considerably broader availability in competition coverage and existing vendors. Usually, to develop spatiotemporal analysis on top of tracking data, we require the latter to be integrated with event data to make sense of human-annotated observed actions and events, such as passes, ball drives, shots, and goals, among many others.

2.2 Soccer analytics

2.2.1 Spatial dominance

Spatial dominance models seek to estimate the degree of control a given player or team has on any given location in a continuous or discrete space. Its application in soccer analytics comes from the necessity of finding mechanisms to quantify the ownership of space, a recurring concept in soccer tactical analysis. A soccer field has an average dimension of 105 x 68 meters and involves 22 players, providing a wide range of complex dynamics to arise from the interaction of these players in such a vast space. The ability to effectively quantify space ownership opens the door to approach the impact of off-ball actions, understood as the behavior and actions of players that are not in control of the ball. Being that the average amount of ball possession for every player is nearly 3 minutes per match, this becomes a fundamental element to master to thrive in this game.

We split the spatial dominance models found in the literature into two groups: distance-based and physics-based models. Distance-based domi-

nance models are those considering the distance of players exclusively to every given point as the determinant factor to decide who controls space. Most of these models make heavy use of the concept of Voronoi tessellations.

Voronoi A Voronoi tessellation partitions space by assigning every location to the closest player. Provided a distance function $d_m^i(t)$ between player i's location and a given location m on the field at time t, the level of control of player i over any given location can be defined as in Equation 2.1.

$$C_m^i(t) = \begin{cases} 1 & i = \text{argmin}_j \, d_m^j(t) \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

When using Voronoi tesselations, a region is assigned to a player if that player is the closest to that region, independently from the player's velocity or any other spatial or contextual factor. A region can also be assigned to a team if a player belonging to that team owns the region, following the previous definition. Voronoi tessellations have been extensively applied in sports for characterizing the continuous space ownership of players (Kim, 2004; Memmert et al., 2017), to characterize the spatial control dynamics of teams (Fonseca et al., 2012; Perl and Memmert, 2016; Masheswaran et al., 2014), and even for understanding the cooperative behavior of robots trained for playing soccer (Kaden et al., 2013; Schiffer et al., 2006; Prokopenko et al., 2014; Law, 2005). A variation of distance-based models uses a weighted version of Voronoi tesselation, where a weighting function is used to account for the relative level of influence in a given location (Cervone et al., 2016a). Following the definition presented in Bornn et al. (2018) we express a weighted Voronoi pitch control model as in Equation 2.2, where $w_m^i(t)$ is the weighting function.

$$C_m^i(t) = \begin{cases} \frac{1}{1+w_i^m(t)} & i = \text{argmin}_j \, w_m^j(t) \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

On the other hand, a physics-based model seeks to calculate the ownership of space, taking into account physical dynamics related to players' and ball movement. One approach presents a player-acceleration-based model (Taki and Hasegawa, 2000). A similar one extends from this concept and includes a force model to consider both accelerations and decelerations, and also provides bounds for players' maximum speed (Fujimura and Sugihara, 2005; Gudmundsson and Horton, 2017). Another approach defines a statistical model for determining the player that most likely will reach the ball if

it travels to a certain location dependent on the time it would take a player to reach that location, and the time it would take to intercept the ball by the opponent (Spearman et al., 2017). The parameters of the model are learned from observed passes in professional soccer matches.

2.2.2 Pass probability

Passes are the most frequently observed event during soccer matches, with an average of 1000 passes per game. Passes allow transferring the ball control from one player to another, allowing the ball to travel both short and long distances effectively and quickly. A pass probability model seeks to estimate the likelihood of a given pass being successful, usually provided information related to the origin and destination location of the pass. Pass probability has been modeled following different approaches. A physics-based modeling of pitch control has been built from observed passes and was used to calculate the probability of successful passes employing spatiotemporal tracking data (Spearman et al., 2017). First-order logic and logic programming have been used to calculate pass probabilities based on a series of constraints derived from qualitative reasoning about the factors involved in the execution and success of passes (Vercruyssen et al., 2016). Other approaches include the use of dominant regions to determine which player is most likely to control the ball after a pass (Gudmundsson and Horton, 2017), and the application of machine learning algorithms on top of handcrafted features to predict pass probabilities, otherwise referred to as pass risk (Power et al., 2017).

2.2.3 Pass selection

Action selection refers to the problem of predicting the following action to be taken by a given player. An extension of this definition is to consider the action destination of the action based on a coarse representation of all the possible destination locations towards which the pass can be taken. Regarding passes, we define the estimation of pass selection as a prediction of the probability of a pass being chosen as a subsequent action. Additionally, some pass selection models include the estimation of the pass destination location, being this either the location of one of the teammates or any other given location of the field. Very few approaches can be found in the literature regarding action selection in soccer and team sports. One approach uses deep convolutional networks from low-level information and structured to consider the location of teammates, opponents, and the level of pressure on each of the teammates (Hubáček et al., 2018). The output of this model

is the probability of passing towards the location of one of the other ten teammates of the team in possession of the ball. Since this model is designed as a multinomial classification problem, there is an inherent difficulty of properly learning pass selection from this approach which is in the choice of teammates order in the output vector. This ordering might be set by the distance of players to the ball handler or the position of the players, among other few qualitative reasons. However, the intricacies of the spatiotemporal dynamics of soccer provide reasons to believe that any ordering of this kind can be prone to overfitting or improper learning. Also, the model does not consider the usual nature of passes to be taken towards a location in the space and not the exact location of players. A different approach in basketball uses a coarsened process for modeling macro-transitions, defined as the prediction of occurrence of discrete actions in basketball within the next interval of time (Cervone et al., 2016b).

2.2.4 Expected goals

Arguably, the most popular metric in soccer analytics is the estimation of the goal expectation, or xG. The xG is usually modeled as the probability of observing a goal after a shot is taken at a given game situation. Given the assumption that a shot is attempted from a given location, the main conceptual challenge of xG models resides in the selection of the factors influencing the outcome of a shot action. Despite its popularity, xG literature is scarce, especially in scientific publications. xG models can be split into two groups: event data-based models (Analysis, 2017; 11Tegen11, 2015; Caley, 2015b,a) and tracking data-based models (Lucey et al., 2014; Eggels, 2016). Most models within these two groups make strong use of features derived from spatial information, and some contextual features. For event data-based models, spatial information is limited to the location of observed on-ball events. On the other hand spatio-temporal data-based events extend the event data location information with rich handcrafted features based on the location of the 22 players and the ball, allowing to produce features related to spatial pressure and density, shot interceptability and player motion information. Both event data and spatio-temporal data sources usually include information about contextual situations preceding the shot such as corners, direct or indirect free kicks, open-play or penalty kicks, or specific events such as dribble, cross pass, long pass, rebound or possession regain (Eggels, 2016). Also, contextual features can be extracted from the full resolution spatio-temporal data to provide information about more specialized game dynamics at time steps close to the moment of the shot, such as the

identification of the attack types including counterattack, fast-attack and organized attack (Lucey et al., 2014). Additionally, player-related summary statistics have been included in order to condition the model to individual player skills (Eggels, 2016).

2.2.5 Player movement estimation

Another area of interest is the estimation of player movement in time, based on spatial and contextual information. The challenge of this task resides in the consideration that the estimation of the next locations where a player can move to is expected to be influenced by a broad set of complex factors such as the positioning of the rest of the players and the ball, the current context of the game, player's role within the team, the score at a given time, and the evaluation of the impact of future actions by the player dependent on the action to be taken next. A recent study approaches the credit assignation to player movements in a continuous time-scale, associated with the likelihood of the attacking team reaching specific locations on the field thanks to these off-ball actions (Dick and Brefeld, 2019). Another approach attempts to find similar player movements using tracking data and through the development of scale, translation, and rotation-invariant representations of movement, and dynamic time warping for comparing trajectories (Haase and Brefeld, 2013).

The estimation of players' movement has also been focused from a team-perspective in a model referred as "ghosting", applied both in soccer and basketball (Le et al., 2017; Seidl et al., 2018). Here, the idea is to predict collective team behavior in a continuous-time scale based on historical information of players' trajectories and dependent on the spatial and contextual game situation at a given time. The problem is modeled through a recurrent neural network using long short-term memory units to learn a single-player behavioral model based on the player's position and a multi-player model to estimate team-level movement behavior.

2.2.6 Action-value and EPV models

The value estimation of individual actions in relation to the future outcome of ball possession has recently gained attention within sports analytics research, especially in soccer. Given the relatively low frequency of goals in soccer in comparison with match duration and the frequency of other events such as passes and turnovers, it becomes very challenging to evaluate indi-

vidual actions within a match. Different approaches have been attempted to learn a valuation function for both on-ball and off-ball events related to goal-scoring. Here, we refer to two types of modeling approaches, action-value, and EPV models. Conceptually, both approaches attempt estimating a long-term expected reward, typically the probability of scoring a goal, given a state representation. When the state representation consists of information specifically related to an observed action, many times including the destination location and action type, we will call these models action-value models. On the other hand, when the state representation is not dependent on the action being taken, but it consists of action-agnostic spatiotemporal information describing the current game situation (i.e., the possession's game state representation), we will refer to these as expected possession value models.

Handcrafted features based on the opinion of a committee of soccer experts have been used to quantify the likelihood of scoring in a continuous time range during a match (Link et al., 2016). The designed features in this approach consider four main aspects for a given game situation: zone, control, pressure, and density. These features are combined through an empirically designed linear equation to produce a valuation of actions during a match. Another approach uses event data also to estimate the value of individual actions during the development of possession (Decroos et al., 2019). Here, the game state is represented as a finite set of consecutive observed discrete actions and, a Bernoulli distributed outcome variable is estimated through standard supervised machine learning algorithms. In a similar approach, possession sequences are clustered based on dynamic time warping distance, and a gradient boosting model (Friedman et al., 2000; Chen and Guestrin, 2016) model is trained to predict the expected goal value of the sequence, assuming it ends with a shot attempt (Bransen and Van Haaren, 2018). A different approach creates a coarsened representation of a soccer field that is learned in an unsupervised way, and then each cluster within the representation is assigned a field value based on pass, shots, and turnover actions observed in a vast dataset of event data. Gyarmati and Stanojevic (2016) calculate the value of a pass as the difference of field value between different locations when a ball transition between these occurs. The estimation of the expectation of a shot within the next 10 seconds of a given pass event has also been used to estimate the reward of a pass, based on spatial and contextual information is used to represent the state in any given game situation Power et al. (2017). Rudd (2011) uses Markov chains to estimate the expected possession value using discrete transition

matrix of 39 states, including zonal location, defensive state, set pieces, and two absorbing states (goal or end of possession), and considering a finite set of on-ball actions. A similar approach named expected threat uses Markov chains and a coarsened representation of field locations to derive the expected goal value of transitioning between discrete locations (Singh, 2019). Beyond the quantification of on-ball actions, off-ball position quality has also been quantified based on the goal expectation. In Spearman (2018), a physics-based statistical model is designed to quantify the quality of players' off-ball positioning based on the positional characteristics at the time of the action that precedes a goal-scoring opportunity. This model allows to rank the quality of opportunities, highlight individual player off-ball positioning, and highlight available potential regions to take advantage of in similar situations.

All of these previous attempts on quantifying action value in soccer assume a series of constraints that reduce the scope and reach of the solution. Some of the limitations of these past works include simplified representations of event data (consisting of merely the location and time of on-ball actions), using strongly handcrafted rule-based systems, or focusing exclusively on one specific type of action. However, a comprehensive EPV framework that considers both the full spatial extent of the soccer field and the space-time dynamics of the 22 players and the ball has not yet been proposed and fully validated. In this work, we provide such a framework and go one step further estimating the added value of observed actions by providing an approach for estimating the expected value of the possession at any time instance.

Action evaluation has also been approached in other sports such as basketball and ice hockey by using spatiotemporal data. The expected possession value of basketball possessions was estimated through a multiresolution process combining macro-transitions (transitions between states following coarsened representation of the game state) and micro-transitions (potential player movements), capturing the variations between actions, players, and court space (Cervone et al., 2016b). Also, deep reinforcement learning has been used for estimating an action-value function from event data of professional ice-hockey games (Liu and Schulte, 2018). Here, a long short-term memory deep network is trained to capture complex time-dependent contextual features from a set of low-level input information extracted from consecutive on-puck events.

2.3 Deep neural networks

2.3.1 Deep feedforward network

A feedforward network defines a mapping $y = f(x;\theta)$, where f is a function approximation of some function $y = f^*(x)$ and θ is a set of parameters resulting in the best function approximation (Goodfellow et al., 2016). This type of model is essentially a composition of functions where a piece of given input information, corresponding to noisy approximations of f^*, is passed through a layered chain of functions (network) and produces an output that is close to y. These different components are commonly referred to as input, hidden, and output layers, respectively. The length of that chain of function compositions representing the hidden layers is commonly referred to as the depth of the network. Each hidden layer is represented by a vector of elements that loosely resembles the current neuroscientific understanding of the role of neurons in the human brain. These elements are referred to as units and resemble neurons in the sense of receiving input from many other neurons and computing its own activation value (Goodfellow et al., 2016). The deep learning term typically refers to network architectures with a depth higher than one. The main objective of this consecutive stack of layers is to map the original input to a more meaningful representation for the specific problem by applying non-linear transformations. The transformation of the original data obtained at the last hidden layer is intended to be a more rich feature set that can then be made sense of through a linear model, following a conceptually similar process than that of the kernel trick (Schölkopf, 2001; Roth and Steinhage, 2000).

Formally, a deep feedforward network is composed by neurons computing the expression $\sigma(w^\top x + b)$ where σ is an activation function and w and b are parameters. The network consists of stacked layers of neurons where the output of each layer can be represented as $\sigma(W^\top x + b)$ where W is the matrix of weight parameters corresponding to each neuron (Goodfellow et al., 2016). When sequentially connecting the layers in a network for k number of layers, where $k \geq 2$, the output of the network can be expressed as

$$\sigma_k(W_k^\top \sigma_{k-1}(W_{k-1}^\top \dots \sigma_2(W_2^\top \sigma_1(W_1^\top x + b_1) + b_2) \dots + b_{k-1}) + b_k) \quad (2.3)$$

where W_i, b_i and σ_i are parameters of the i-th layer, for a network of depth k (Shamir, 2018).

The parameter set θ of deep feedforward networks is usually learned through iterative non-convex gradient-based optimization procedures, where the network parameters are updated according to the negative direction of the gradient of the loss curve. Gradient-descent can be conceptually structured in three types: batch gradient-descent, where the network is updated only after all training samples are fed; stochastic gradient descent, where updates are performed after each example is passed through the network; and mini-batch gradient descent, where the update occurs after batches of examples are passed through the network. Different algorithms apply different strategies to overcome learning-related issues of gradient-descent. The most popular are: stochastic gradient descent (SGD) with momentum, which helps to accelerate gradient-descent by following the vector direction of persistent reduction of the error, Nesterov accelerated gradient, which controls the vector magnitude of momentum (Dozat, 2016), and Adagrad, Adadelta and adaptive moment estimation (ADAM) (Duchi et al., 2011; Zeiler, 2012; Kingma and Ba, 2014), all variations of methods that provide adaptive modifications of the learning rate through the learning process.

When designing a deep neural network, we need to decide about the architecture of the network, the optimization method, the types of units in each layer, and the loss function deciding how to measure the difference between f and $f*$ at each step. In the following sections, we describe a broad set of components that have been successfully incorporated in neural networks for learning complex spatial-aware representations for spatiotemporal problems, which are strongly related to the methods used in this work.

2.3.2 Convolutional neural networks

Convolutional neural networks (convnets) are a structured type of neural network specifically designed to capture spatial structures in data that has been proven to be particularly effective for problems where such spatial structures are expected to exist, such as images, speech, and time-series (Scherer et al., 2010; LeCun et al., 1995). Also, address the incapacity of non-structured networks to deal with translation invariance or local distortions of the input (LeCun et al., 1995). More specifically, convnets provide a way of learning complex spatial local features that preserve the topology of the input by restricting the receptive fields to be local. In general, convnets have proven to be successful in data sources with a Euclidean structure, which makes them an appealing approach for reaching a better level of understanding of the complex spatial interactions of players in soccer.

Meaningful features in a local subspace are expected to also be helpful in other regions of an input matrix. Because of this, convnets reuse a broad set of parameters, allowing a considerable decrease in the size of the network, which makes learning feasible.

One of the main objectives of this work is to provide a clear visual interpretation of the models and their results. convnets have been proven to learn what are sometimes more powerful visual features than handcrafted ones, even given large receptive fields and weak label training Long et al. (2014). In most common applications of convnets in images, the final layer of spatially aware features is condensed into a fully connected feedforward network for classification. Such is the case of object recognition, action recognition, object detection, and object tracking. In these cases, convolutional filters are used as powerful spatial feature extractors, but in the end the spatial relationships are lost by merging them to form the input of standard feedforward networks. Nevertheless, other problems, such as image segmentation, require a pixel-level or location-label labeling of the input data. One type of convnet approaching this problem is the family of fully convolutional networks that replace the last fully-connected layers with 1-d convolutions and can produce a full prediction surface. Fully convolutional networks have been extensively applied to semantic image segmentation, specifically for the pixel-labeling problem (Long et al., 2015; Ronneberger et al., 2015; Chen et al., 2017; Badrinarayanan et al., 2017; Papandreou et al., 2015) to successfully detect wide pixel areas associated with an object in images. These types of networks are particularly interesting for our work in the sense that they provide a comprehensive framework for approaching a two-fold problem: accurate prediction of actions and a visually interpretable surface of predictions that can be mapped to 2d soccer field.

2.3.3 Problem-specific components in neural networks

Several processing components have been built to solve a broad set of specific issues related to neural networks, such as avoiding over-fitting and improve generalization in large architectures, achieving invariance to translation, rotation, and local noise, obtain features at different spatial scales, dealing with the input resizing caused by the application of convolutional filters, and obtaining robust upsampling. In this section, we describe the components most related to the problems addressed in this work.

Pooling

Pooling is an operation performed over convolutional filters that seek to provide representations with invariance to a small translation of the input. Essentially, it provides an aggregation of low-level features over a small neighborhood (Scherer et al., 2010), by replacing the output of the net at a certain location with a summary statistic of the nearby outputs (Goodfellow et al., 2016). There are several types of pooling operations depending on the type of aggregation performed. Given a rectangular neighborhood, max-pooling outputs the maximum value of every pixel, average-pooling produces an average of each value in the neighborhood, which in case of being weighted is called weighted-average pooling (typically regarding the distance to the central pixel) and L^2 norm pooling returns the L^2 norm of the neighborhood.

Upsampling

While pooling can help learn features at different scales, it also reduces the dimension of the original input. However, specific problems require the outputs at different architecture steps to have a greater size than the down-sampled one, for example, to match the original input dimensions. Also, either the down-sampled feature set might be too small to be visually evaluated, or an eye-pleasing surface is desired to be produced. We can learn a composition of convolutions to perform linear or non-linear upsampling for all of these cases. A common choice is to use transposed convolutions or deconvolutions, which refer to an operation in the opposite direction of a convolution that maintains its connectivity patterns (Dumoulin and Visin, 2016). A transposed convolution upsamples the size of a rectangular neighborhood by using a larger region where the missing locations are typically filled with zeros (zero-padding) and applying an analogous operation than that of convolution (Shi et al., 2016). Despite their popularity in problems requiring an output of larger resolution than previous inputs, it has been shown how transposed convolutions are prone to produce visual artifacts that can hurt both the prediction capacity and the visual results of the model (Odena et al., 2016). A usually better working solution is to apply a linear upsampling method such as nearest neighbors upsampling and passing the output through a stack of convolutional layers of arbitrary receptive fields. This stack of convolutions will learn a smoothing function on top of the previously upsampled output. If non-linear activation functions are used, a non-linear upsampling operation can be learned directly (Odena et al., 2016).

Skip connections

The first introduction of skip connections in deep learning architectures is presented within an architecture designed for achieving image segmentation through fully convolutional neural networks (Long et al., 2015). The concept of skipping refers to combining outputs of the network produced different depths within the architecture. The original idea for image segmentation was to merge coarse and finer features at different scales to combine different detail levels. Basically, lower layers skip to higher ones, resulting in more nuanced layers making sense of finer-scale predictions and coarse layers making sense of higher scale predictions, managing to find a tradeoff in the preservation of both local and global structure (Long et al., 2015). Recently, skip connections have been incorporated in other architectures to achieve to provide an easier gradient flow from output layers to layers closer to the input (Wang et al., 2017; He et al., 2016; Goodfellow et al., 2016).

Dropout and Batch normalization

The concept of dropout has become a popular component in modern deep neural network architectures as a mechanism for addressing overfitting. In the original paper, it was shown to provide performance improvement for a wide range of applications in supervised learning, including computer vision, speech recognition, document classification, and computational biology (Srivastava et al., 2014). Applying dropout means temporarily removing units from the network by randomly sampling from a binomial distribution of fixed probability p. Dropout can be considered as a stochastic regularization approach, where the expected loss function is minimized under a noise distribution (Srivastava et al., 2014).

An alternative regularization method that has been increasingly incorporated in recent convolutional neural networks is batch normalization. Batch normalization addresses the internal covariate shift problem, which consists on changing the distribution of each layer's input during training, which slows down training and increases the need for careful parameter initialization (Ioffe and Szegedy, 2015). This mechanism learns normalization parameters that modify the mean and variance of layer inputs, allowing to normalize the input of subsequent layers and also reducing the dependence of gradients on the scale of the parameters or initial values provided (Ioffe and Szegedy, 2015). In mini-batch gradient descent, normalization parameters are learned by mini-batch but can be further generalized to unique

parameters after training, so it is not batch-dependent during inference.

2.3.4 Loss functions

A critical component of any machine learning algorithm is the loss or cost function $L(f(x;\theta), y)$, providing a numerical estimation of the difference between the predicted and ground-truth values. Through an optimization procedure, we attempt to either minimize or maximize a loss function, whose overall estimate for the available data provides a numerical evaluation of the model's predictive ability. The selection of the loss function has direct implications on how the model parameter's θ are learned and how the overall performance of the model is evaluated. In this section, we describe the characteristics of a set of loss functions that are used throughout this work. For defining these functions, we will assume that $\hat{y}$ represents the set of predictions resulting from the learned model $f(x;\theta)$, y represents the corresponding set of observed outcomes for each example of a set of N examples, and M represents a set of discrete outcome classes.

Cross-entropy

When we train our model using maximum likelihood estimation, a standard approach in modern neural networks, we are essentially minimizing the cross-entropy between the distributions of the training data and the models' output (Goodfellow et al., 2016). The cross-entropy function is usually referred to as the negative log-likelihood and can be defined as a loss function following Equation 2.4.

$$\mathcal{L}(\hat{y}, y) = \mathrm{H}(\hat{y}, y) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} \cdot \log(\hat{y}_{ij})) \tag{2.4}$$

Logarithmic loss or binary cross-entropy

When we have binary outcomes, a commonly used function is the log-loss, which is essentially the cross-entropy function for $M = 2$. The binary cross-entropy can then be expressed as in Equation 2.5.

$$\mathcal{L}(\hat{y}, y) = \mathrm{logloss}(\hat{y}, y) - \frac{1}{N}\sum_{i=1}^{N}(y \cdot \log(\hat{y}) + (1-y) \cdot \log(1-\hat{y})) \tag{2.5}$$

The binary-cross-entropy is often optimized in machine-learning models when we aim to produce probabilistic predictions for a discrete set with binary outcomes. A usual case is the logistic regression, a type of generalized linear model where the outputs are constrained to a $[0, 1]$ range by applying the logistic function to the linear combination of coefficients and inputs. Here, the coefficients are estimated using maximum likelihood estimation, minimizing the binary cross-entropy.

Mean squared loss

A commonly used loss function in regression problems is the mean squared loss, where we compute the average of the squared difference between observations and predictions.

$$\mathcal{L}(\hat{y}, y) = \text{MSE}(\hat{y}, y) = \frac{1}{N}\sum_{i}^{N}(y_i - \hat{y}_i)^2 \tag{2.6}$$

The mean squared error (MSE) assumes the outcomes are normally distributed. This loss function is sensitive toward outliers and penalizes the larger errors more than, the smaller ones.

2.3.5 Activation functions

In Section 2.3.1 we described how activation functions are used to produce the output of the neurons in both the hidden and output layers in feed-forward neural networks. Activation functions can be either linear or non-linear. While linear activations are helpful for providing an exact mapping of the inputs and weights in each layer, the selection of non-linear activations allows producing non-linear transformation of the inputs, thus providing the neural network the capacity of approximating complex non-linear functions. A critical aspect of a non-linear activation function is that it has to be differentiable so it can be used for calculating and propagating the loss within gradient descent-based optimization methods Nwankpa et al. (2018). Here, we refer exclusively to the family of ridge activation functions, where the function is applied to a linear combination of the input variables, unlike other classes of functions such as radial basis functions.

The selection of the activation functions in each layer plays a fundamental role in ensuring the proper propagation of the gradient and fine-tuning of the weights across the network to ensure better generalization. Additionally, the capacity of selecting activation functions provides neural networks

with higher customization capabilities than standard machine learning models. This is particularly important for the output layer, where different activation functions allow different outcome distributions and thus higher flexibility for using appropriate loss functions.

In this section, we describe the set of activation functions used across this work. We will refer to each function as $\sigma(x)$, where x represents the linear combination of weights and inputs $w^T x + b$ and σ corresponds to the activation function, following the definition presented in Section 2.3.1.

Linear activation

Equation 2.7 presents the linear activation function, often called the identity function. Following the expression, we can see that the function produced an output identical to the linear combination of the inputs and weights received. This function becomes particularly useful when no additional transformation needs to be applied to a given layer's inputs.

$$\sigma(x) = x \tag{2.7}$$

Sigmoid activation

The sigmoid activation consists in applying the logistic function defined in Equation 2.8. This type of function is usually called a "squashing function" since the input is limited to a defined range of values. In the case of the sigmoid activation, the output is constrained to the $[0, 1]$ range.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.8}$$

The main characteristics of the function are being a differentiable function accepting real values and with positive derivatives at every point Nwankpa et al. (2018). Sigmoid activations are often discouraged for its use in hidden layers of deep neural networks, for presenting issues such as gradient saturation and slow convergence, among other problems (Nwankpa et al., 2018).

Softmax activation

The softmax activation uses the softmax function defined in Equation 2.9. The function receives a vector v of real numbers, and produces a vector of the same size where each value is transformed into the $[0, 1]$ range, and

such that the sum of all the values in the vector is equal to 1. Essentially, the softmax function "squeezes" a real-valued vector of arbitrary size into a probabilistic representation of its inputs.

$$\text{softmax}(v)_i = \frac{e^{v_i}}{\sum_{j=1}^{K} e^{v_j}} \text{for } i = 0, \ldots, K \tag{2.9}$$

The softmax function is often used in multi-class classification as an activation function for the output layer, with each value of the output vector representing the probability of the example to belong to the corresponding class. The softmax function becomes particularly useful for calibrating the output of neural networks when using the temperature scaling calibration method, presented in Section 2.5.

Rectified Linear unit (ReLu) activation

The ReLu activation function is defined in Equation 2.10. Since its introduction by Nair and Hinton (2010) it has become a widely popularized activation function within deep learning literature. Especially effective as an activation function for the hidden layers, the ReLu function has been reported to present a better gradient propagation and allowing faster training of deep networks (Dahl et al., 2013)

$$\sigma(x) = \max(0, x) \tag{2.10}$$

2.3.6 Stochastic gradient descent

One of the main advantages of using deep learning models for approaching complex problems is the capacity of these networks to approximate nonlinear functions. In most linear models and other machine learning approaches, we are able to achieve convergence to optimal solutions after optimization, either through linear equation solvers or using convex optimization methods Goodfellow et al. (2016). However, the non-linearity of neural networks involves loss functions that are usually non-convex, where global convergence can not be guaranteed. Based on this, iterative gradient-based optimization methods have been successful for training deep neural networks, where the network parameters are iteratively updated in proportion with the gradient of the loss function selected. Essentially, gradient-descent methods calculate the derivative of the loss function and move small steps in the opposite direction of the derivative, to produce slight changes towards the minimization of the function. For deep learning architectures, where the number

of input parameters is expected to be greater than one, we calculate the partial derivatives, according to each input variable, which is generalized in the concept of gradient, which composes the derivative relative to a vector Goodfellow et al. (2016). The proportion of the step down the gradient is known as the learning rate, a critical parameter to be tuned when training neural networks. In principle, when optimizing this function, we aim to obtain a global minimum or the point where we obtain the lowest value of the function. However, depending on the initialization of the network's weights, the selection of the learning rate, and other conditions of the learning procedure, the optimization is subtle to reach local minimum points. At these local minima, the optimization is not able to further decrease the function by making infinitesimal steps.

Usually, we require large training sets for training deep learning models and achieving a good generalization, especially given the high number of parameters. However, this increasingly high number of examples and operations becomes training more computationally expensive, up to the point where propagating the loss for each example because infeasible. For dealing with this scenario, we can estimate an expectation of the error with a smaller set of examples instead of using the entire dataset, which conforms the basis of the SGD method. For doing so, we can sample a subset of examples (usually referred to as minibatch) drawn uniformly from the training set and propagate the average of the gradient for the minibatch only (Goodfellow et al., 2016). This approach allows training larger networks and increasing the number of examples in the training set while keeping the training times feasible.

2.4 Markov decision process

The concept of Markov Decision Process (MDP) provides a mathematical framework for modeling the decision-making of an agent in a given environment. A MDP is formally defined by a tuple (S, A, π, V) representing a set of states S and actions A, a transition probability function π and a function V evaluating the expected value of transitioning from one state to another by taking a given action. Let A be the set of actions an agent can take from any state of the set S; the objective is to learn the optimal policy $\pi(s, a)$ (i.e., the probability of taking action $a \in A$ at the state $s \in S$) that maximizes the long-term reward from the current state, assuming the Markov property. A

state-value function $V(s)$ is then defined as the expected return (sum future discounted reward) from state s, based on policy π, while action-values can be defined as the expected return of taking action a, at state s, $Q(s, a)$.

Temporal difference learning methods have shown to be successful for learning $Q(s, a)$ (from now on Q), specifically by addressing problems with large and non-finite states by using function approximation methods. Approaches found in literature usually rely on the Bellman equation, which expresses the value of a decision in a given state as an aggregation between the value at that state and the expected value of its successors states (Sutton et al., 1998). The two main learning approaches for finding optimal policies are policy iteration and value iteration. Both use a form of the Bellman equation, following Bellman's principle of optimality by breaking the decision problem into optimal sub-problems. The main difference is that in value iteration, the value function is iteratively updated, and the optimal policy is derived from the optimal value function, while in policy iteration, the policy is iteratively updated and reused while updating the value function (which applies the updated policy). Value-iteration gave rise to the concept of Q-learning, where the Q function estimation is adjusted based on a balance between immediate reward obtained by the action and the estimation of future reward, weighted by a discount factor. In practice, this equation is used as an update rule or loss within a function approximation procedure.

In practice, reinforcement learning is typically applied to problems where one of a pre-defined set of actions at a given state can be tested, and rewards can be observed. Based on this, the future expected reward can be calculated by applying the function's current parameterization to obtain the expected value of each pre-defined action and the next state. Ideal problems that account for these characteristics are found in most board games and video games, where game simulators can be used to explore the search space in this way. These kinds of environments have been exploited in recent years for research on reinforcement learning approaches (Mnih et al., 2013; Silver et al., 2016, 2017). However, when dealing exclusively with historical data, as in the case of soccer tracking data, we need to figure out a way of learning the desired distribution (Q, π or V depending on the objective) with a finite set of observed actions and rewards, which are expected to be a less comprehensive set of events. Additionally, the selection of long-term and short-term rewards from data is less straightforward in soccer than in other games. The difficulty resides in finding an objective metric of success, so the reward is expressed accordingly. While the long-term reward can be defined

based on the observed goals or shots in the future, the short-term reward is more challenging to define since fewer observed actions can be attributed to a numerical reward value as objective as scoring or conceding goals.

2.5 Probability calibration

Probability calibration deals with the task of producing reliable estimates that the examples of a distribution belong to a given class (Rüping, 2006). Depending on the complexity of the problem being modeled, the direct classification of an example with a class might be insufficient or uninformative from a practical standpoint. This situation can be even more pressing in problems where the classes in the target distribution are unbalanced. For example, in a sport like soccer, where goals are scarce, the task of predicting whether a shot will produce a goal lacks practical value for coaches. A more helpful outcome would be to understand the probability of that shot producing goal, which, on the other hand, will allow evaluating situations as highly advantageous, despite producing seemingly low probabilities. Here, we address two concepts related to producing calibrated probabilities: evaluating a model's calibration and the post-hoc application of methods for calibrating a model once trained.

2.5.1 Assessing model calibration

Regarding the evaluation of the calibration of a model producing probabilistic outputs, a common approach is using a calibration reliability plot, also referred to as reliability diagram (Wilks, 1990). These kinds of plots compare the average predicted probability for a set of values in several ranges within the extent of the outcome distribution and the average observed outcome or class. Typically, the predicted values are grouped into sets of bins, following either a uniform or a quantile-based binning approach. These diagrams are usually validated visually, where the objective is to assess how far the points are plotted in a 2-dimensional (2D)-axis diagram from the ideal calibration line described by the function $y = x$. While this visual validation might be useful to obtain a quick assessment of the segments where the predicted probabilities range might differ from the average outcome, having a quantitative metric for calibration becomes useful to compare calibration across different models. One metric of this kind that has been used successfully for evaluating the calibration of neural networks (Guo et al., 2017; Nixon et al., 2019) is the expected calibration error (ECE) presented in Naeini

et al. (2015). ECE measures the weighted average of the difference between the accuracy and confidence of the predictions and observed outcomes across a series of bins. Borrowing notation from Nixon et al. (2019), let n_b be the number of predictions in bin b, $\text{acc}(b)$ and $\text{conf}(b)$ the accuracy and confidence of bin b, respectively, and N the number of examples, the ECE metric can be expressed as in Equation 2.11.

$$\text{ECE} = \sum_{b=1}^{B} \frac{n_b}{N} |\text{acc}(b) - \text{conf}(b)| \tag{2.11}$$

2.5.2 Post-hoc model calibration

After a model is trained, a model's calibration assessment might show poorly calibrated probabilities along with the set of bins. A common approach for dealing with this is performing a post-hoc calibration, where the predicted probabilities and the observed outcomes are employed to train a model that learns how to transform the predicted probabilities to improve calibration. Two methods have become increasingly popular in machine learning literature, Platt calibration (Platt et al., 1999) and isotonic regression (Niculescu-Mizil and Caruana, 2005). Platt's method applies a sigmoid transformation on the predictions, which can essentially be approached as learning a logistic regression on top of the predicted probabilities (Niculescu-Mizil and Caruana, 2005). Literature shows that this kind of calibration has been proved useful when the calibration reliability plot shows an s-shaped calibration curve. On the other hand, isotonic regression aims to find a transformation function that has to be monotonically increasing (i.e., isotonic). An example of this is the pair-adjacent violators' method (Niculescu-Mizil and Caruana, 2005) which finds a step-wise non-decreasing function that transforms ranges of predicted probabilities into single values.

Regarding deep neural networks, Guo et al. (2017) show that many architectures larger than standard two-layer feedforward networks produce poorly calibrated models. Specifically, they identify that depth, width, weight decay, and Batch Normalization considerably influence the calibration of the final learned models. To deal with this problem, Guo et al. (2017) introduce the temperature scaling method, which can be applied to networks having a softmax activation as output. Temperature scaling consists of dividing the vector of logits passed to a softmax function by a constant temperature value T_p. This product modifies the scale of the probability vector produced by the softmax function. However, it preserves each element's ranking, im-

pacting only the distribution of probabilities and leaving the classification prediction unmodified.

An important observation for performing post-hoc model calibration is that the calibration functions or models should not be learned directly from the training data to avoid bias between the predictions and observed data. A hold-out dataset should be used instead.

2.6 SHAP (Shapley additive explanations)

The ability to interpret the predictions of a model becomes a critical factor when the results of a model are intended to be used in practice by field experts with a non-scientific background. This is particularly important in sports analytics, where most complex statistical models being built address questions or concerns elevated by practitioners, and its results are intended to be applied by these in practice. This set of practitioners include coaches, game analysts, physical coaches, sports executives, and even players. However, to grasp most team sports' complex relationships, we usually require to build models with increasing complexity in its structure, which often produce accurate predictions but lower interpretability. Additionally, a deep understanding of the factors impacting the models' outcomes would support the development of improvements or refinements in the modeling process.

Lundberg and Lee (2017) present a unified framework for interpreting model predictions, named SHAP. The authors identified that a set of six different methods used for providing explanations on different types of models, are strongly framed within the additive feature attribution group of methods. The main approach followed in this methods is constructing an explanation model such that each feature gets attributed an effect and the sum of all these effects approximate the output the model, thus providing a simplified approach for understanding the size of the effect of each feature (Lundberg and Lee, 2017). The authors introduce the concept SHAP values which are strongly based on the other concept of Shapley values (Shapley, 2016), an approach strongly based on game-theory for estimating the mentioned effects that comply with the properties of local accuracy, missingness, and consistency, all critical element for additive feature attribution models, and further explained in (Lundberg and Lee, 2017). Essentially, SHAP values consist of Shapley values that attribute a weight to a feature based

on the observed change in the model prediction when only that feature is available. Interestingly, we can learn how much the model deviates from the expected value of the model (average prediction) when that feature is available, providing a deeper understanding of the impact of each specific feature on the model's predictions.

In this work, we use the Kernel SHAP method, a model-agnostic approach for estimating SHAP values on any prediction model. The technical implementation is borrowed from the Github repository developed by the authors of the original paper (Lundberg, 2020). From the calculation of SHAP values for each feature and sample, we make substantial use of two different plots. One of these visualizations represents the average impact of a feature on a model's output, presented through a horizontal bar plot, which essentially represents a feature importance plot. The average importance of each feature is a valuable piece of information for understanding the relevance of each feature, identifying features that might not be needed (or redundant), or even identifying a solid focus (and sometimes bias) of the model to certain types of features. An additional visualization used is the representation of a bee swarm plot where for each feature and sample on the hold-out dataset, we represent the SHAP value of the feature for that example.

Chapter 3

A theoretical framework for the expected possession value (EPV)

This chapter presents a theoretical framework for modeling EPV in soccer. We first provide a general definition of EPV and introduce a series of crucial aspects that must be clearly defined when developing an EPV model. Then, we propose a modeling approach where EPV is addressed as a Markov decision process, and the primary EPV expression is decomposed into a series of soccer-specific components to gain greater interpretability of the model for practical purposes. Finally, we provide an overview of how the different chapters of the thesis contribute to theoretical and experimental work for modeling the proposed EPV framework, using spatiotemporal tracking data from professional soccer matches.

3.1 Defining the expected possession value

The expected possession value is essentially an estimate of which team will score the next goal at any given time. Let $G \in \{1, -1, 0\}$, where the values represent the team in control of the ball scoring next, the other team scoring next, or the match half ending, respectively; the EPV corresponds to the expected value of G. This value can be seen as an instantaneous estimate of the future reward that a team can expect to receive, given all the information available at a given time. Having this, it becomes critical to decide what kind of soccer-related information would be necessary to provide accurate estimates of EPV, ensuring that it encompasses the essential concepts taken

into account by soccer experts (e.g., coaches) when interpreting the game. While many factors might be considered to influence EPV to a certain degree (e.g., weather, teams' physical condition, historical rivalry, teams' strength, among others), we consider that the spatiotemporal information derived from players and ball's locations is the fundamental piece of knowledge for successfully modeling EPV. The presented model is designed to be applied on top of spatiotemporal tracking data and assuming the data is accompanied and synchronized with event data, consisting of annotated events observed during the match, indicating the location, time, and other possible tags.

3.1.1 EPV as a Markov decision process

This problem can be framed as a MDP. Let a player with possession of the ball be an agent that can take any action of a discrete set A from any state of the set of all possible states S; we aim to learn the state-value function $EPV(s)$, defined as the expected return from state s, based on a policy $\pi(s, a)$, which defines the probability of taking action a at state s. By approaching EPV in this way, we are essentially focusing on the problem of estimating the long-term reward (EPV) that a team in possession of the ball might expect, according to the game situation (state) at any given time. To estimate this, we need to represent the game state with soccer spatiotemporal data, define the series of discrete actions that a player can take at any time (A), and estimate how probable it is that a player takes that action ($\pi(s, a)$), given the game state. In contrast with typical MDP applications, our aim is not to find the optimal policy π (i.e., what is the best action the player can take), but to estimate the expected possession value from an average policy learned from historical data (i.e., which are the most likely actions).

Let Γ be the set of all possible soccer possessions, and $r \in \Gamma$ represents the full path of a specific possession. Let Ψ be a high dimensional space, including all the spatiotemporal information and a series of annotated events, $T_t(r) \in \Psi$ is a snapshot of the spatiotemporal data after t seconds from the start of the possession. And let $G(r)$ be the outcome of a possession r, where $G(r) \in \{1, -1, 0\}$, with 1 being a goal is scored by the team in control of the ball, -1 being a goal is conceded, and 0 being that the match half ends.

Definition 3.1.1. *The expected possession value of a soccer possession at time t is $EPV_t = \mathbb{E}[G|T_t]$*

Following Definition 3.1.1, we can observe that EPV is an integration

over all the future paths a possession can take at time t, given the available information at that time, T_t. Here, T_t is essentially a subset of data from all the possible spatiotemporal information that could be available (Ψ), taken at time t. At this modeling stage, we want to express that, while $T_t(r)$ could take many different shapes depending on the implementation and data sources available, it essentially represents the available data that the outcome of the possession G will be conditioned to, when estimating $\text{EPV}_t = \mathbb{E}[G|T_t]$. Note that since the probability of a team scoring equals the opponent team conceding probability and vice versa, we can estimate and express the EPV from either team's perspective. Following this, G could equivalently be parameterized as the home team scoring next, the away team scoring next, or the half ends. However, we stick to the perspective of the controlling team throughout for ease of narrative.

This model design approach of EPV as a MDP shares similarities with previous approaches in other sports, such as basketball (Cervone et al., 2016b), American football (Yurko et al., 2020), and ice-hockey (Schulte et al., 2017). Aside from a shared high-level goal, our approach is drastically different from these others, driven by the underlying differences between the mentioned sports. In Section 3.1.2 and Section 3.2 we present the particularities of our proposed approach for EPV in soccer.

3.1.2 The foundational elements of the EPV for soccer

Before designing a theoretical framework for EPV in soccer, we need to understand the game's main characteristics and how these impact the likelihood of scoring or conceding goals. Below we list a series of considerations that drive the selection of critical components of the model, such as the definition of the state space, the action space and the reward function, the type of data to use during the learning and inference stages, and the factors that condition the overall EPV expression.

3.1.3 Goals are objective but scarce

The only unobjectionable metric of success in soccer is scoring goals. However, observing a goal is a rare event, typically expected to happen in the range of $[2, 2.6]$ times per match, or once every 60 times after a team regains ball control. This is a considerably lower frequency of the score in comparison with other sports such as basketball, where the likelihood of scoring points is considerably higher, and the game rules impose a time-constraint

to possessions. In the initial definition of EPV presented in Section 3.1 we are implictly associating a higher "value" of a possession with a higher likelihood of scoring a goal. More specifically, we are focusing on the estimating the expectation of observing goals in the future. Following this, the mentioned imbalance and scarcity of goals must be considered when defining an estimand for the long-term reward of a possession. Alternative metrics of success could be considered, other than goals, with the advantage of increasing the number of positive outcomes but with an added risk of introducing noise. Some examples of these metrics are: observing a shot within the possession(Power et al., 2017) or after fixed amount of on-ball actions (Decroos et al., 2019), or the estimated xG of a shot observed within the ball control of a given team (Bransen and Van Haaren, 2018).

3.1.4 Soccer possessions and long-term rewards

Although there is no definition of possession within soccer rules, this concept is frequently used in game analyses as units encompassing sequences of actions. The standard approach is to consider possessions the time slots in which a team controls the ball. However, this definition may vary depending on the problem and the analysis approach. To estimate the expected possession value, we need to provide a clear definition of soccer possessions. In the context of the EPV, we require possessions to have three well-defined elements: the starting time, the ending time, and an observed outcome. Let Q be a directed acyclic graph that has a set of possible initial states and absorbing states (i.e., a state that, once entered, cannot be left), we will say that a soccer possessions is represented by a finite directed path q, generated from Q. The starting and ending time of the possession is defined by the time either an initial or absorbing state is reached, respectively. When we reach an absorbing state, we will say the possession resets, and the outcome or reward of the possession is defined by the absorbing state.

In this work we assume that possessions start from a single initial state represented by kick-offs (i.e., the first event taking place after a half starts or a goal is scored). Regarding the outcome, we assume three main possible absorbing states: one of the two teams scores a goal or that a match half ends. To facilitate the learning process, we consider an additional absorbing state that is reached when a goal is not observed after a fixed amount of time ϵ, from the start of the possession. The value of ϵ limits the time span in which an observed action influences a future reward. Notice that when $\epsilon = \infty$ we return to the original definition of three absorbing states. In

other action-value approaches related to the EPV , possessions are defined as consecutive sets of actions of the same team and assume narrower absorbing states (Power et al., 2017; Decroos et al., 2019; Bransen and Van Haaren, 2018), such as the alternative definitions of long-term rewards described above in Section 3.1.3. We propose a broader definition of possessions, where both teams' on-ball actions can be included within the possession time range, and the outcome of the possession is set according to the next team scoring a goal or a match halve ending.

3.1.5 Selecting a finite set of actions

In theory, the set of possible actions that a player with the ball can take at any time instance is infinite. There is, however, a small and discrete set of actions that soccer practitioners frequently refer to when describing the game, including passes, shots, ball touches, take-on, and aerial duels, among many others. Additionally, human-annotated events for these and other action types are easily accessible for professional soccer matches. To achieve model tractability and increase the understanding of the model by using familiar terms for coaches, the definition of a finite action-space becomes useful. While a large set of events might provide a more fine-grained EPV model, it might also be more challenging to learn and interpret the model.

In this work, the action space is composed by three main types of actions: passes, ball drives and shots. We will consider a pass any action where a player intends to transfer the ball's control to a teammate (including successful and missed passes). Shots are all the actions where a player kicks the ball intending to score a goal. We will broadly define a ball drive as the action of a player maintaining control of the ball before the next action is observed or the game stops (e.g., dead ball or half end). In general, we consider that these three represents the smaller set of actions that encompasses most observable soccer actions. The vast majority (if not all) of named soccer actions are derived from these three main actions. For example, crosses can be seen as a type of pass that origins near the opponent's box and attempted toward the box. Similarly, other types of passes that coaches usually refer to, such as passes breaking a formation line, long passes, or a pass back, can be seen as different types of passes with different contextual conditions. Similarly, ball drives are expected to encompass more specific types of ball carries, such as dribbles, ball drives breaking lines, and the long or short ball carries, in general. A similar explanation follows shots. In general, we considered that a finite and reduced set of actions would facilitate the model's

interpretation for the decomposed model approach presented in Section 3.2.

3.1.6 Comprehensive spatiotemporal information

Given the large dimensions of the soccer field (up to $104m \times 68m$) and the high number of players (twenty field players and the two goalkeepers), the success in a soccer game is highly influenced by the off-ball dynamics of the players. The legendary player and coach Johan Cruyff once said that "players only have the ball three minutes on average in a match, what matters most is what they do when they do not have the ball". Following this, a complete EPV model should make sense of soccer's complex spatiotemporal dynamics, including the player's positioning, motion information, team-structure, ball-pressure, and spatial dominance. For doing so, the state representation of the model must be as comprehensive as possible, including both spatial and contextual information capturing the mentioned dynamics. In this sense, spatiotemporal tracking data, providing the location of players and the ball at a high frequency rate, would allow to extract seemingly meaningful information that is missing in on-ball event data. Additionally, event data is a fundamental source to include the observed on-ball actions and add completeness when learning an EPV model. In this work, we will make a strong focus on leveraging both tracking and event data to obtain a comprehensive EPV model.

3.1.7 Passes can go anywhere on the field

A fundamental concept in soccer practice is that, most often than not, passes are attempted toward an unoccupied space on the field rather than directly to a player's location. Given the large size of the field and the difficulty of being precise either when passing or controlling the ball, the management of the changing space dynamics becomes a critical factor influencing a possession's outcome. Specifically, regarding passes, the success and expected reward of two passes with the same origin and destination location may vary drastically depending on factors such as the pressure on the ball carrier, the density near the destination location, or the motion dynamics of the players at the moment of the pass. Beyond the evaluation of observed passes, an EPV model capable of considering every other possible destination location would provide rich information for assessing off-ball performance, evaluate risk and reward balance on decision-making, and identify potential actions with a better expected outcome.

3.2 A decomposed approach

Both the estimation of the EPV at a given time instance and the time-series of EPV values during a time range can provide useful information for soccer practitioners. Some examples are evaluating the goal-scoring probability created during an attack, identifying valuable actions during a possession (by detecting large slope changes), or even the generation of highlight reels. However, beyond the single-value estimation of EPV, practitioners would benefit from gaining more detailed information about the factors that influence a certain EPV value to be higher or lower. More specifically, and in order to comply with the fundamental soccer characteristics presented in Section 3.1.2, we seek an interpretable model that provides rich analytical information about the impact that players' positioning in space and both observed and potential on-ball actions have on the overall EPV estimate. Additionally, being soccer a game where space and time management is critical, a comprehensive EPV model should contemplate the information available at every location on the field. To reach this granularity, we propose to express EPV as a decomposed model, where the different components represent a series of fundamental concepts that influence the long-term expected value of the possession. By doing so, we can take into account the different spatiotemporal considerations mentioned above and gain greater insight into the impact that smaller parts have on the EPV estimate.

Formally speaking, to obtain this desired structured modeling, we will further decompose Definition 3.1.1 following the law of total expectation and considering a discrete set of on-ball actions. We assume that the space of possible actions $A = \{\rho, \delta, \varsigma\}$ is a discrete set where ρ, δ, and ς represent pass, ball drive, and shot attempt actions, respectively. We can rewrite Definition 3.1.1 as in Equation 3.1, where the expected value of G is conditioned to the set of possible actions A, and weighted according to the probability of selecting an action, expressed by $\mathbb{P}(A = a|T_t)$.

$$\mathrm{EPV}_t = \sum_{a \in A} \mathbb{E}[G|A = a, T_t] \quad \overbrace{\mathbb{P}(A = a|T_t)}^{\text{Action selection probability}} \tag{3.1}$$

Additionally, to consider that passes can go anywhere on the field, we define D_t to be the selected pass destination location at time t and $\mathbb{P}(D_t|T_t)$ to be a transition probability model for passes. Let L be the set of all the

possible locations in a soccer field, then $D_t \in L$. On the other hand, we assume that ball drives (δ) and shots (ς) have a single possible destination location (the expected player location in the next second and the goal line center, respectively). Following this, we can rewrite Equation 3.1 as presented in Equation 3.2. This expression ensures that both the components estimating the expected value of passes and the pass selection probability are conditioned to consider every possible destination location on the field.

$$\begin{aligned} \mathrm{EPV}_t = (\sum_{l \in L} \overbrace{\mathbb{E}[G|A=\rho, D_t=l, T_t]}^{\text{Joint expected value surface of passes}} \overbrace{\mathbb{P}(D_t=l|A=\rho, T_t)}^{\text{Pass selection probability}}) \, \mathbb{P}(A=\rho|T_t) \\ + \overbrace{\mathbb{E}[G|A=\delta, T_t]}^{\text{Expected value of ball drives}} \mathbb{P}(A=\delta|T_t) \\ + \overbrace{\mathbb{E}[G|A=\varsigma, T_t]}^{\text{Expected value from shots}} \mathbb{P}(A=\varsigma|T_t) \end{aligned} \tag{3.2}$$

The expected value of passing actions, $\mathbb{E}[G|D, A=\rho]$, can be further extended to include the two scenarios of producing a successful or a missed pass (turnover). We model the outcome of a pass as O_ρ, which takes a value of 1 when a pass is successful or 0 in case of a turnover. We can then rewrite this expression as in Equation 3.3. In this step, we are enforcing the expression to consider the impact of the action-outcome, as well as conditioning this outcome to the selected destination location.

$$\begin{aligned} \mathbb{E}[G|A=\rho, D_t, T_t] = \overbrace{\mathbb{E}[G|A=\rho, O_\rho=1, D_t, T_t]}^{\text{Expected value of successful/missed passes}} \overbrace{\mathbb{P}(O_\rho=1|A=\rho, D_t, T_t)}^{\text{Probability of successful/missed passes}} \\ + \mathbb{E}[G|A=\rho, O_\rho=0, D_t, T_t] \, \mathbb{P}(O_\rho=0|A=\rho, D_t, T_t) \end{aligned} \tag{3.3}$$

Equation 3.4 represents an analogous definition for ball drives, having O_δ be a random variable taking values 0 or 1, representing a successful ball drive or a loss of ball control following that ball drive, which we will refer

as a missed ball drive.

$$\begin{aligned}\mathbb{E}[G|A=\delta,T_t] = &\overbrace{\mathbb{E}[G|A=\delta,O_\delta=1,T_t]}^{\text{Expected value of successful/missed ball drives}}\overbrace{\mathbb{P}(O_\delta=1|A=\delta,T_t)}^{\text{Probability of successful/missed ball drives}} \\ &+\mathbb{E}[G|A=\delta,O_\delta=0,T_t]\,\mathbb{P}(O_\delta=0|A=\delta,T_t)\end{aligned} \tag{3.4}$$

Finally, the expression $\mathbb{E}[G|A=\varsigma]$ is equivalent to an xG model, described in detail in Section 2.2.4. In Figure 3.1 we present how the outputs of the different components presented in this section are combined to produce a single EPV estimation, while also providing numerical and visual information of how each part of the model impacts the final value. All the proposed components represent concepts that are familiar to soccer practitioners. Ideas such as identifying that a particular pass might have higher scoring value or lower likelihood to be completed, that certain shot attempts are more likely to become goal than others, or that the next action to select might be impacted by the location of the 22 players and the ball, are part of the analysis mindset of professional soccer coaches. By providing coaches with a tool for both analytical and visual interpretation capabilities that considers these familiar concepts, we expect to ease the integration of data-driven analysis within professional coaching staff.

Essentially, this structured approach allows us to express the future expected outcome of a possession as a combination of the expected impact of taking a pass, attempting a ball drive, or shooting to the goal. Additionally, by further decomposing the model considering both successful and unsuccessful outcomes for any of these actions, we provide our EPV framework with capabilities for assessing the risk associated with every possible action. This granular structure enables one to develop separated and more fine-grained models for each of the problems comprised in these components. This decomposed nature allows the models to be extended or adapted independently, which becomes a useful feature for integrating new data sources or usage-specific constraints without re-estimating the full model. For instance, the xG component, $\mathbb{E}[G|A=\varsigma]$, could be re-estimated when there is the availability of new observed shot events, without needing to re-estimate the rest of the models related to passes and ball drive actions. In other examples, the action selection probability might be re-calculated to account for team-specific passing tendencies (e.g., short pass vs. long pass playing styles), or the pass probability component could be extended to include new

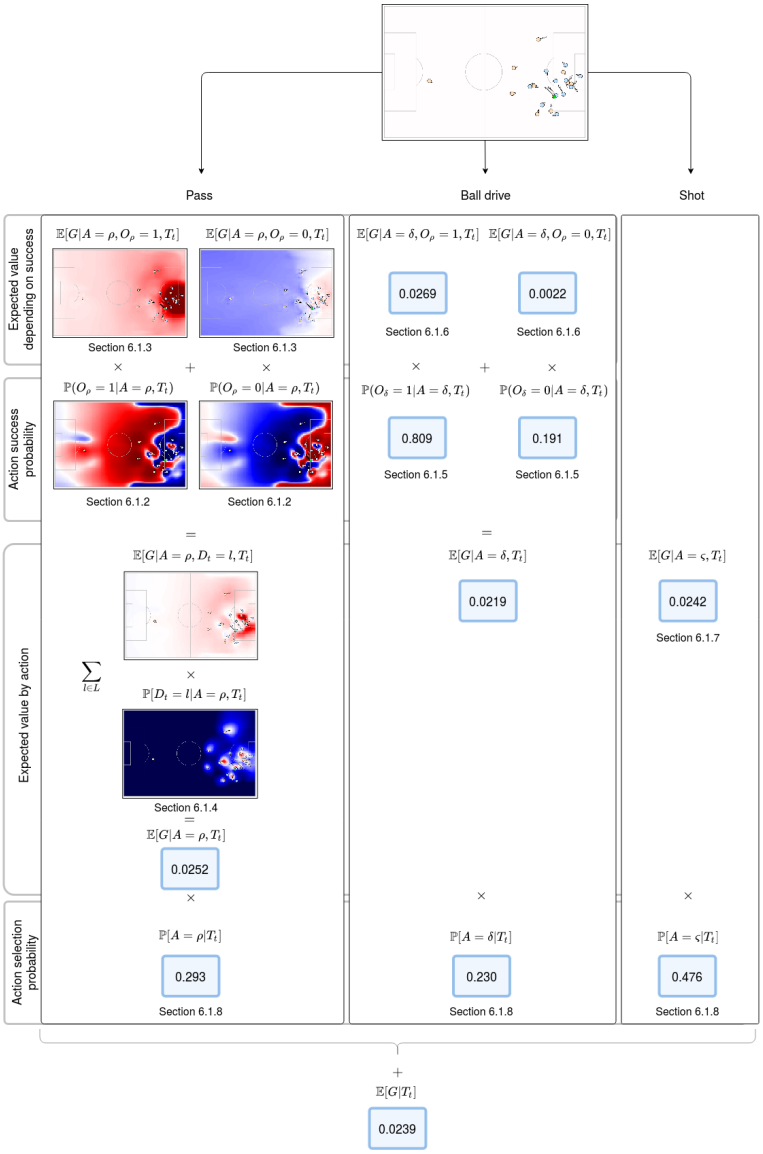


Figure 3.1: Diagram representing the estimation of the EPV for a given game situation. The final EPV estimation of 0.0239 is produced by combining the expected value of three possible actions the player in possession of the ball can take (pass, ball drive, and shot) weighted by the likelihood of those actions being selected. Both pass expectation and probability are modeled to consider every possible location of the field as a destination. The predicted surfaces for successful and unsuccessful potential passes and the surface of destination location likelihood are presented.

features related to players' body orientation when passes are taken (Arbués-Sangüesa et al., 2020a).

3.3 Developing a comprehensive model

To develop a functional model for the EPV framework presented in this chapter, we need to estimate each of the separated components presented in Section 3.2. In Section 3.1 and Section 3.1.2, we argued that a comprehensive EPV model that can be used in practice to provide information to coaches necessarily requires capturing a set of fundamental spatiotemporal dynamics of soccer. For any game situation, this includes considering the location of the 22 player and the ball, the idea that passes can be attempted towards any location on the field and incorporating contextual information that is expected to impact the player's decision making.

In the following chapters, we present the development process followed in this thesis to estimate each of the proposed decomposed approach components, using machine learning and statistical models on top of spatiotemporal tracking data from professional soccer matches. In Chapter 4 we develop a statistical model for estimating the spatial control from both teams' and players' perspectives. From this approach, we derive a set of spatial features, including the ball-carrier pressure and player's relative density across the field, used for estimating several of the EPV components. In Chapter 5 we introduce a deep learning architecture, named SoccerMap, that provided the groundwork to estimate full probability surfaces for the pass probability, pass selection likelihood, and pass expected value components. In Chapter 6 we first present a set of novel spatial and contextual features derived from spatiotemporal tracking data that are used as representative features for estimating each of the components. We then present the learning methodology and the experiments carried out to build calibrated models for all the EPV components and the joint EPV estimation. Between these chapters and Chapter 7 we present over ten different practical applications derived from this framework providing detailed examples of how a soccer practitioner can develop a fine-grained analytical and visual interpretation of real soccer situations.

Chapter 4

Developing spatiotemporal features for soccer

In this work, we approach the estimation of EPV by following the comprehensive framework presented in Chapter 3, where EPV is defined based on a series of components. Estimating these components necessitates thorough game state representations to capture the complex spatiotemporal relationships involved in each of these problems. For example, the pass probability component would benefit from features providing information about all the players' location and their velocity to estimate the probability of attempting a pass into an open space. Similarly, the shot expected value component could produce better estimations if we provide information about the number of players potentially blocking the shot or opponents' density in the shooting area. Following this idea, ball drives success might be influenced by the level of spatial pressure on the ball carrier, and the expected value from ball drives might vary drastically depending on the relative location of the opponent team.

This chapter presents a detailed definition of a comprehensive set of features derived from tracking and event data. Figure 4.1 presents a visual representation of the main features described in this section. These features constitute the essential building blocks for game representations used across the different models built for estimating the expected possession value framework (presented in Chapter 5 and Chapter 6). The features developed can be either low-level or high-level. Low-level features are those that can be directly derived from tracking and event data, such as the locations and velocity of players or the distance and angle between the players and the event

or goal location. On the other hand, high-level features comprised a series of soccer-specific knowledge and were developed in close collaboration with a group of professional soccer analysts from FC Barcelona. Conceptually, we group all the developed features into two groups: spatial and contextual features. In this chapter, we present the technical details followed for developing these features and a series of practical applications where these can be used independently for advanced analysis of soccer situations. Finally, we present the methodology followed within the collaboration with the soccer analysts that contributed to this work.

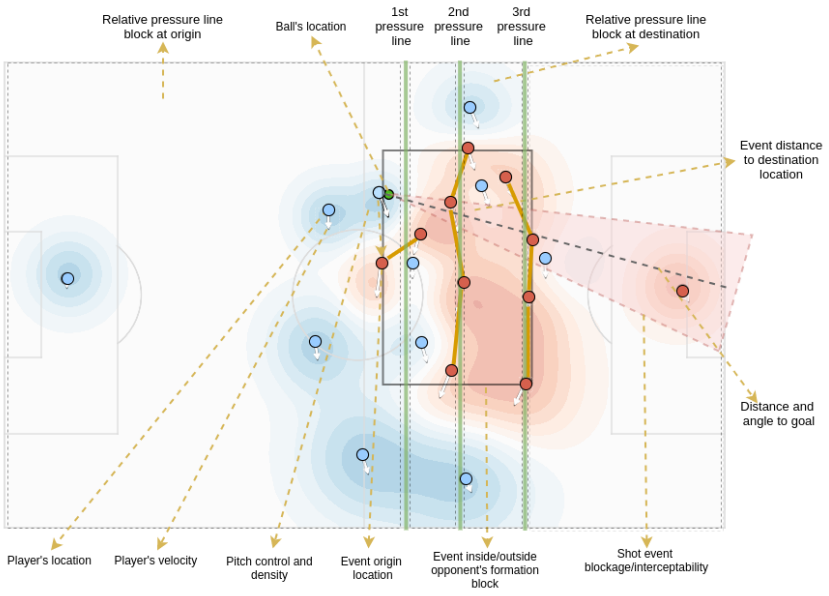


Figure 4.1: Visual representation of a series of spatial and contextual features in a soccer match situation. Blue and red dots represent the attacking and defending team players, respectively, while the green dot represents the ball location. The blue and red surface represents the pitch control of each team along the field. The grey rectangle covering the red dots represents the defending team's formation block. The green vertical lines represent the defending team's vertical dynamic formation lines, while the polygons with solid yellow lines represent the players clustered in each pressure line. The black dotted rectangles represent the relative locations between dynamic formation lines. Dotted yellow lines and associated text describe the main extracted features

4.1 Normalizing spatiotemporal data

The different models presented in this work are developed on top of two main types of spatiotemporal data in soccer: tracking data and event data. The main characteristics of both data types are explained in detail in Section 2.1. While the specific sampling rate and event types available might vary from one data provider to another, we will assume that a set of minimum information is provided for both data sources. For event data, we assume that the origin location, the event type, and the time the event occurs are available. Additionally, we will assume that the destination location is available for passes. In the case of tracking data, we assume the data includes the location of all the players and the ball, the team to which player belongs, and that the data is provided at least once for every second. For both data sources, we assume the x and y location take values in a $[0, 1]$ interval, and the dimensions of the soccer field where the match took place are available (normally available in public sources as well). Also, the data must include a "half start" event indicating the start of every half of the match. We will refer to both data sources as spatiotemporal data.

Following the decomposed model presented in Section 3.2 we are approaching the estimation of the expected possession value from the team's perspective in possession of the ball. Specifically, we are evaluating the long-term reward of the possession based on the actions that the player in control of the ball can take at any moment. From the perspective of a given team, we can easily observe that the perceived value of moving the ball close to the opponent's goal is different from moving the ball close to their own goal in a mirror location. Based on this, we normalize the spatiotemporal data from the perspective of the team carrying the ball at any given time, so the defending team's goal is always on the rightmost side of the field, and their attacking team's goal is on the leftmost side of the field. Following this, for every available action, the data is normalized based on the perspective of the team taking that action.

Note that to apply this normalization, we first need to identify each team's own goal, given that teams change side in every half of the game. For every half start event, we count the number of players of each team in both halves of the field and label each teams' goal as the goal location belonging to the half with a majority number of players. Figure 4.2 presents three game situations of a match in the first half for two teams: yellow and blue team. We can see that at the half starting time, the yellow team and

blue teams defending goals are identified based on the count of players on each side of the field. A pass event from the yellow team was observed at minute 2, and data is normalized to ensure left to right attacking direction. In the last image, a pass from the blue team was observed in minute 8 of the same half, and the data is normalized to adjust the teams attacking directions.

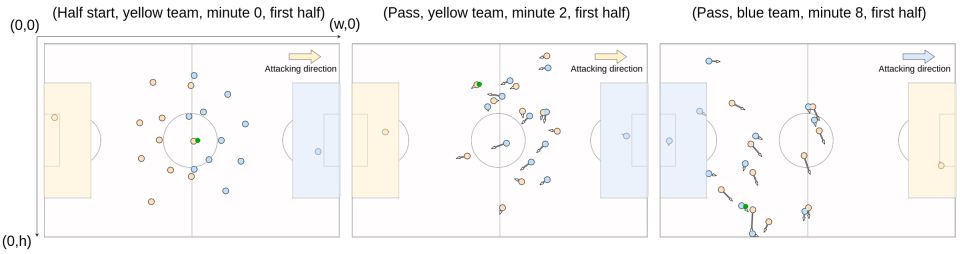


Figure 4.2: Three images representing the normalization of spatiotemporal data based on the team attempting the next action. The left plot shows the identification of both teams' goals based on their location at the time of the half-start event. The center and right plots show how the locations are normalized based on the team taking the next action to ensure left to right attacking direction. The reference system is represented by the two axis on top of the left plot.

On the other hand, while all the spatiotemporal data locations are assumed to be defined in a $[0, 1]$ range, many of the features used in this work would benefit from expressing the locations using a coordinate system easier to interpret. The allowed length and width in professional soccer fall in the $[90\text{m}, 120\text{m}]$ and $[45\text{m}, 90\text{m}]$ ranges, respectively. Given that the soccer field's sizes can vary from one venue to another, it becomes critical to consider this in the data transformation since this will considerably impact the calculation of distances and velocities. In this work, we transform the spatiotemporal locations to the metric system. Given the length and width of a soccer field (w, h) and a given location (x, y) in the $[0, 1]$ coordinate system, we calculate the relative coordinates in meters x_m and y_m following the expression $(x_m, y_m) = (x \times (w - 1), y \times (h - 1))$.

4.2 Spatial features

We consider spatial features directly derived from the players' spatial location and the ball in a given time range. These can be obtained for any game situation regardless of the context and comprise mainly physical and spatial

information. Table 4.1 presents the main types of spatial features explored in this work, and from which the specific features used in Chapter 5 and Chapter 6 are derived. The main spatial features obtained from tracking data are related to the location of players from both teams, the velocity vector of each player, the ball's location, and the location of the opponent's goal at any time instance. From the player's spatial location, we also produce a series of features related to the control of space and players' density along the field. In the following sections, we describe the technical details for deriving features from each of these spatial concepts.

Table 4.1: Description of a set of spatial concepts derived from tracking data

Concept type	Description
(x,y) location	Location of a player, the ball, or attempted action, normalized in the [0,w) and [0,h) ranges, respectively, where w and h correspond to the length and width of the soccer field. The location (0,0) corresponds to the top-left corner.
Distance between locations	Distance in meters between two locations.
Angle between locations	Angle in degrees between two locations.
Player's velocity	Player's velocity vector in the last second.
Pitch control	Probability of controlling the ball in a specific location.
Pitch influence	Degree of influence of a set of players in a specific location.
Interceptability	Features related to ball interception.

4.2.1 Distance, angle, and velocity

Before describing the approaches for calculating distances, angles and velocities, lets define the formal notation followed. Let p and q be two locations in a soccer field such that $\mathbf{p} = (p_x, p_y)$ and $\mathbf{q} = (q_x, q_y)$ where p_x and q_x represent the x-axis coordinate and p_y represents the y-axis coordinate. Let w and h be the length and height of a soccer field, then $p_x, q_x \in [0, w)$ and $p_y, q_y \in [0, h)$.

We calculate the distance between two locations p, q in terms of the

Euclidean distance d expressed in Equation 4.1.

$$d(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\| = \sqrt{(q_x - p_x)^2 + (q_y - p_y)^2} \tag{4.1}$$

We will represent a player's velocity in terms of the change of the location in the x-axis and y-axis in the last second, relative to the opponent's goal. Let p^t and p^{t-1} be the location of a player at time t and time $t-1$, representing one second before, we express the velocity of v at time t following Equation 4.2.

$$\mathbf{v} = p^t - p^{t-1} = (p_x^t - p_x^{t-1}, p_y^t - p_y^{t-1}) \tag{4.2}$$

Another spatial feature used in this work is the angle between a player's location and the opponent's goal. Let $\mathbf{g} = (w, h/2)$ be the location of the opponent's goal, and $\mathbf{p}$ the players' location, the angle to the goal is calculated as the angle between two locations in the plane $\mathbf{z} = (1, 0)$. This calculation is expressed in Equations 4.3 and 4.4.

$$\mathbf{w} = \frac{g - p}{\|g - p\|} \tag{4.3}$$

$$\theta = \text{atan}_2(\|\mathbf{w} \times \mathbf{z}\|, \mathbf{w} \cdot \mathbf{z}) \tag{4.4}$$

We also calculate the angle between a player's velocity vector and every other location on the field $\mathbf{r}$. Let $\mathbf{u}$ be a player's velocity vector centered at a players' current location $\mathbf{p}$, and $\mathbf{v}$ be a vector such that $\mathbf{v} = \mathbf{r} - \mathbf{p}$ then the angle between the two vectors is defined by Equation 4.5.

$$\theta = \text{atan}_2(\|\mathbf{u} \times \mathbf{v}\|, \mathbf{u} \cdot \mathbf{v}) \tag{4.5}$$

4.2.2 Quantifying spatial influence and control

One of the most recurrent aspects in coaches' analysis of soccer situations is the concept of space. Being that the typical size of a soccer field is 7140 m^2 and that the outcome of the game results on the interaction of 22 players within this large area, one can understand that the management of space in time is a critical aspect for successfully disentangling the complexity of this game. The movement dynamics of players in space is one of the most unexplored topics in soccer analytics, yet coaches consider this aspect a fundamental one. There is at least one statistical reason for that: on average, players spend only 3% of their playing time without being in contact of the ball. In the words of one of the most renowned players and coaches in the

history of soccer, Johan Cruyff, "it is statistically proven that players have the ball 3 minutes on average. So, the most important thing is: what do you do during those 87 minutes when you do not have the ball? That is what determines whether you are a good player or not".

Either to measure the success probability of on-ball actions, to estimate the goal expectation of a given play, or to understand the effect of players' movement in time, among many others, almost any kind of spatiotemporal analysis in soccer would greatly benefit from a model that can estimate players' influence and ownership of any location on the field, typically referred as pitch control. Pitch control is a recurrent concept in the analysis of space dominance in team sports. It can be defined as the probability of controlling the ball that a player has on a given location in time, as thoroughly explained in Section 2.2.1. Most of the models found in literature, such as the Voronoi tesselation-based approaches, are designed based on the assumption that a given location on the field is exclusively dominated or influenced by only one specific player. This idea disregards the concept that ownership of space is continuous, not discrete, with uncertainty in who controls areas between players. The distance between players and the ball is also believed to influence the relative positioning and degree of space control, especially for sports with wider playing spaces such as soccer; however, this is not considered by the mentioned approaches.

We propose a novel pitch control model that considers the location, velocity, and distance to the ball for all the players, providing a smooth surface of control for each team. Every player's influence is computed and summarized for any given location, resulting in a probability of control. An additional objective of this approach was to provide a model that could be applied from the information available in a single data frame without requiring any other prior data for learning its parameters. Also, such a model would allow easier reproducibility. Regarding visual interpretation, this model can produce entire pitch control surfaces at $140Hz$ (140 frames per second), allowing coaches to receive a real-time visual representation of space ownership during a match.

In this section, we present the technical details of a pitch influence model, measuring the degree of space reachability or ownership of individual players, and a pitch control model quantifying a team-level degree of control for any location on the field. These two models are used to produce spatial control, and spatial pressure features used extensively in the models presented in

Chapters 5 and 6.

Quantifying players' spatial influence

Before calculating pitch control from a team perspective, we first define a model for measuring individual player influence on the field. We approach this concept of influence as the probability of a player reaching any location on the field, conditioned to the ball's location and player's velocity. There are several reasons to consider these two elements. Depending on their location in time, players are expected to have different influence levels on nearby zones. When a player is far away from the ball, his level of influence can be understood as a wider area, based on the reasoning that if the ball moves towards the player, he would have more time to reach the ball within a larger space. On the opposite, when closer to the ball, the player has less possibilities of reaching the ball if it moves away from its current location. Also, the player's velocity plays an essential role in defining this area of influence. A player moving at running speed might have more influence in the direction of speed than if they were walking or jogging. Furthermore, the player may have higher levels of influence in nearby spaces than in farther spaces.

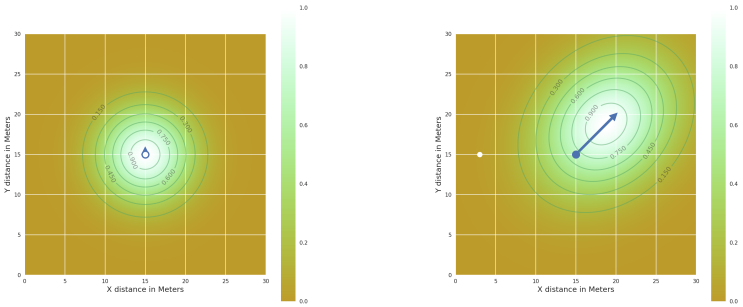
Based on this reasoning, we propose defining the player influence area through a multivariate normal distribution, whose shape can be adjusted to account for the player's location, velocity, and relative distance to the ball. A degree of influence or control can be calculated at any given location through the distribution's probability density function. As explained later in Section 4.2.2 this model was developed in close collaboration with professional soccer coaches. The reason for deciding a prior distribution resides on the need to provide a model as interpretable and flexible as possible, where we could directly introduce the expert considerations on modeling the control of space.

For a given location in space p at time t, the probability density function of player's i reach is defined by a standard multivariate normal distribution f, with mean $\mu_i(t)$ and covariance matrix $\sum_i(t)$ given the player's velocity $\vec{s}$ and angle θ. The mean and covariance matrix are calculated from the player's velocity and location and the ball's location. The details of this calculations are presented in Section 4.2.2. Once we have a model for a players' reach surface model, we define the player's influence model I for any location p, as the value of f at that location normalized by the value of f at player's current location p_i, as shown in Equation4.6, which provides a

degree of influence within a $[0, 1]$ range.

$$I_i(p,t) = \frac{f_i(p,t)}{f_i(p_i(t),t)} \tag{4.6}$$

Figure 4.3 presents the player influence area in two different situations regarding the player's distance to the ball and velocity. Here we can observe how depending on the ball's distance, the player's range of influence varies. Also, the distribution of player influence is reshaped to be oriented according to the direction of movement and stretched in relation to the speed. If the player is in motion, the distribution is translated so the higher level of influence is near points where the player can reach faster, according to his speed. This model can easily be expanded to handle player-specific movement characteristics, such as acceleration and maximum speed.



(a) Player influence area for player in control of the ball and a speed of 1 m/s.

(b) Player influence area for player 12 meters away from the ball, running at 6.36 m/s in a 45 degrees angle

Figure 4.3: Two situations representing the player influence area.

Employing coach-led priors for estimating player influence

This model's development was supported by the expert advice of a series of professional coaches from FC Barcelona (see details on the work methodology in Section 4.5). In this section, we provide technical details for calculating a player's influence surface at any given time and explain the series of soccer-specific considerations that were introduced in the model.

As presented in Section 4.2.2, the influence degree I_i value presented in Equation 4.6 is expressed in terms of the probability density function of a bivariate Gaussian distribution defined by Equation 4.7

$$f_i(p,t) = \frac{1}{\sqrt{(2\pi)^2 \det \Sigma_i(t)}} e^{(-\frac{1}{2}(p-\mu_i(t))^T \Sigma_i(t)^{-1}(p-\mu_i(t)))} \tag{4.7}$$

Given this expression, we have two main parameters that can be adjusted to model the player's influence area: the mean (μ) and covariance matrix (Σ) of the distribution. From a practical perspective, we are modeling the probability of a player reaching a location before the ball. Based on this, we want to model the μ and Σ parameters, so the reach distribution considers the distance of the ball and the player's velocity vector. Specifically, we want to consider that the farther away from the ball, the larger the influence area of the player, and also that the direction and magnitude of the player's velocity will impact that angle and the magnitude of the principal axes of the distribution.

We will first decompose the covariance matrix into a rotation and scale matrix whose parameters are easier to interpret and customize. Using the singular value decomposition algorithm, we can express the covariance matrix as a function of its eigenvectors and eigenvalues as expressed in Equation 4.8, where W is the matrix whose columns are the eigenvectors of Σ, and L is the diagonal matrix whose non-zero elements are the corresponding eigenvalues (Spruyt, 2014). Let $A = W$ and $S = \sqrt{L}$, we can define A as a rotation matrix and S as a scaling matrix, allowing to express the covariance as in Equation 4.9. Based on this, the rotation matrix and scaling matrix can be defined as Equations 4.10 and 4.11, where θ is the rotation angle of the velocity vector and, s_x and s_y are the scaling factors in the x and y direction.

$$\Sigma = WLW^{-1} \tag{4.8}$$

$$\Sigma = ASSA^{-1} \tag{4.9}$$

$$A = \begin{bmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{bmatrix} \tag{4.10}$$

$$S = \begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix} \tag{4.11}$$

Following this decomposition of the covariance matrix, we can now decide the parameters influencing the distribution's shape using soccer-specific information. The rotation matrix is defined by a single angle parameter (θ), which can be obtained by calculating the angle of the velocity vector with respect to the x-axis. By defining θ in this way, the distribution orientation will follow the velocity vector's direction.

On the other hand, we need to define how to calculate the parameters s_x and s_y which represent the spread of the distribution in its two principal components. For defining these scaling factors, we want to take into account both the player's magnitude of speed $V_i(t)$ (in meters per second) and the distance to the ball $d_i(t)$. We will say that both scaling factors are composed by two main elements: a stationary reach function $R_i(d_i, t)$ and a speed expanding factor $E(R_i(d_i, t), V_i(t))$. Given two weighting parameters λ_x and λ_y, then s_x and s_y are defined as in Equations 4.12 and 4.13.

$$s_x = R_i(d_i, t) + \lambda_x E(R_i(d_i, t), V_i(t)) \tag{4.12}$$

$$s_y = R_i(d_i, t) + \lambda_y E(R_i(d_i, t), V_i(t)) \tag{4.13}$$

The stationary reach function R_i estimates a player's influence radius when the player starts from rest, conditioned to the distance to the ball's distance $d_i(t)$. Based on expert soccer analysts' opinion and through experimental observations, we set $R_i(d_i, t) \in [4, 10]$ as the minimum and maximum distance in meters for the reach radius. Specifically, $R_i(d_i, t)$ follows the transformation function shown at Figure 4.4.

Once we have the stationary reach function, we need to decide how the distribution spread changes according to the ball's distance and the player's velocity. Following the observations of the soccer experts we will consider an inversely proportional relationship between s_x and s_y, by setting $\lambda_x = 1$ and $\lambda_y = -1$. Intuitively, as the player's speed increases, the higher the reach in the direction of the velocity and the lower the spread of the y-axis, representing an increased difficulty to changing direction. We will define the expanding factor $E(R_i(t), V_i(t))$ so that it represents an estimate of how much would the stationary reach increase in a given direction, provided the player's current speed. Setting $13m/s$ as the maximum possible speed reachable, we calculate the ratio between player's speed and the maximum speed, expressed by $\text{Srat}_i(V, t)$, as shown in Equation 4.14. Finally, given the two

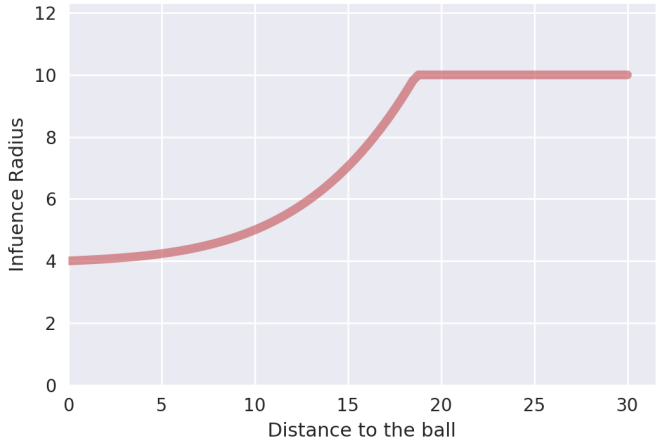


Figure 4.4: Player influence radius relation with distance to the ball

weighting factors γ_x and γ_y, the scaling matrix is defined as in Equation 4.15.

The last parameter we need to define is the distribution mean value $\mu_i(t)$. To take into account player's velocity we will calculate $\mu_i(t)$ by translating the players location p_i at time t in the direction of the speed vector $\vec{s}$, weighted by constant factor γ_μ, as expressed in Equation 4.16.

$$\text{Srat}_i(V,t) = (\frac{V_i(t)}{13})^2 \tag{4.14}$$

$$S_i(d_i,t) = \begin{bmatrix} R_i(d_i,t) + \gamma_x(R_i(d_i,t)Srat_i(V_i(t))) & 0 \\ 0 & R_i(d_i,t) - \gamma_y(R_i(d_i,t)Srat_i(V_i(t))) \end{bmatrix} \tag{4.15}$$

$$\mu_i(t) = p_i(t) + \gamma_\mu \vec{s}_i(t) \tag{4.16}$$

In the final model, the remaining constant parameters are set in the following way: $\lambda_x = 1$, $\lambda_y = -1\lambda_x$, $\gamma_x = \gamma_y = 0.5$, $\gamma_\mu = 0.5$.

4.2.3 Estimating team-level pitch control

We then define pitch control as an aggregation of the influence that player's from both teams have at every location of the field, providing a value of

control within a continuous range. Equation 4.17 presents a team's pitch control at location p and time t, where i and j refers to the index of the player in each opposing team, σ is the logistic function, δ_a and δ_d are weight parameters to allow balancing the overall influence of the attacking and defending team, respectively, and γ_{pc} is an adjusting factor for the logistic function. Here, the logistic function σ transforms the subtraction of each team's accumulated individual influence area into a degree of control within the $[0, 1]$ range. Figure 4.5 presents this probabilistic pitch control surface on a given soccer situation, while Figure 4.6 presents the surface of influence of each player in the attacking team. Pitch influence and pitch control provide a rich summary of players' spatial distribution and impact along the playing surface and can enrich the information on locations where players might have some influence, despite not being directly present. The model used along this work sets the constant parameters to the following values: $\gamma_{pc} = \delta_a = \delta_b = 1$.

$$PC(p,t) = \gamma_{pc}\sigma(\delta_a \sum_i I_i(p,t) - \delta_d \sum_j I_j(p,t)) \tag{4.17}$$

4.2.4 Space quality and value

While the control of space is a useful feature for understanding the level of spatial dominance of player or a team, some concepts such as space creation require to add an understanding of the quality of the controlled space. In other words, the quality of positioning of a given player can be associated with having the best possible control of the space, and doing so for spaces with higher value. We could then express the quality of owned space Q as a function of the level of ownership (control) PC and the value of space V, as presented in Equation 4.18.

$$Q(t) = PC(t)V(t) \tag{4.18}$$

The level of ownership at any location of the field, PC, can be modeled directly using the pitch control model presented in Section 4.2.3. On the other hand, to model the pitch value V, we present in this section an early approach for quantifying the value of any location on the pitch, relative to the location of the defending team. Here we will focus only on the technical characteristics of this model, and later in Section 7.2.3, we present a series of practical applications employing the quality of space owned, Q, where we introduce the concepts of space occupation and generation.

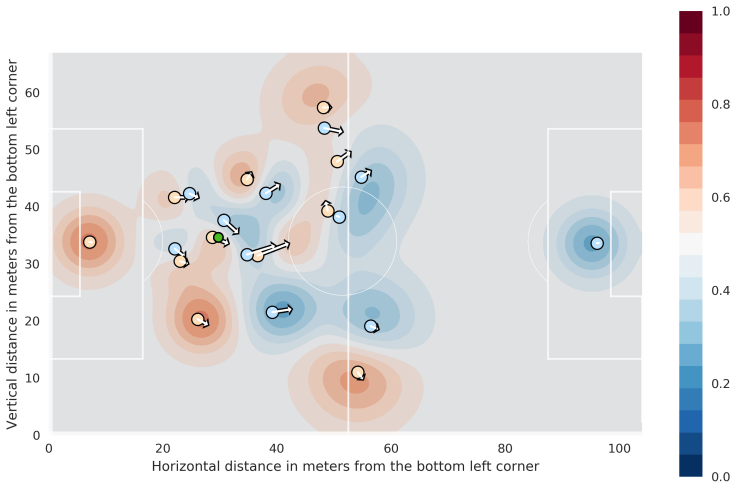


Figure 4.5: A probabilistic pitch control surface for two teams in a soccer game situation. The circle corresponds to the players' location where the attacking team's players appear in blue and the opponent team's players in red. White arrows show the direction of player's velocity vector, ending at the expected location in one second. Pitch control is calculated from the attacking team's perspective, so the higher the value, the higher the control of the attacking team.

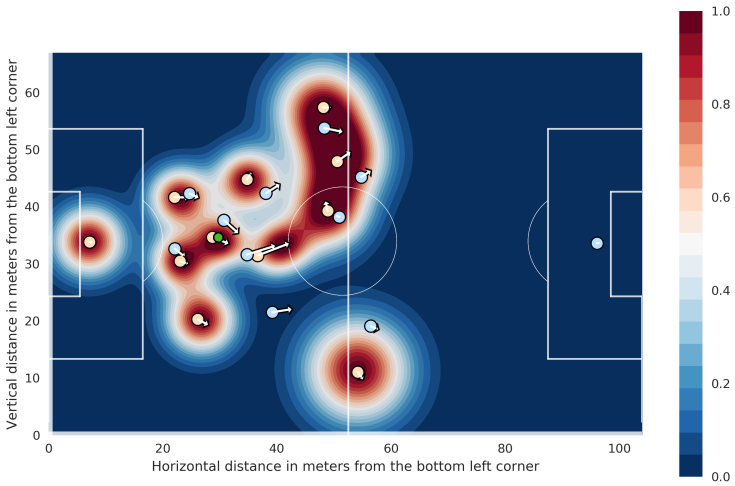


Figure 4.6: The sum of player pitch influence in every location for every player in the attacking team. The circles correspond to players' location where the attacking team's players appear in blue and the opponent team's players in red. White arrows show the direction of player's velocity vector, ending at the expected location in one second.

The sole fact of moving for finding better passing options is an advantage itself. However, it can be easily argued that not every space has the same value. A trivial method for determining the value of space is its distance to the opponent's goal. Its well known that spaces near the goal have an increased value, given the advantage that would provide to dominate them. But exploring more deeply into the dynamics of soccer, and based on the opinion from F.C. Barcelona expert analysts, it can be also argued that the value of space changes dynamically depending on multiple positional factors, such as the location of the ball and the players. In order to quantify in a detailed way the value of the space generated or occupied we provide a model for finding the relative pitch value on every position of the field, depending on the location of the ball. This link presents a video where the dynamic evaluation of pitch value depending on the ball location can be observed, following the pitch value model presented in this section: http://www.lukebornn.com/sloan/field_value.mp4.

We base the development of this model in the following hypothesis: con-

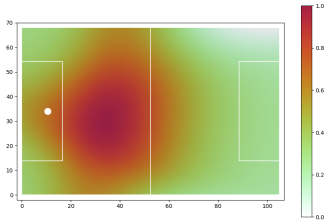
sidering a sufficiently high number of situations, the defending team distributes itself throughout the field in a manner which covers high value spaces. Although it is clear that at any given point defenders will deviate based on overloads, specific offensive player positioning, and other scenarios, in general, most defenders will remain close to high value areas. Based on this, we propose to develop a model capable of estimating the sum pitch influence that a defensive team would have in a given location on the field, given the location of the ball. Formally speaking, we want to learn a function f with parameters θ that is able to estimate the space value at any given location $p_l(t)$, provided the ball location $p_b(t)$. The space value at location V is expressed in Equation 4.19.

$$V_l(t) = f(p_b(t), p_l(t); \theta) \tag{4.19}$$

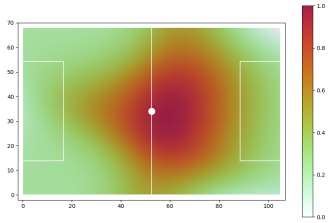
For learning such a model, we use a standard feed forward neural network architecture, which is trained to estimate the defender's team pitch influence at given location p_l, given the ball location $p_b(t)$. We employ Metrica Sports tracking data for 20 matches of the first (La Liga) and third (Segunda B) Spanish divisions, to build a dataset with 2.4 million examples, consisting of the location of the ball and the pitch influence for the defending at any other location in the field. The dataset is constructed in three steps. First, a series of game situations are selected where a given team is in possession of the ball, and the opponent team is identified as the defending team. Second, the sum of pitch influence for every defending player at every location l within a 21×15 grid representing the soccer field, is calculated. Finally, for each of the locations in the field grid, one example of the dataset is produced, consisting on the location of the ball at that time, and the calculated pitch influence for the defending team at that location. To avoid a high correlation between events close in time, we pick time instances within the possession time, that are three seconds away from each other. The pitch influence of the defending team at a location p_l and time t, expressed by $D_l(t)$, is defined in Equation 4.20. The target pitch value $\hat{V}_l(t)$ used during learning is presented in Equation 4.21. This target value equals the defending team's pitch influence value, but is restricted to the $[0, 1]$ range, by imposing a maximum value of 1 to this density.

$$D_l(t) = \sum_d I_d(p_b(t), p_l(t)) \tag{4.20}$$

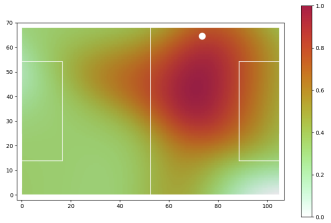
$$\hat{V}_l(t) = \begin{cases} 1 & D_l(t) > 1 \\ D_l(t) & \text{otherwise} \end{cases} \tag{4.21}$$



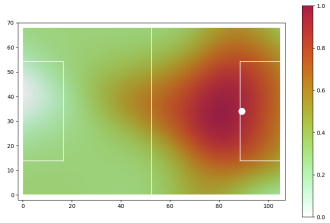
(a) Pitch value for ball vertically centered at the first quarter of the field



(b) Pitch value for ball at the center of the field



(c) Pitch value for ball at the third quarter of the field on top of the left lane

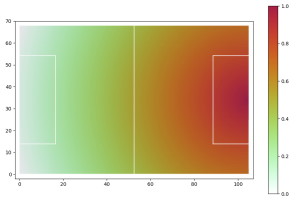


(d) Pitch value for ball vertically centered in the fourth quarter of the field

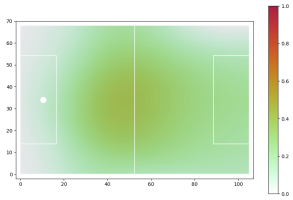
Figure 4.7: Predicted pitch value in a [0,1] range for the ball location represented by a white circle

We randomly sample the available data into a training and test set consisting following a 3:1 ratio. We carry out a 10-fold cross-validation process, using ADAM optimization, and minimizing the MSE loss. We set the β_1 and β_2 parameters to 0.9 and 0.999, and perform a grid search on the learning rate ($\{1e{-}3, 1e{-}4, 1e{-}5, 1e{-}6\}$), and batch size parameters ($\{16, 32\}$). We select the model that minimizes the average error across the 10 folds. Given a ball location, we can now query the model to predict the pitch value at any location on the field. Figure 4.7 shows three different ball position scenarios and the obtained field valuation.

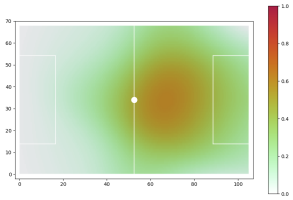
This model has learned that nearby locations to the ball have increasing value for a certain range, while understanding effectively how to translate this value depending on ball position. The model still lacks from the natural intuition that the cumulative value of space is higher when further up the field, closer to the opponent's goal. In order to adapt to this intuitive



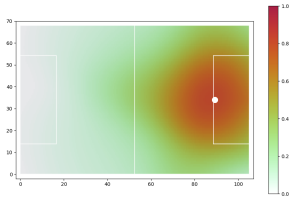
(a) Distance to goal pitch value normalization surface



(b) Normalized pitch value for ball vertically centered at the first quarter of the field



(c) Normalized pitch value for ball at the center of field



(d) Normalized pitch value for ball vertically centered the fourt quarter of the field

Figure 4.8: Predicted pitch value in a [0,1] range for given ball location, represented by a white circle, normalized by a distance to goal model

thinking we normalize the obtained pitch value by the distance to the goal of every location normalized on a $[0, 1]$ range. Figure 4.8 presents the normalization surface and three different pitch value situations, where the results still adapt to ball location but show a more consistent valuation of the pitch which adjusts for the threat of the ball location, according to expert analysts. We see that when one's own goalkeeper has the ball, the overall value of space is limited, but when in the opponent's box, space is much more valuable alongside the looming threat of a shot on goal.

4.2.5 Block count and interceptability

The last spatial concept presented in Table 4.1 refers to the capacity of intercepting the ball. The features derived from the interceptability concept are expected to play an essential role in capturing the opponents' spatial influence near shooting options, allowing us to produce a more detailed model for estimating the expected value from shots. Here, we will focus on two

interceptability features: shot blockage and shot pressing.

The shot blockage feature refers to the number of players blocking a shooting option. We calculate this value by counting the number of defending players located inside the triangle formed by the two posts' location and the ball carrier's location (i.e., the player attempting the shot). To calculate the number of players located within the triangle, we use the point-in-polygon ray casting algorithm described in Hacker (1962) and Haines (1994). Following this algorithm, we say that a player located in a given (x, y) location (for $x, y \in [0, 1]$) is inside the triangle, if the line constituted by the player's location and the point $(1, 0)$ intersects any edge of the triangle an odd number of times, otherwise we say the point is outside the triangle. On the other hand, for the player pressing feature, we count the number of defending players less than 3 meters away from the ball carrier. The professional coaches selected the 3 meters value when being consulted about the higher ball carrier distance considered ball pressure. Additionally, since goalkeepers can touch the ball with their hands, we consider they show different blocking and pressing dynamics, so we do not include then in either of the features.

Figure 4.9 presents two game situations where a ball carrier attempts a shot. In the left case, the defenders are located within the triangle between the ball carrier and the two posts, producing a blockage count of three, while there is no pressure on the ball. In the situation on the right, three players are pressing the ball carrier, including a defender that is also potentially blocking the shot. If we would employ only event data, these two situations would only contain information about the ball carrier's location and would overestimate the expected value of the shot.

4.3 Contextual features

To provide more comprehensive state representations, we include a series of features derived from soccer-specific knowledge, which provides contextual information to the model. Table 4.3 presents the main concepts from which multiple contextual features are derived.

The dynamic lines capture alignment groups between the players of a given team and provide a way of contextualizing the pitch locations relative to the players' locations. For example, by identifying the defending team's

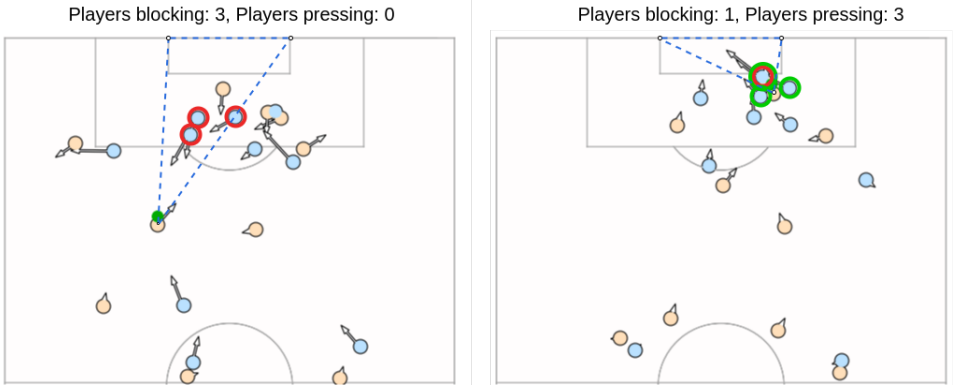


Figure 4.9: Two game situations where a shot event was observed. Yellow and blue circles represent the attacking and defending team, respectively. A red contour indicates a player is located within the triangle formed between the two posts and the ball and can potentially block the shot. A green contour indicates the player is less than 3 meters away from the ball and pressing the ball carrier. We do not show the location of the goalkeepers in either plot.

formation lines at any given time, we can capture contextual information such as potential passes breaking lines or the current phase of the possession, which would allow differentiating spatiotemporal dynamics within a possession. In Section 4.3.1 we present the technical details for calculating the dynamic formation lines from tracking data, while in Section 4.4 we present a series of practical applications that are directly derived from this concept. From the concept of outplayed players, we can derive features such as the number of opponent players to overcome after a given pass is attempted or the number of teammates in front of or behind the ball, among many similar derivatives. This is explained in more detail in Section 4.3.3. Additionally,in Section 4.3.2 we present the technical details for developing the baseline xG model used in Chapter 6.

4.3.1 Dynamic formation lines

The concept of dynamic formation lines refers to players being aligned with their teammates within different alignment groups. For ease of narrative, we will refer to a defending team's formation lines as "pressure lines." A typical conceptualization of pressure lines in soccer would be the groups formed by the defenders, the midfielders, and the attackers, which tend to keep a consistent alignment. By identifying the pressure lines, we can obtain every

Table 4.2: Description of a set of contextual concepts derived from tracking data

Concept type	Description
Dynamic formation lines	Relative positioning of players according to the team's current formation or the opponents.
Baseline event-based models	Models built on top of event data and used as a baseline to enrich the learning of tracking data-based models.
Outplayed players	the number of players that are surpassed after an action is attempted.

player's opponent-relative location, which provides high-level information about players' expected behavior. For example, when a player controls the ball and is behind the opponent's first pressure line, we expect a different pressure behavior and turnover risk than when the ball is close to the third pressure line and the goal.

We will refer to two types of groups: vertical and horizontal formation lines. Vertical formation lines are obtained by clustering players based exclusively on the x-coordinate of their location, while horizontal formation lines only uses the y-coordinate. We approach the identification of the formation lines through a complete-linkage clustering, where the distance between clusters is given by the distance between the two farther away points in each cluster. Formally, given a set of n player locations $P = \{p_1, ..., p_n\}$, and let $d(p, q)$ be the Euclidean distance between p and q, and $D(L_1, L_2)$ the distance between clusters L_1 and L_2, the set L of dynamic formation lines is conformed by the average locations of the player's belonging to the complete-linkage clustering of P in k partitions, such that for $L_1, L_2 \in L$ and $D(L_1, L_2) = \max_{p^{L_1}, p^{L_2}} d(p^{L_1}, p^{L_2})$. When $p_i = (x_i, y_i) = (x_i, 0)$ we call L the set of vertical dynamic formation lines, and when $p_i = (x_i, y_i) = (0, y_i)$ we call L the set of horizontal formation pressure lines.

Figure 4.10 presents an example of three groups of vertical and horizontal dynamic lines identified in a specific game situation. We can observe that, in this situation, the vertical lines clustered four players in the back, four in the middle, and two in the forward line. Typically, soccer coaches might refer to this grouping as a "4-4-2" formation, where the four players in the back are expected to be defenders, and the two players in the front

are expected to be forwards. However, in the changing dynamics of a soccer match, players can change position indistinctively, such that a player in the "left-back" defending position might appear close to the opponent's goal, aligned with the forwards. By identifying these groups of players at each frame, we can provide instantaneous contextual information to our models. On the other hand, the horizontal lines (specifically the first and third) provide a width boundary for the team's formation. The block formed by the first and third vertical and horizontal lines represents the "inside block" of an opponent's team, which coaches typically refer to as a valuable playing zone.

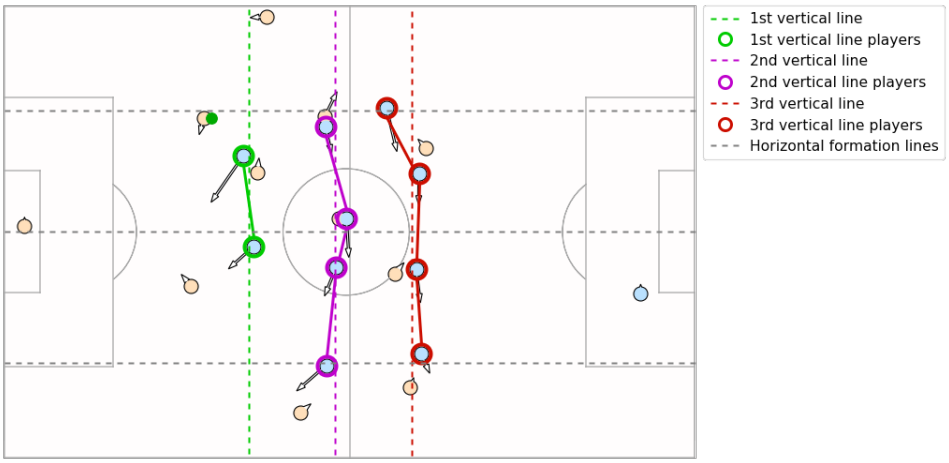


Figure 4.10: Three formation lines detected for the defending team (blue circles) in a match situation. The dotted lines show the average position of three vertical formation lines and two horizontal formation lines. The green, purple and red contour around the defending team players indicate the vertical formation line where each player was clustered in.

In this work, we set $k = 3$ to identify vertical formation lines, which conceptually represent forwards, midfielders, and defenders. For horizontal formation lines, we set $k = 3$, which will tend to define the breadth-wise borderlines of the team formation block and split the inside of the block into two parts. Note that while the optimal value of k could be learned for each situation, using a constant value allows for more straightforward interpretation from a practical perspective since coaches can directly understand what each of the three vertical and formation lines represents.

4.3.2 Baseline expected goals model

In Section 2.2.4, we describe the characteristics of existing xG models, representing an estimation of the probability of scoring goals from shots. The component of the EPV framework defined by the expression $\mathbb{E}[G|A = \varsigma]$, presented in Chapter 3, essentially represents an xG model. A subset of the presented spatial and contextual variables, derived from tracking data, are used in Chapter 6 to estimate this expression. However, a common limitation for building xG models from tracking data is the reduced amount of shot events available compared to larger and more easily available event data sources. To provide robustness to our model estimating $\mathbb{E}[G|A = \varsigma]$, we developed a baseline xG model from a larger dataset based on event data, presented in this section. This model's objective is to produce a baseline estimation of xG, which can be used as a strong prior for the tracking data-based model. This baseline estimation is also introduced as a feature in the action selection model presented in Chapter 6.

To produce a calibrated baseline estimation of xG, we use a wide dataset of event data provided by *OPTA*, which contains 117,948 shot events and 12,266 goals as detailed in Table 4.3.2. This dataset is considerably larger than the 13,735 shots available in the tracking data dataset used in Chapter 6. Event data has been used successfully in previous work to obtain a calibrated estimation of xG (Eggels, 2016).

Table 4.3: Total count of matches and shot events included within the event data dataset

Data Type	Source	# Total	# Training	# Test	% Goals
Match	Event	4,679	3,509	1,170	-
Shot	Event	117,948	87,980	30,645	10.4

We use a set of spatial features consisting of the event location and the distance and angle between the ball location and the goal. Contextual features are composed of a one-hot encoded vector indicating the attacking type at the moment of the event (i.e.,open-play, set-piece, free-kick, corner, penalty), and a boolean variable indicating whether the action is taken with the head or not. The matches are split into a training and test set, and model selection is performed through a K-fold cross-validation procedure on the training set, with $K = 10$. For every shot in the dataset, we label the outcome as 1 for the shots resulting in a goal, and 0 otherwise. We build the

model using the extreme gradient boosting algorithm XGBoost (Chen and Guestrin, 2016), and we perform an exhaustive grid-search on the following hyper-parameters of the model: number of trees ($\{50, 100, 250\}$), learning rate ($\{1e-3, 1e-2, 1e-1\}$), and maximum depth ($\{3, 5, 10\}$). All the features are standardized, obtaining a scaled feature set where each variable has a mean of 0 and a unitary standard deviation.

The best model presented a log loss value of 0.2540 and a calibration ECE value of 0.00594 in the test set. The parameters of the best model where: 100 trees, a maximum depth of 3, and a learning rate of $1e-1$. Figure 4.11 presents a calibration plot where the x-axis represents the average prediction in a set of 10 equally-sized bins, and the y-axis the average number of goals in the dataset for each bin. We can observe in this plot that the baseline xG model produces calibrated estimations. Some slight deviations from the optimal calibration line are observed due to small sample sizes in the later bins.

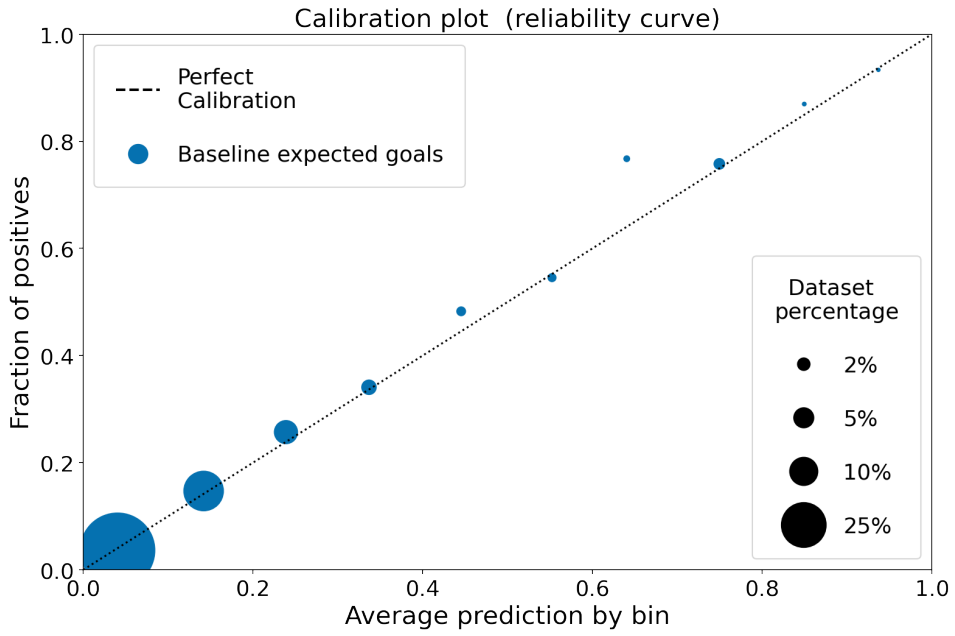


Figure 4.11: Calibration plot of the baseline xG model. Values in the x-axis represent the average prediction in a set of 10 equally-sized bins, while the y-axis represents the average number of goals observed for the examples of each bin. The circle size represents the percentage of examples contained in each bin with respect to the total.

4.3.3 Outplayed players

Two of the first tracking data-derived metrics to be popularized where the "packing" and "impect" (Schaper, 2021) metrics. Both metrics are related to the number of players surpassed by a pass or a ball drive. To calculate this metric in observed actions, the authors count the difference between the number of players in front of the ball when the action is attempted and the ball is received, respectively. When this count only includes player's from the defending team, the metric is called "packing", while when only players from the defending team considered to have the role of defenders (e.g., center-back or left-back), the metric is called "impect". In this work, we approach this idea more broadly and refer to the features derived from counting the difference of players between locations in time as outplayed players-features. Specifically, for the component estimating the expected value from passes, we calculate the number of opponents to be surpassed (or "outplayed") for every location of the field. We provide location-wise information indicating the pass's expected impact in terms of the number of opponents surpassed. We produce an analogous metric indicating the number of team players that would remain behind the ball since this would provide information relative to the expectation of conceding goals (e.g., when a pass back leaves a player as the last defender before the goal).

While the number of outplayed players might add some noise when including players that are far away from the ball and whose influence might be residual, this feature's inclusion is expected to add more information when combined with other contextual features. For example, in combination with the opponent's formation block location, we can obtain information about whether the pass is headed towards the inside or outside of the formation block and how many players are to be surpassed. Intuitively, a pass that outplays several players and that is headed towards the inside of the opponent block is more likely to produce an increase of the EPV than a pass back directed outside the opponent's block that adds two more opponent players in front of the ball.

4.4 Exploration of the developed features

In this Section, we provide a more profound exploration into the developed spatial and contextual features. We focus on how these features can separately provide rich information about spatiotemporal dynamics in soccer, such as the effect of pitch control in attempting a successful pass or ball

drive, the relationship between pass distance and location to the long-term expectation of scoring and receiving goals, and the value provided by the identification of dynamic pressure lines. We also provide a series of practical applications, showing how some of these features can allow gaining a deeper understanding of player behavior and performance according to context and space.

4.4.1 The long-term expected value from actions

The main focus of the proposed EPV framework is gaining a better understanding of the long-term expected value of actions in soccer. A significant addition of our approach compared to previous EPV approaches is to acknowledge that the outcome of actions is not only associated with the increase of of the probability of scoring goals, but these also impact the probability of conceding goals. In other words, our approach considers the risk and reward balance of decision-making. Here, we provide a broad exploration of how the long-term expected value of actions is distributed along the soccer field. Figure 4.12 presents a comparison of the long-term expected value of successful and missed pass and ball drives. Specifically, the left plot shows the average goals scored and conceded after a successful pass or ball drive was attempted from that location. The right plot is analogous, showing the average value at the destination location of missed passes and ball drives. The average goals are calculated in the following way: if a goal is scored by the team taking the action, at least 15 seconds after the action is attempted, the action's outcome is labeled 1. Otherwise, if a goal is conceded within the next 15 seconds, the outcome is labeled as -1. If no goal is observed, the action is labeled as 0. The average of those labels is calculated by each cell in a coarsened representation of a soccer field, consisting of 26×17 locations.

We can observe that the long-term value of successful actions correlates with the distance to the goal. The closest the action is to the opponent's box, and in particular, to the center of the goal, the higher the likelihood to observe a goal after 15 seconds. We can also note that the expected value of successful actions in the first third of the field tends to 0, while successful actions near the penalty point negatively affect the average. On the other hand, the average long-term outcome of missed actions presents a different behavior. Missed actions in the first half of the field tend to produce a negative outcome on average (i.e., conceding a goal), particularly when the ball is lost towards the center of the field and near the box. More interestingly, missed actions in the last third tend to produce a positive outcome on av-

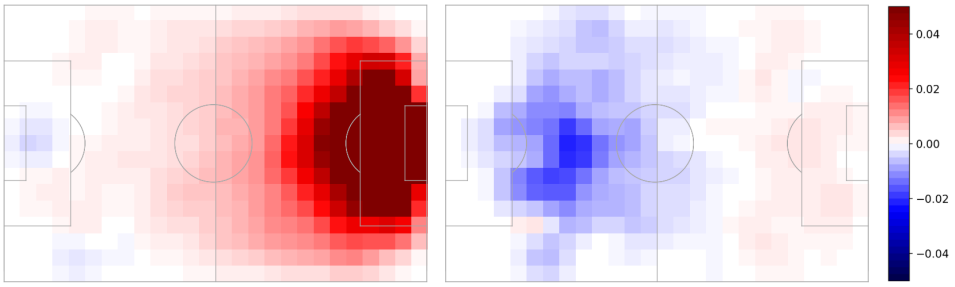


Figure 4.12: Comparison of the average goals scored and conceded within 15 seconds after a pass or ball drive is attempted, for the EPL seasons 13/14 and 14/15. The image on the left shows the average value at the origin location of successful actions, and the image on the right the average value at the destination location of missed actions

erage. This surprising scenario describes a well-understood soccer pattern: teams that take risks close to the box might lose the ball but then counter-press quickly to recover the ball and create scoring chances. This situation might lead to scoring chances with high probability because when a team recovers the ball, it gets wider to start attacking, and if they lose the ball while starting this defense-attack transition, they might be more disordered than when defending in an organized way. Another situation that explains this observation is the case of teams that play long-balls towards the last third: while the initial pass is often headed by an opponent, producing a missed pass, the attacking team will apply intense pressure nearby the ball to recover the ball quickly, and near the opponent's box.

These observations provide an idea of the importance of defining EPV taking into consideration that any action may produce an increase in the probability of either scoring or conceding a goal (as presented in Chapter 3), regardless of whether it is successful or not. Most of the previous EPV approaches in soccer and other sports constraint the long-term expected reward of actions into a $[0, 1]$ range, which difficult to make sense of the intricacies of risk and reward according to context.

4.4.2 Exploring success and long-term outcome of passes

Passes are the most frequent actions in soccer and constitute a fundamental element for understanding how teams and players succeed in this sport. This Section explores the relationship between pass distance and the success and long-term average reward from passes. First, we cluster all the passes available for the EPL seasons 13/14 and 14/15 into ten groups, according to

pass distance, using a K-means clustering procedure. Figure 4.13 presents the frequency and the average success of all the passes, grouped by distance. We can observe that the longer the pass distance, the lower the expected success probability of passes. Also, the frequency of attempted passes follows a similar distribution, where the majority of attempted passes have a distance lower than 30 meters. This Figure essentially shows that passes of short to medium distance (below 30 meters) present a considerably higher success probability (from 80% to 90%) than long passes (above 30 meters), which show to be more difficult. Note that 0 to 6 meters-long passes present a lower success probability than those between 6 and 20 meters. This shows an important characteristic of the observed data, where the vast majority intercepted passes near the pass taker are clustered in the $0-6$ group since the destination location label is kept as observed in the actual matches. While we could attempt predicting the expected location of passes, we considered this could impact the development and analysis of the intended EPV model and chose to leave the data as observed.

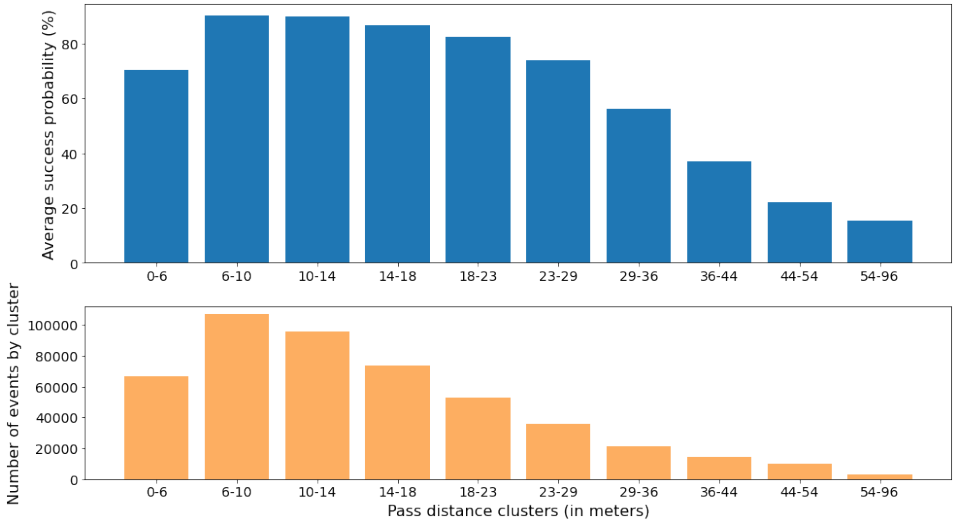


Figure 4.13: On top, a comparison of the average success of passes for the EPL seasons 13/14 and 14/15, clustered in ten groups according to pass distance. On the bottom, the frequency of passes by cluster

Another interesting behavior is the relationship between pass distance and the average long-term value, presented in Figure 4.14. At first glance, we can observe that long passes can provide a considerably more significant

likelihood of scoring a goal in the next 15 seconds when successful and a low likelihood of conceding a goal when missed. However, as seen before, the probability of success of such passes is relatively low. On the other hand, passes below 30 meters present a stable tendency, where a goal is scored after successful passes 0.75% of the time, while missed passes below this threshold can lead to a probability of conceding a goal, down to 0.25% of the time. Additionally, and following the discussion in Section 4.4.1, long-distance passes (such as the case from 29 to 36 meters in the available data) can even provide a higher probability of scoring than conceding when missed.

Figure 4.14: On top, a comparison of the average goals observed within 15 seconds for passes in the EPL seasons 13/14 and 14/15, clustered in ten groups according to pass distance. On the bottom, the frequency of passes by cluster

4.4.3 The effect of pressure in action success

One of the main contributions of the pitch control model presented in Section 4.2.3 is producing features that allow us to better understand the impact of spatial pressure on the likelihood of successfully attempting an action. In particular, the success of passes and ball drives is considered to be highly influenced by the level of pressure received by the player carrying the ball. When a player is marked closely by opponents, the likelihood of losing control of the ball or an attempted pass being intercepted is expected to increase considerably. This likelihood of losing the ball is also expected to increase

when the number of opponents pressing is high, and the distance to the ball carrier is low. In Figure 4.15 we present two plots showing the distribution of the pitch control of the ball carrier at the time a ball drive (left) or a pass (right) is attempted conditioned to the outcome of the action.

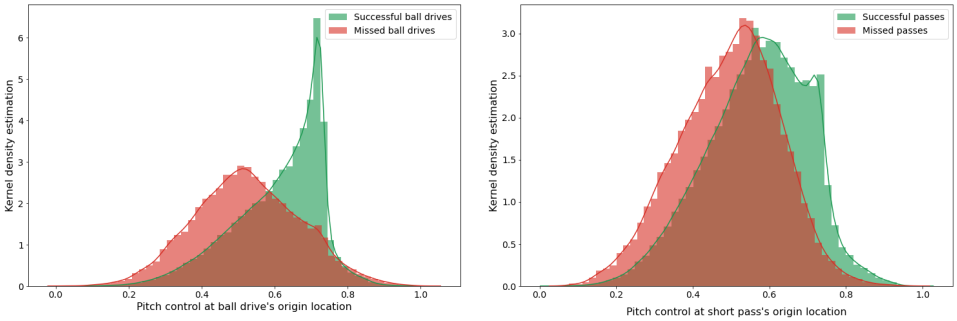


Figure 4.15: Comparison of the distribution of pitch control of the ball carrier for successful and unsuccessful actions at the time the action is taken. The plots show the distribution for ball drives (left) and passes (right)

We can observe that the effect of pitch control is considerably high in the likelihood of keeping control of the ball, where the higher the pitch control, the higher the probability of attempting a ball drive successfully. We can also observe that for lower values of pitch control (i.e., higher spatial control of the opponent team), the frequency of ball drives attempts is reduced considerably, and the probability of success drops to its lower values, possibly reflecting that players avoid ball drives when the spatial pressure is high. In the case of passes, a pitch control above 0.5 tends to favor attempting successful passes, while values below 0.5 decrease the probability of success. While the effect can be observed clearly, we can also see that the pitch control of the ball carrier alone is not sufficient to predicting pass success probability, where other factors such as the spatial control of destination zone, the relative location in the field, and the velocity of players is expected also to influence this probability.

4.4.4 Understanding context through dynamic pressure lines

The concept of dynamic lines presented in Section 4.3.1 allows one to contextualize the location and impact of actions according to the opponent's positioning at any given time. As explained in the mentioned section, two passes with the exact origin and destination location, but attempted in two different opponent formation configurations, may have a different meaning

and interpretation. This section presents three practical examples of how the frame-by-frame detection of dynamic pressure lines can allow one to develop a more fine-grained and rich interpretation of players' and teams' passing dynamics.

Detecting player's passing tendencies relative to the opponent

The origin and destination location of a player's attempted and received passes can provide information about the field zones that the player tends to exploit, either on-ball or off-ball. However, the large size of the soccer field and the broad set of possible defensive set-ups that teams can exhibit when defending provides the need for employing a richer set of features for understanding a player's passing tendencies beyond the location of his actions. Figure 4.16 compares the passes attempted and received by two FC Barcelona right-wingers, Messi and Dembélé, in a match against Athletic Club Bilbao in January of 2021. The Figure presents four plots by player. The first two plots correspond to a kernel density estimation of the origin location of the passes attempted by each player; while the first one uses the observed origin location of passes, the second one presents the locations relative two the opponent's pressure lines. The third and fourth plots present the total percentage of passes attempted (using origin location) and received (using destination location) by each player, respectively, summarized according to the opponent's pressure lines.

We can observe that the pass location changes considerably when we recalculate the locations relative to the opponent's line. The first plot shows us that Messi tended to play in a central position around the midfield, with a wide coverage spread. However, when observing the second plot, we can see that most of Messi's passes occurred between the opponent's second and third pressure line (a highly threatening attacking position) while the opponent presented an advanced block positioning. The third plot shows that Messi can frequently pass towards locations inside the opponent's block and behind the defender's back, exhibiting the tremendous associative capacity and generating danger of the player. The third plot shows a recent positioning behavior of Messi, where he tends to play lower than what is usual for a winger to receive between the first and second lines and participate more in the playmaking and ball circulation.

Dembelé, on the other hand, showed different positioning and passing tendencies. While the first heatmap might provide an idea that the player is

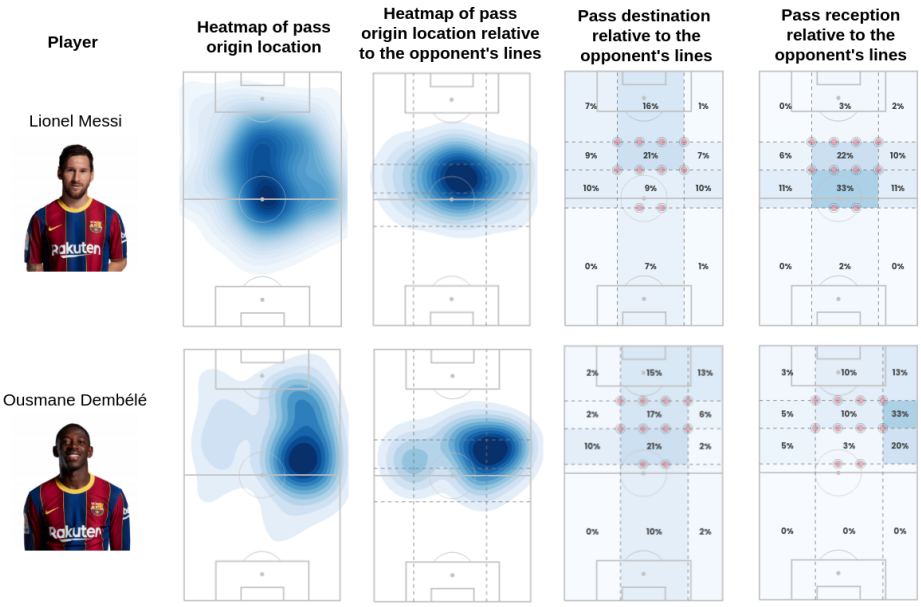


Figure 4.16: Comparison of two player's passing selection and reception dynamics in a single match, providing absolute and pressure lines-relative location of passes

positioned in a lower part of the field (while wingers tend to play in higher positions), the relative location heatmap shows the player is touching the ball frequently between the second and third line, by the outside of the opponent block. Like the case of Messi, with the relative lines, we can understand that the opponent is playing in a high block position, providing a very different read than the one represented in the original heatmap. The third and fourth plot shows that while the player tends to receive in the outside, he tends to pass the ball towards the inside of the formation block and can make successful passes between all the lines. Also, the plot shows that the player has been able to receive the ball beyond the defensive line, a critical performance element for wingers.

In this analysis, we have observed how the simple incorporation of relative positioning lines can provide a more rich interpretation of observed locations of passes and provide deeper inspection into a player's passing selection and reception dynamics. Employing this information, we can translate the vast amount of observed passes into more straightforward and holistic pieces of information for coaches.

The value of passes breaking lines

A concept frequently discussed by coaches is the ability to break lines, either through passing or driving. This idea refers directly to the capacity of making a pass or a ball drive that starts behind a pressure line and ends successfully above that line. While actions breaking lines are typically interpreted as valuable, there is no quantification of the value of these actions using tracking data. In order to better understand the value of passes breaking lines, we employ the dynamic pressure lines model to compare the average goals observed within 15 seconds after the formation line is broken through passes, differentiating between the inside and outside of the opponent's block. This comparison is presented in Figure 4.17, where the plot below presents the percentage of all passes that correspond to a series of passes breaking lines, and the plot above presents the mentioned long-term reward of each of these passes.

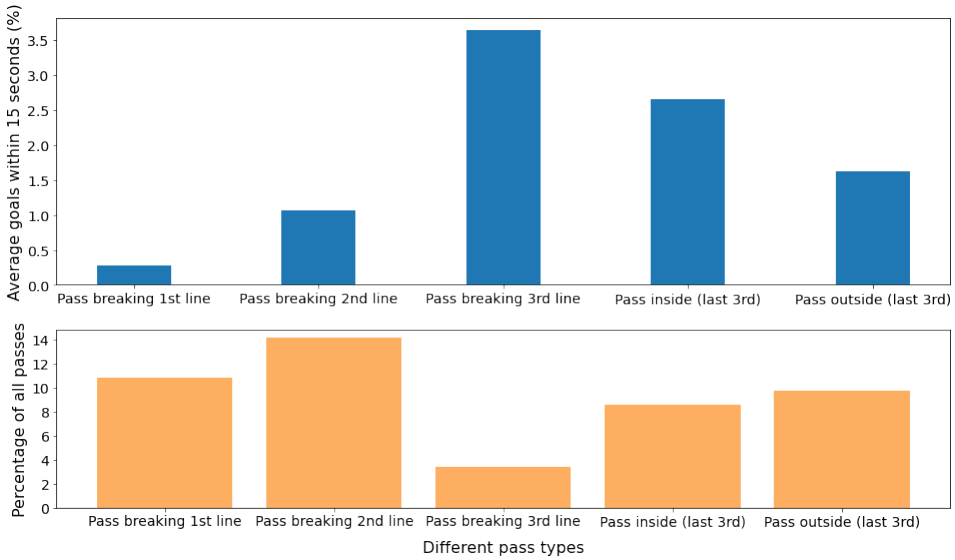


Figure 4.17: Comparison of the average goals observed within 15 seconds that after passes breaking lines, and passes into the last third of the field are successfully attempted

We can observe that all the passes breaking the second line and above present an average amount of goals within 15 seconds larger than the average of 0.75% observed in most of the passes presented earlier in Figure 4.13. Passes breaking the third line are shown to be valuable, providing an average probability of scoring goals of 3.5%. While we can observe that passes

attempted towards the last third (which might not break lines) provide a high expected long-term value, this is considerably lower than when the defenders' line is broken directly. Also, we can observe that passing the ball towards the inside of the opponent's block in the last third (as seen for Messi's passes in the last Section) has a larger expected value than passes towards the outside of the block. Naturally, passes breaking the first and second line are considerably more frequent that passes breaking the third line, which also provides an idea of the difficulty of finding single actions with high value in soccer.

Identifying game phases

During a soccer possession the dynamics of the game can change drastically depending on the location of the ball and the 22 players. In order to provide a structured approach for analyzing the development of possessions, coaches tend to group different parts of the possession into phases, according to contextual information. A typical approach for analyzing organized possessions in professional is defining three possible phases that possession might go through: buildup, creation, and finishing phases. The buildup is a phase where a possession starts to develop from one's half, typically with most of the opponent team behind the ball. Once the first pressure line or the midfield is reached, the possession is considered to reach a creation phase, where the objective is to keep progressing towards the opponent's goal. Then, when the possession reaches zones near the box, the possession enters a finishing stage, where the objective is to score. During the same possession, a team could return to a previous phase. While these stages can take many names and forms, the underlying idea is that possession goes through different stages where the contextual characteristics differ and are not straightforward to define. Also, for each stage the characteristics of on-ball and off-ball actions might vary, given that the objective of each phase varies as well.

Following this definition, we can employ the dynamic pressure lines to identify the different phases that possession goes through with high precision. Following the advice of the professional coaches that contributed to this work, we will say the possession enters a buildup phase when the ball is behind the first pressure line. After the first pressure line is overcome, we will say the possession enters a creation phase. This phase might end if the ball moves behind the first pressure line or if the ball is above the second pressure line and in the last third when we will say the possession

enters a finishing stage. While this simplified version can be improved by evaluating the time the ball spends between lines and some other movement dynamics, we consider this definition is sufficient to present the value of this contextualization.

To provide an idea of how the passing dynamics change between phases, we present in Figure 4.18 a comparison of FC Barcelona and Athletic Club de Bilbao's passing networks during their buildup and creation phases, in a match played in January 2021. We can observe that the players' distribution along the field changes considerably between the buildup and creation phase. While in the buildup phase, the teams' blocks tend to be longer and broader, in the creation phase, we observe a more compact block, with differences in the occupation of spaces. These changes in positioning directly impact the ball-passing dynamics and the players involved in each phase. We can expect wider spaces during the buildup phase and a higher risk of losing the ball given the closeness to the attacking team's goal. On the other hand, during the creation phase, the need for short passes, faster ball circulation, and in general, the ability to play in spaces with higher density and value increases considerably. These dynamics can also change drastically, depending on the type of pressure imposed by the opponent team, which can also be identified through the dynamic lines algorithm. For example, buildup phases against a low block (opponent players placed in their side field) propose a completely different game situation than a buildup phase where the opponent defends with a high block and intense pressure. By identifying the dynamic lines of both teams, we can provide a fundamental building block for deriving features that allows us to develop a richer representation of the spatiotemporal dynamics according to context.

4.5 Methodology of the collaboration with coaches

One of the fundamental goals of sports analytics is to ensure that research can be directly applied by sports practitioners in professional environments. Soccer coaches are the best example of a type of practitioner that has key decision-making responsibilities with a direct impact in team performance and development. Based on this idea, this thesis was aimed to be developed at the facilities of FC Barcelona, one the most renowned teams in the world, with the objective to work closely with professional soccer coaches, and ensure the applicability and practical relevance of the different models developed. In this Section we describe the nature of the collaboration with

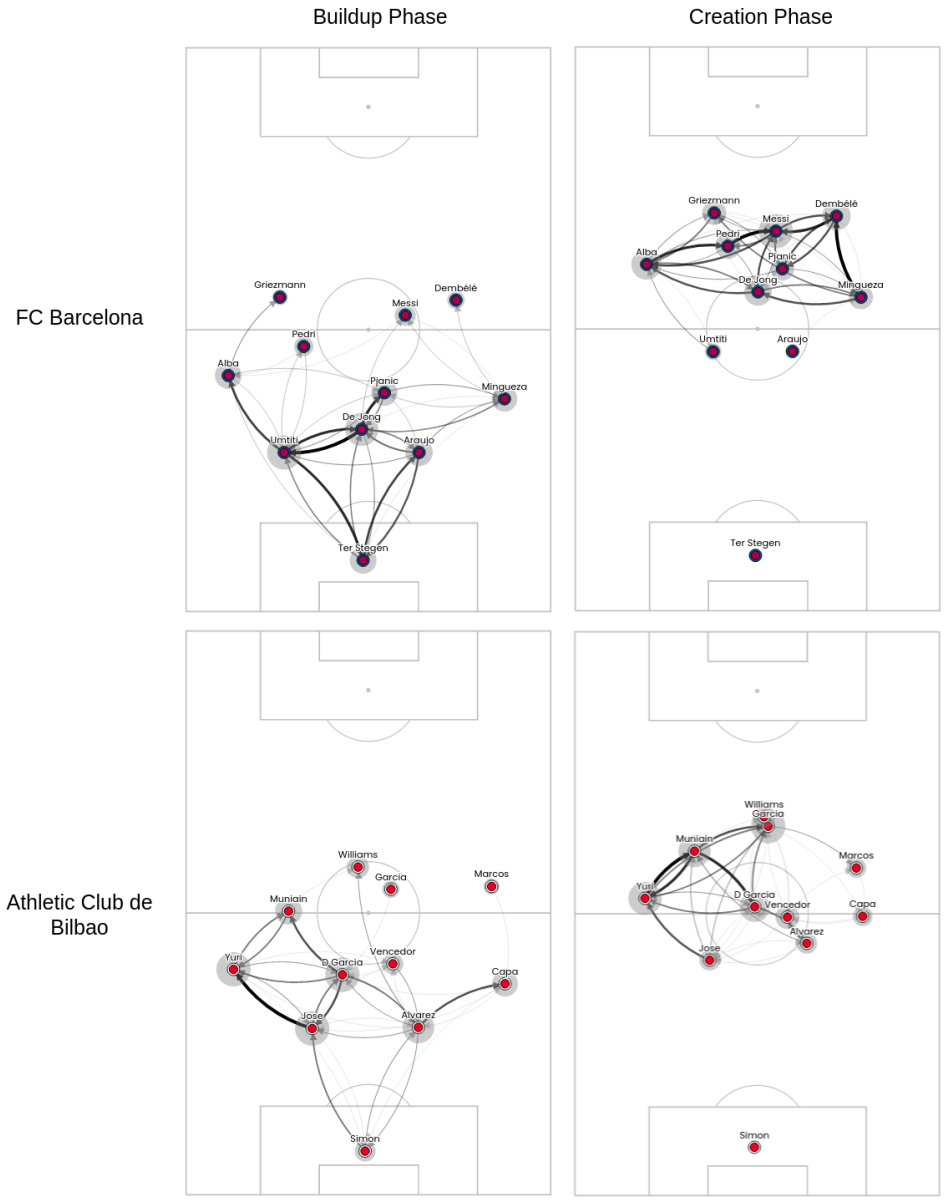


Figure 4.18: Comparison of the pass maps for FC Barcelona and Athletic Club de Bilbao, during the buildup and creation phases, in a match played in January 2021. Arrows size indicates the percentage from all the team passes, and the circles around players are proportional to the percentage of all the passes received by the team players. Players are placed in the average location of all their attempted passes

FC Barcelonaś coaches, and the collaboration methodology followed across the development of this thesis.

From the beginning of this work, the club identified the role of game analysts as the best type of coach to further this collaboration. Game analysts are professional soccer coaches whose worked is focused on providing rich tactical analysis of the performance of both teams and players. Most of their work is carried through the detailed analysis of match videos, and the manual tagging of the events observed during the match. Right now, almost every professional team in the European soccer clubs counts with at least one game analyst within their coaching staff. Specicially, we created a collaboration grouped conformed by game analysts of the most part of soccer teams at FC Barcelona. The game analysts most actively involved in this group where Raúl Peláez (first team), Javier Molina (second team), Guillem Escriu (under-19 team), and the cross-team analysts Xavier Pavo and Dídac Soler.

The first year of this collaboration involved meetings in a weekly basis with the objective of understanding the terminology and the building blocks of the analysts' daily work. At the same time, we introduced the types of data available in soccer (especially event and tracking data), and the main characteristics of the most common algorithmic approaches to develop spatiotemporal analysis from this data. During this process, we analyzed one-by-one most of the tags that analysts produce in each match. Starting from more simple concepts such as the definition of attack and defense, up to more complex ones such as the identification of game phases or the dynamic variation of defense pressure types, we designed a series of rule-based algorithms providing automated labeling of these concepts. The automatic labeling algorithms provided video clips that the group could watch together each other week to validate and refine the concepts. This first stage provided the pillar for allowing a faster collaboration with these professionals in the rest of the years of the thesis.

During the next years we started developing the different spatial and contextual features presented in this Section, as well as the different applied models presented in Chapters 5, 6 and 7. The feedback and observations of these coaches where critical for designing and fine-tuning the different models, especially given their keenness in identifying when the results of a model will be applicable and easily digestible within professional soccer staffs.

In order to accelerate and facilitate the joint analysis of different models, we designed a web-based infrastructure that integrates spatiotemporal data, the developed statistics, and match videos, into a single analysis tool. Figure 4.19 presents a snapshot of this web-based tool which become a fundamental piece for improving the understanding of the models developed during this work, as well as providing a ground-truth validation of the different findings. The infrastructure consists of three main elements: a data processing backend, a web services middleware, and frontend web presentation. For the data processing backend, we designed an algorithm development framework providing the groundwork for quickly developing new models and a seamless integration between different data sources. This framework currently constitutes the backbone for the development of algorithms and statistical models at the FC Barcelona's data analysis department.

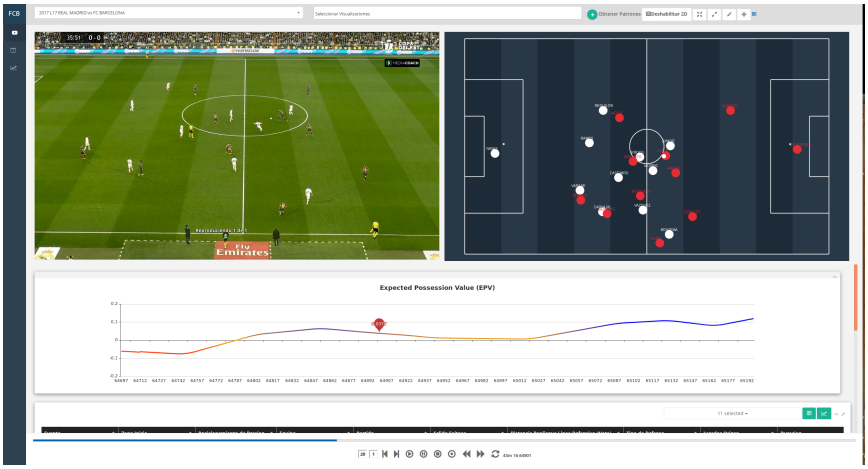


Figure 4.19: Web-based tool integrating spatiotemporal tracking data, calculated statistics and synchronized video footage.

Chapter 5

SoccerMap: learning probability surfaces from raw tracking data

The majority of existing research in soccer analytics has focused on analyzing the impact of either on-ball events, such as goals, shots, and passes, or the effects of players' movements and match dynamics (Gudmundsson and Horton, 2017). Most modeling approaches share one or more common issues, such as: heavy use of handcrafted features, little visual interpretability, and coarse representations that ignore meaningful spatial relationships. We still lack a comprehensive approach that can learn from lower-level input, exploit spatial relationships on any location, and provide accurate predictions of observed and unobserved events at any location on the field.

This chapter presents SoccerMap, a deep learning architecture that allows calculating full probability surfaces from low-level spatiotemporal tracking data. SoccerMap receives layers of low-level inputs and learns a feature hierarchy that produces predictions at different sampling levels, capturing both coarse and fine spatial details. By merging these predictions, we can produce visually-rich probability surfaces for any game situation that allows coaches to develop a fine-grained analysis of players' positioning and decision-making, an as-yet little-explored area in sports. We show how this architecture can be easily adapted to provide practical solutions for challenging problems such as the estimation of pass probability, pass selection likelihood, and pass expected value surfaces. We describe all the components of the SoccerMap architecture and show that it successfully solves

the challenging problem of learning a full prediction surface when there is only a single-pixel correspondence between ground-truth outcomes and the predicted probability map.

5.1 Methodology

In this section, we describe the characteristics of the SoccerMap architecture, including the general structure of the input data, the different feature extraction components, and the design considerations for learning a prediction surface from single-location outcome labels.

We seek an architecture that can produce a full probability surface from a given game state representation constructed from spatiotemporal tracking data. Specifically, the architecture must be able to learn both more refined features related to the influence of close locations and features considering information on a greater spatial scale. For a SoccerMap architecture trained to predict pass events' success probability, some examples of local features that the network might learn are the likelihood of nearby players reaching the destination location or information about local spatial pressure. On the other hand, higher scale features might consider the player's density and interceptability of the ball in its path from the location of origin.

In this work, we use the term "probability surface" in a broad way to refer to a matrix of values (generally in [0,1]), which complies with a size such that it can be extrapolated to a soccer field, unlike a single value estimate. From a design perspective, SoccerMap is designed to be applied as a spatial-aware feature extractor for estimating probability surfaces for a broad set of problems, instead of being tailored to solve a specific problem in soccer (e.g., predicting pass selection probability or predicting the expected value of on-ball events). Essentially, to build a network for estimating probability surfaces one is only required to define a matrix-like game state representation, provide the location and outcome of a given observed event, and define both an activation and a loss function that are appropriate for the distribution of the outcome variable. During training, the network weights are adjusted to learn the set of characteristics that best correspond to the specific problem being set up.

5.1.1 Defining SoccerMap formally

We seek a model that can produce accurate predictions of the probability of observing a given event for each location on a $l \times h$ coarse representation of a soccer field, given a matrix representation of the game state of size $l \times h \times c$. Based on this, we formally define SoccerMap as in Definition 5.1.1.

Definition 5.1.1 (SoccerMap)**.** *Let* $\{x|x \in \mathbb{R}^{l \times h \times c}\}$ *be the set of possible game state representations derived from a tracking data snapshot* T_t*, at time* t*, where* $l, h \in \mathbb{N}_1$ *are the height and length of a coarse representation of soccer field, and* $c \in \mathbb{N}_1$ *the number of data slices. A SoccerMap is a function* $f(x;\theta), f : \mathbb{R}^{l \times h \times c} \rightarrow \mathbb{R}^{l \times h}_{[0,1]}$*, where* f *produces a probability map, and* θ *are the function parameters.*

Here we identify three main elements that must be clearly defined for developing a SoccerMap model: the game state representation, the function parameters, and the learning approach. In the next sections, we describe each of these elements' characteristics in detail, where we propose a fully convolutional neural network architecture for learning the function parameters, which we refer to as the SoccerMap architecture.

5.1.2 Representing the game state

In order to capture the complex spatial relationships that are required for learning accurate probability surfaces in soccer, we want our model to make sense of spatiotemporal tracking data. In Section 3.1.1 we defined Ψ to be a high dimensional space representing all the spatiotemporal data available, and $T_t \in \Psi$ as a subset of that data at time t. We refer to T_t as a snapshot of spatiotemporal data. In the context of SoccerMap, we will employ spatiotemporal tracking data to construct snapshots representing the game state at any given time. Intuitively, the game state representation should comprise all the spatiotemporal information that is considered to influence the probability at any location. Given that we aim to learn a probability surface, it is convenient to structure the game state representation in a matrix form resembling a soccer field's dimensions (and the dimensions of the output surface). Following Definition 5.1.1 we will define the game state representation as a stack of c matrices of size $l \times h$, each representing a subset of the available spatiotemporal information. The specific choice of information for each of these c slices might vary depending on the problem being solved. In general, these slices might be constituted by either spatial or contextual information of any kind, including those defined in Chapter

4, as long as a value is defined for each cell of the matrix. In Section 5.2 we show that we can estimate pass probability surfaces by using a game representation constituted by low-level information slices including the locations of the players of each team, the magnitude of velocity of the players in each direction of a 2D plane, the angle between players and the ball, and the distance to the ball and the defending team goal location. In Section 5.2 we show that, although most of these slices constitute sparse matrices, SoccerMap can make sense of this information on both local and global scales. In Chapter 6 we present different types of game representations that are specifically adapted for estimating the pass probability, pass selection, and pass EPV components of the decomposed EPV model through the SoccerMap architecture.

5.1.3 SoccerMap architecture design

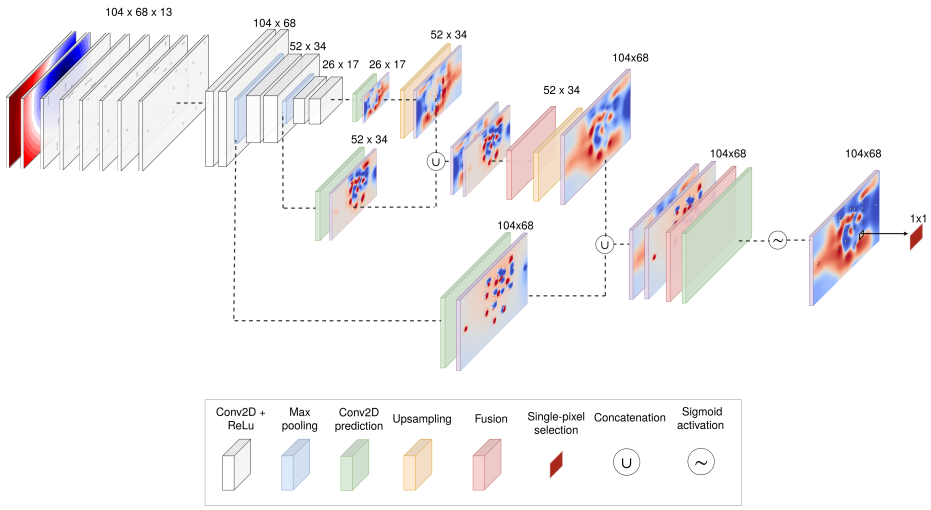


Figure 5.1: SoccerMap architecture with an input game state representation of 104×68 and 13 input channels, trained for predicting pass probability surfaces.

The SoccerMap architecture design is approached as a fully convolutional neural network. Figure 5.1 presents a visual representation of the main components of the proposed architecture, trained to predict pass probability surfaces and receiving an input data snapshot of $104 \times 68 \times 13$. An input game state representation is processed by the deep neural network that creates a feature hierarchy by learning convolutions at $1x$, $1/2x$, and $1/4x$ scales while preserving the filters' receptive fields. Predictions are produced

at each of these scales and then upsampled nonlinearly and merged through fusion layers. An activation function is applied to the latest prediction layer to produce a single probability estimations at every location and preserving the original input scale ($l \times h$). During training, a single-location prediction associated with the destination of a sample event is selected to compute a selected loss function that is backpropagated to adjust the network weights.

5.1.4 The reasoning behind the choice of layers

The diagram in Figure 5.2 presents a detailed representation of all the components of a standard SoccerMap architecture for a soccer field representation of sizes 104×68, and a variable number of layers c.

The network incorporates different layers: max-pooling, linear, ReLu and linear activation layers, and 2-dimensional convolutional filters (conv2d) for feature extraction, prediction, upsampling, and fusion. In this section, we present a detailed explanation of the reasoning behind the choice of these layers and the architecture design.

Convolutions for feature extraction At each of the $1x$, $1/2x$, and $1/4x$ scales, two layers of 2D convolutional filters with a 5×5 receptive field and stride of 1 are applied, each one followed by a ReLu activation function layer, in order to extract spatial features at every scale. To keep the same dimensions after the convolutional filters, we apply symmetric padding to the input matrix of the convolutional layer. We chose symmetric-padding to avoid border-image artifacts that can hinder the model's predicting ability and visual representation (Odena et al., 2016).

Fully convolutional network There are several conceptual and practical reasons for considering convnets for this problem. Convolutional filters are designed to recognize the relationships between nearby pixels, producing features that are spatially aware. convnets have been proven successful in data sources with a Euclidean structure, such as images and videos, so a 2D-mapping of soccer field location-based information can be expected to be an ideal data structure for learning essential features. Also, these features are expected to be non-trivial and complex. convnets have been proven to learn what are sometimes more powerful visual features than handcrafted ones, even given large receptive fields and weak label training (Long et al., 2014). Regarding the architecture design, we are interested in learning the full $l \times h$ mapping of probabilities covering the extent of a soccer field, for which fully convolutional layers are more appropriate than standard neural

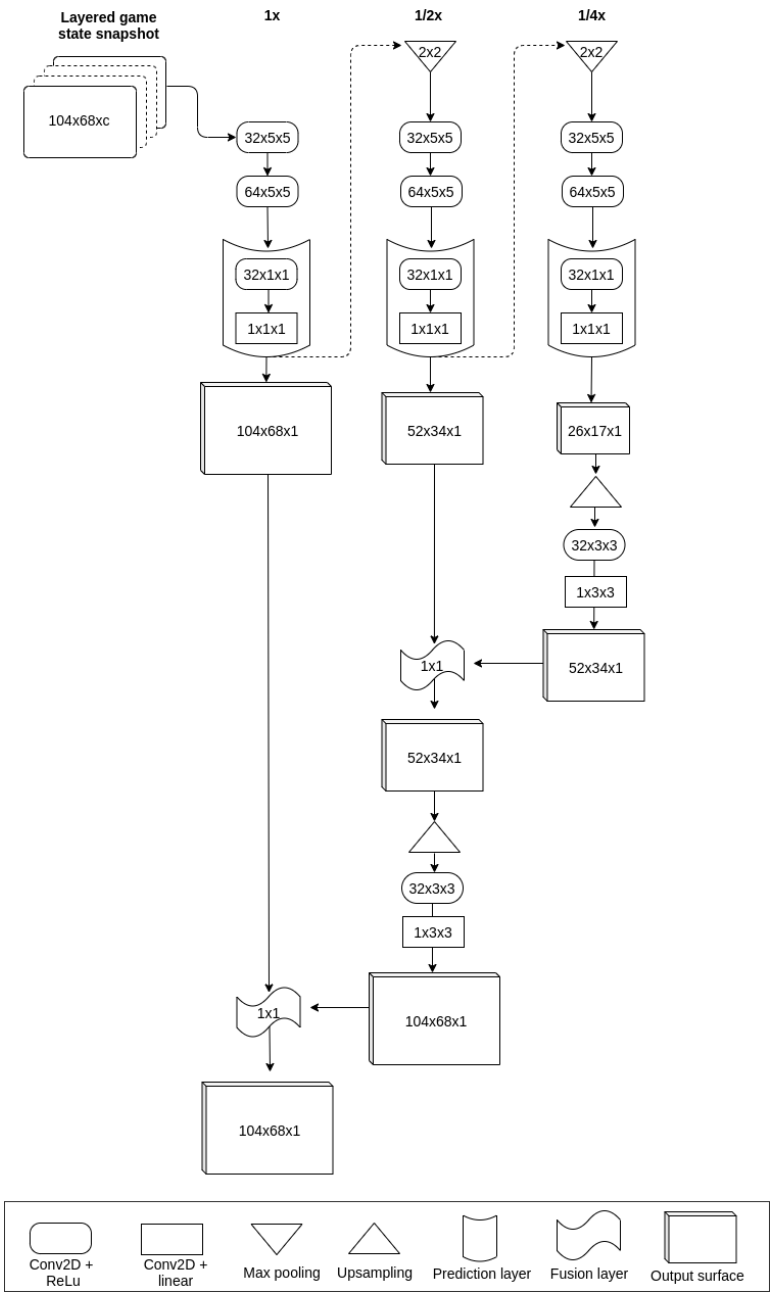


Figure 5.2: Components of the SoccerMap architecture. A layered input of a game state snapshot is fed to a network that produces prediction surfaces at 1x, 1/2x, and 1/4x sampling scales to capture both local and global features. Outputs at different sampling rates are merged and upsampled to produce a single prediction surface.

networks built for classification by changing dense prediction layers for 1x1 convolution layers.

Pooling and upsampling The network applies downsampling twice through max-pooling layers to obtain the $1/2x$ and $1/4x$ representations. Since the activation field size is kept constant after every downsampling step, the network can learn filters of a wider spatial extent, leading to the detection of coarse details. We learn non-linear upsampling functions at every upsampling step by first applying a $2x$ nearest neighbor upsampling and then two layers of convolutional filters. The first convolutional layer consists of 32 filters with a 3×3 activation field and stride 1, followed by a ReLu activation layer. The second layer consists of 1 layer with a 3×3 activation field and stride 1, followed by a linear activation layer. This upsampling strategy has been shown to provide smoother outputs and to avoid artifacts that can usually be found in the application of transposed convolutions (Odena et al., 2016).

Prediction and fusion layers Prediction layers consist of a stack of two convolutional layers, the first with 32 1×1 convolutional filters followed by an ReLu activation layer, and the second consists of one 1×1 convolutional filter followed by a linear activation layer. Instead of reducing the output to a single prediction value, we keep the spatial dimensions at each step and use 1×1 convolutions to produce predictions at each location. The stack learns a non-linear prediction on top of the output of convolutional layers. To merge the outputs at different scales, we concatenate the pair of matrices and pass them through a convolutional layer of one 1×1 filter.

5.1.5 Learning from single-location labels

From a practical standpoint, training a SoccerMap architecture has the intrinsic difficulty that ground-truth data of full probability surfaces are not available in most problems where it could be applied. In contrast, we usually only have an observed binary value at a single location. For example, for the pass probability problem, ground-truth data only provides the observed binary outcome (i.e. successful or unsuccessful) of the pass at the destination location. For more challenging problems such as estimating the expected possession value of passes, the single-value outcome is usually observed in a much more distant time than the moment of attempting the pass.

Considering this, we approach the model training as a weakly-supervised learning task, where the ground truth labels only correspond to a single location in the full mapping matrix that needs to be learned. The target-location loss presented in Definition 5.1.2 essentially shrinks the output of a SoccerMap f to a single value by selecting the prediction at the destination location of the event and then computes the loss between this single prediction and the ground-truth outcome.

Definition 5.1.2 (Target-Location Loss)**.** *Let y_k be the observed outcome of an event at time t, for a game state x_k, d_k the destination location of the event k, f a SoccerMap with parameters θ, and L a loss function, we define the target-location loss as*

$$L(f(x_k;\theta), y_k) = L(f(x_k;\theta)^{d_k}, y_k)$$

5.2 Experiments and results

In this section, we approach estimating pass probability surfaces representing the success probability of attempting a pass to any location of the field, given a game situation. We first present the characteristics of the dataset used to train a SoccerMap architecture and two other benchmark models used for performance comparison. Then, we describe the game state representations used in each of the models, and the experimental framework applied. Finally, we present the logistic loss, calibration, and inference time for each of the models. Additionally, we present an ablation study performed on the SoccerMap architecture trained for this problem, where we assess the impact of each of the main components of the architecture.

5.2.1 Dataset

We use tracking data, and event data from 633 EPL matches from the 2013/2014 and 2014/2015 seasons, provided by *STATS LLC*. Each match contains the 2D location for every player and the ball sampled at 10Hz. The event data provides the origin location, destination location, time, player, team, and outcome for 480,670 passes. From this data, we extract the tracking data snapshot described in Section 5.2.2 to produce a coarse $(104, 68)$ representation of a soccer field, and obtaining a dataset of size $480,670 \times 104 \times 68 \times 13$. There are 382,806 successful passes and 97,864 missed passes.

5.2.2 A game state representation for estimating pass probability

We construct a game state representation for the SoccerMap model constituted by matrices of size $104 \times 68 \times 13$. The dimensions 104×68 correspond to the maximum length and width allows for a soccer field in professional soccer [1]. Following these dimensions, each cell approximately represents $1m^2$ of the field. The snapshot of tracking data is constituted by 13 slices of low-level spatiotemporal information at the time a pass is attempted. The game-state representation is composed of the following slices:

- Six sparse matrices with the location and the two components of the velocity vector for the players in both the attacking and defending teams.
- Two dense matrices where every location contains the distance to the ball and goal location.
- Two dense matrices containing the sine and cosine of the angle between every location to the goal and the ball location, and one dense matrix containing the angle in radians to the goal location.
- Two sparse matrices containing the sine and cosine of the angle between the velocity vector of the ball carrier and each of the teammates in the attacking team.

The tracking data is normalized left to right, where the field's rightmost location is where the team taking the pass scores goals. Since we are using a discrete field representation for learning probability surfaces from convolutional neural networks, this normalization becomes convenient for the network to differentiate between the cells in the matrix that are closer to where goals are scored or conceded, which is expected to influence the risk associated to the pass.

5.2.3 Benchmark models

We compare our results against a series of benchmark models of increasing levels of complexity. We define a baseline model *Naive* that for every pass, outputs the known average pass completion in the entire dataset (80%)

[1] A usual length of soccer fields is 105; however it is convenient to use 104, which is an even number and can be divided by 2 and 4 (following the 1/2x and 1/4x downsampling scales).

following a similar definition in Power et al. (2017). We build two additional models *Logistic Net* and *Dense2 Net* based on a set of handcrafted features built on top of tracking data. Logistic Net is a neural network with one hidden layer, followed by a linear activation layer and a single sigmoid output unit. Dense2 Net is a fully connected neural network with two hidden layers, followed by ReLu activations and a sigmoid output unit.

Handcrafted features We employ several of the spatial features presented in Chapter 4 for building the dataset to be used for training the benchmark models. Specifically, for all the available passes, we include origin and destination location, pass distance, angle to goal at origin and destination, and the pitch influence of the attacking and defending teams at both the origin and destination location.

5.2.4 Experimental framework

This section describes the experimental framework followed for testing the performance of the proposed architecture for the pass success probability estimation problem.

Training, validation, and test set We randomly sample the available matches and split them into a training, validation, and test set with a $60:20:20$ distribution. The events in the training dataset are randomly shuffled to avoid introducing bias related to the closeness of the occurrence of these events in time.The validation set is used for model selection during a grid-search process. The test set is left as hold-out data, and results are reported on performance for this dataset. For the benchmark models, datasets are built by extracting the features described in Section 5.2.3, and an identical split is performed. Features are standardized column-wise by subtracting the mean value and dividing by the standard deviation.

Optimization Both the SoccerMap network and the baseline models are trained using the ADAM optimization. Model selection is achieved through grid-search on learning rates of 10^{-3}, 10^{-4} and 10^{-5}, and batch sizes of 1, 16 and 32, while β_1, β_2 are set to 0.9 and 0.999, respectively. We use early stopping with a minimum delta rate of 0.001. Optimization is computed on a single Tesla M60 graphical processing unit (GPU) and using Tensorflow 1.5.0. During the optimization, the negative log-loss is minimized.

Metrics For each of the models we report the log-loss, following the definition presented in Equation 2.5. We validate the model's calibration using a variation of the ECE presented in Section 2.5.1. For obtaining this metric, we distribute the predicted outcomes into K bins and compute the difference between the average prediction in each bin and the average expected outcome for the examples in each bin. Equation 5.1 presents this variation of the ECE metric, where K is the number of bins, and B_k corresponds to the set of examples in the k-th bin. Essentially, we are calculating the average difference between predicted and expected outcomes, weighted by the number of examples in each bin. In these experiments, we use uniform binning to obtain K bins in ascending order. A perfectly calibrated model would have an ECE value of 0. Additionally, we provide a calibration reliability plot showing the mean confidence for every bin B_k. Note that Equation 5.1 extends directly from the more global definition of ECE presented earlier in Section 2.5.1 and Equation 2.11.

$$\text{ECE} = \sum_{k=1}^{K} \frac{|B_k|}{N} \left| \left(\frac{1}{|B_k|} \sum_{i \in B_k} y_i \right) - \left(\frac{1}{|B_k|} \sum_{i \in B_k} \hat{y}_i \right) \right| \tag{5.1}$$

5.2.5 Results

Table 5.1 presents the results for the benchmark models and SoccerMap on the pass probability dataset. We can observe that SoccerMap achieves a considerably lower error than the other models. This result is remarkable considering the network uses exclusively low-level spatiotemporal data and makes sense of the full spatial extent of the field. Despite the large number of parameters in SoccerMap, the number of examples per second is near a thousand, being high enough to ensure a real-time estimation for frame rates below 200Hz (i.e., twenty times higher than the available frame rate). Figure 5.3 presents a calibration reliability plot for each of the models. We can see that both the models using handcrafted features and SoccerMap can produce well-calibrated estimations of pass probabilities. Additionally, since there is no strong deviation in any of the ranges of predicted probabilities in relation to observed data, a post-hoc calibration procedure could be applied to fine-tune slight miss calibrations in any of the models. While both the visual evaluation of the calibration and the ECE metric show a good calibration of the models, the considerably lower log-loss reveals that SoccerMap tends to produce predictions that are closer to the observed values than those predictions produced by the benchmark models.

Table 5.1: Results for the benchmark models and SoccerMap on the pass probability dataset

Model	log-loss	ECE	Ex. (s)	Number of parameters
Naive	0.5151	–	–	0
Logistic Net	0.2833	0.0087	22,982	12
Dense2 Net	0.2178	0.0043	22,600	276
SoccerMap	**0.1842**	0.0172	953	401,259

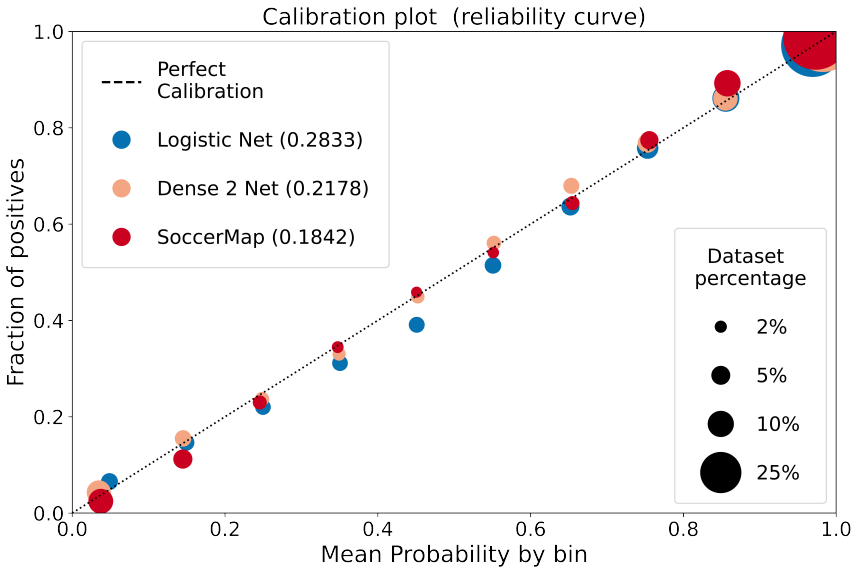


Figure 5.3: A calibration reliability plot, where the X-axis presents the mean predicted value for samples in each of 10 bins, and the Y-axis the fraction of samples in each bin containing positive examples.

Figure 5.4 presents the predicted pass probability surface for a specific game situation during a professional soccer match. We observe that the model can capture both fine-grained information, such as the influence of defending and attacking players on nearby locations and coarse information, such as the probability of reaching more extensive spatial areas depending on the distance to the ball and the proximity of players. We can also observe that the model considers the player's speed for predicting probabilities of passing to not-yet-occupied spaces, a critical aspect of practical soccer analysis.

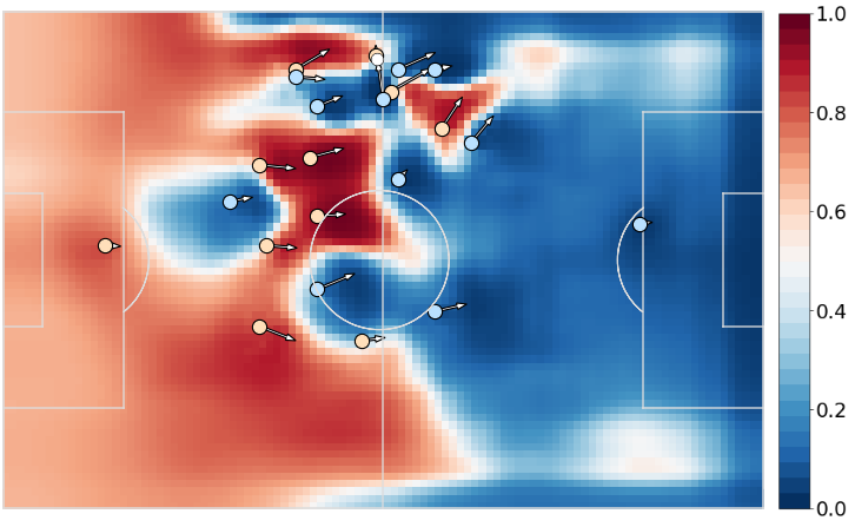


Figure 5.4: Pass probability surface for a given game situation. Yellow and blue circles represent players' locations on the attacking and defending team, respectively, and the arrows represent the velocity vector for each player. The white circle represents the ball location.

Ablation Study

We performed an ablation study to evaluate whether the different components of the proposed architecture allow improving its performance on the passing probability estimation problem or not by testing the performance of different architecture variations. Specifically, the variations of the SoccerMap architecture are created by removing one component at a time from the following set of components: nonlinear upsampling (UP), fusion layer (FL), nonlinear prediction layer (NLP), and skip-connections (SC). For the case of the fusion layer, we substitute the convolutional layers with direct pointwise addition of the predictions at different sampling levels. The removal of the nonlinear prediction layer consists in substituting the ReLu activations by linear activations at the prediction level. When skip-connections are removed, the architecture is modified considerably. For this case, we avoid making predictions at different sampling levels and keep a single sampling level (i.e., the original 104×68 input size). We provide two variations: CNN-D4 and CNN-D8, consisting of a feedforward network with 4 and 8 layers of 32 convolutional filters of size 5×5, respectively, each one followed by layer of ReLu activation functions. Finally, a linear prediction layer is applied to obtain the final predictions.

For each of the configurations, we train the architecture three times, with a fixed learning rate of 10^{-4}, a batch size of 32, a maximum of 20 epochs, and equivalent optimization criteria as presented in Section 5.2.4. Table 5.2 presents the average log-loss obtained on different configurations of the architecture, among the different runs. We can observe that the SoccerMap-based architectures outperform the rest of the configurations that do not use skip-connections. This highlights the importance of producing predictions at different sampling levels to capture relevant relationships at different scales. The full SoccerMap architecture presented the best average performance among the architectures. The nonlinearity at the prediction layers and fusion layers provided a slight performance improvement, provided the higher average loss obtained in the configuration where these are removed. The removal of the nonlinear upsampling provided a worst performance than the other architectures. While with a higher number of iterations, these average performances could improve, a critical aspect of this layer is the ability to produce smoother surfaces. From a practical standpoint, smoother surfaces are more visually appealing for practitioners, easing the communication and translation of the results into practice

Table 5.2: Ablation study for subsets of components of the SoccerMap architecture.

Architecture	SC	UP	FL	NLP	log-loss
SoccerMap	✓	✓	✓	✓	0.1859
SoccerMap-NLP	✓	✓	✓		0.1875
SoccerMap-FL	✓	✓		✓	0.1870
SoccerMap-UP	✓		✓	✓	0.1984
CNN-D4					0.2045
CNN-D8					0.2007

5.3 Related work

From an applied standpoint, our work is related to several other approaches aimed at estimating pass probabilities and other performance metrics derived from spatiotemporal data in soccer, described in Section 2.2.2. While some of the related work has estimated probability surfaces by inference on a set of discrete pass destination locations (Spearman et al., 2017), none

has yet approached the learning of probability surfaces directly. The related problem of pass selection has been approached by applying convolutional neural networks that predict the likelihood of passing to a specific player on the attacking team (Hubáček et al., 2018).

Regarding the technical approach, we leverage recent findings on the application of fully convolutional neural networks for image segmentation. Fully convolutional networks have been extensively applied to semantic image segmentation, specifically for the pixel-labeling problem to detect broad pixel areas associated with objects in images successfully. The approach most related to our work builds a hierarchy of features at different sampling levels that are merged to provide segmentation regions that preserve both fine and coarse details (Long et al., 2015). From a learning perspective, image segmentation has been approached as either supervised (Long et al., 2015), weakly supervised (Pathak et al., 2015), and semi-supervised learning problems (Papandreou et al., 2015). Usually, the available labels are associated with many other pixels in the original image. However, in our case, labels are only associated with a single location in the desired probability map (the destination location of the event), transforming the estimation of a full probability surface into a challenging prediction problem.

5.4 Discussion

The estimation of full probability surfaces provides a new dimension for soccer analytics. The presented architecture allows generating visual tools to help coaches perform fine-tuned analysis of opponents and own-team performance derived from low-level spatiotemporal soccer data. We show how this network can perform remarkably well at estimating the probability of observed passes. By merging features extracted at different sampling levels, the network can extract both fine and coarse details, thereby managing to make sense of soccer's complex spatial dynamics. Chapter 6 shows how the SoccerMap can be trained to learn to estimate two related but different problems, the pass selection probability and the expected possession value from passes. We show these models also produce accurate and calibrated estimations. In Chapter 7 we present several novels practical applications on soccer analytics derived from these SoccerMap-based models, such as evaluating optimal passing, evaluating optimal positioning, and identifying context-specific and team-level passing tendencies. This analysis framework derived from spatiotemporal data could also be applied directly in many

other team sports, where the visual representation of complex information can bring the coach and the data analyst closer.

Chapter 6

Estimating the EPV components

In Chapter 3 we presented a theoretical framework for modeling EPV, where this concept is approached as an estimation of the long-term reward of a possession, given all the spatiotemporal data available at a given time. More specifically, Section 3.2 presents the idea of decomposing the general EPV expression into a series of terms modeling the influence of three types of on-ball actions: passes, ball drives and shots. This decomposition allows us to split out a complex model into more easily understandable parts so the practitioner can both understand the factors that produce the final estimate and evaluate the effect that other possible actions may have had.

In this chapter we employ spatiotemporal tracking data and event data from professional soccer matches to develop a series of models for estimating the components of the EPV, and also producing a single instantaneous estimate of the EPV for any time instance. A visual representation of the output of this approach is presented in Figure 3.1. We propose two different approaches to learn each of the separated models, depending on whether we need to estimate a field-wide probability surface or producing only a single-valued prediction. We adapt the SoccerMap architecture presented in Chapter 5 to produce full prediction surfaces from low-level features, for the passing related components. Specifically, we are able to learn the surfaces for the pass selection, pass probability and pass EPV problems, from very challenging learning set-ups where only a single-location ground-truth is available. Producing these surfaces allows the components related to passes to estimate either the expected value or the probability of attempting a pass

to any other location on the field. On the other hand, for the components related to ball drive or shot actions, we use shallow neural networks on top of a broad set of novel spatial and contextual features, to produce single estimations of the expected value and the success probability of these actions.

This chapter is structured in the following way. We first provide the implementation details for estimating each of the components of the decomposed approach presented in Section 3.2. Then, we present the experimental setup for estimating these components from a large dataset of spatiotemporal tracking data of professional soccer matches, where we show this approach can produce calibrated estimations for each of the components and the final EPV estimate. With this approach we present a comprehensive analysis framework that allows us to develop a wide variety of practical applications in soccer, which includes both on-ball and off-ball performance analysis. Several of these applications are presented in Chapter 7.

6.1 Separate component inference

In this section we describe the approaches followed for estimating each of the components described in Equations 3.2,3.3, and 3.4. In general, we use function approximation methods to learn models for these components from spatiotemporal data. Specifically, we want to approximate some function f^* that maps a set of features x, to an outcome y, such that $y = f^*(x)$. To do this, we will find the mapping $y = f(x;\theta)$ to learn the values of a set of parameters θ that result in an approximation to f^*.

Customized convolutional neural network architectures are used for estimating probability surfaces for the components involving passes, such as pass success probability, the expected possession value of passes, and the field-wide pass selection surface. Standard shallow neural networks are used to estimate ball drive probability, expected possession value from ball drives and shots, and the action selection probability components. In this section we focus on describing the modeling approach for each component, providing details about the selection of features x, observed value y, and model parameters θ. The implementation details regarding the experiments carried out for building these models are explained in Section 6.2. All the spatial and contextual features referred in this section are explained in more detail in Chapter 4, while the detailed list of features used for each model is described in Appendix A.

6.1.1 Estimating pass impact at every location on the field

One of the most significant challenges when modeling passes in soccer is that, in practice, passes can go anywhere on the field. Previous attempts on quantifying pass success probability and expected value from passes in both soccer and basketball assume that the passing options a given player has are limited to the number of teammates on the field, and centered at their location at the time of the pass (Power et al., 2017; Cervone et al., 2016b; Hubáček et al., 2018). However, in order to accurately estimate the impact of passes in soccer (a key element for estimating the future pathways of a possession), we need to be able to make sense of the spatial and contextual information that influences the selection, accuracy, and potential risk and reward of passing to any other location on the field. For doing so, we employ the SoccerMap fully convolutional neural network architecture (see Chapter 5), which is specifically designed to exploit spatiotemporal information at different scales. We extend and adapt this architecture to the three related passing action models we require to learn: pass success probability, pass selection probability and pass expected value.

While these three problems necessitate from different design considerations, we structure the proposed models into four main conceptual blocks: the *game state representation*, a *feature extraction block*, a *surface prediction block*, and a *loss computation block*. For each of these models we detail the selected configuration on each of these conceptual blocks. The game state representation refers to the coarse matrix representation of size $l \times h \times c$, required by the SoccerMap architecture, and described in Section 5.1.2. For all of these problems, the feature extraction block is represented by the SoccerMap fully convolutional network architecture, and uses the parameters presented in Section 5.1.3. The surface prediction block refers to the activation function used at the end of the architecture, and that produces the predicted values for each location and sampling level. For all these three problems, the loss computation block employs the target-location loss presented in Definition 5.1.2, but uses different loss functions. In the following sections, we describe the design characteristics for the three pass-related problems: pass success probability, pass selection probability, and expected value from passes. By joining these models' output, we will obtain a single action-value estimation (EPV) for passing actions, expressed by $\mathbb{E}[G|A = \rho, T_t]$.

6.1.2 Pass success probability

From any given game situation where a player controls the ball, we desire to estimate the success probability of a pass attempted towards any of the other potential destination locations, expressed by $\mathbb{P}(A = \rho, D_t|T_t)$. Figure 6.1 presents the designed architecture for this problem. The input data at time t is conformed by 13 layers of spatiotemporal information obtained from the tracking data snapshot T_t consisting mainly of information regarding the location, velocity, distance, and angles between the both team's players and the goal. The feature extraction block is composed strictly by the SoccerMap architecture, where representative features are learned. This block's output consists of a $104 \times 68 \times 1$ pass probability predictions, one for each possible destination location in the coarsened field representation. In the surface prediction block a sigmoid activation function, defined in Equation 2.8, is applied to each prediction input to produce a matrix of pass probability estimations in the [0,1] continuous range. Finally, at the loss computation block, we select the probability output at the known destination location of observed passes and compute the negative log-loss, defined in Equation 2.5, between the predicted ($\hat{y}$) and observed pass outcome (y).

Note that we are learning all the network parameters θ needed to produce a full surface prediction by the back-propagation of the loss value between the predicted value at that location and the observed outcome of pass success at a single location. We show in Section 6.2.6 that this learning set up is sufficient to obtain remarkable results.

6.1.3 Expected possession value from passes

Once we have a pass success probability model, we are halfway to obtaining an estimation for $\mathbb{E}[G|A = \rho, D_t, T_t]$, as expressed in Equation 3.3. The remaining two components, $\mathbb{E}[G|A = \rho, O_p = 1, Dt, Tt]$ and $\mathbb{E}[G|A = \rho, O_p = 0, Dt, Tt]$, correspond to the expected value of successful and unsuccessful passes, respectively. We learn a model for each expression separately; however, we use an equivalent architecture for both cases. The main difference is that one model must be learned with successful passes and the other with missed passes exclusively to obtain full surface predictions for both cases.

The input data matrix consists of 16 different layers with equivalent location, velocity, distance, and angular information to those selected for the pass success probability model. Additionally, we append a series of layers corresponding to contextual features related to outplayed players'

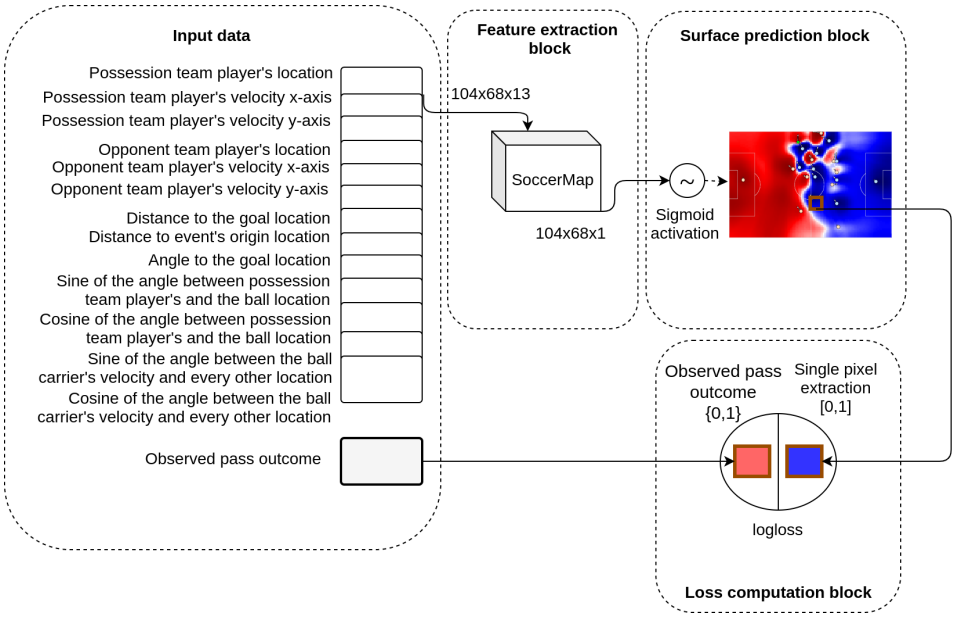


Figure 6.1: Representation of the neural network architecture for the pass probability surface estimation, for a coarsened representation of size 104×68×13. Thirteen layers of spatial features are fed to a SoccerMap feature extraction block, which outputs a 104×68×1 prediction surface. A sigmoid activation function is applied to each output, producing a pass probability surface. The output at the destination location of an observed pass is extracted, and the log-loss between this output and the observed outcome of the pass is back-propagated to learn the network parameters

concepts and dynamic pressure lines. Finally, we add a layer with the pass probability surface, considering that this can provide valuable information to estimate the expected value of passes. This surface is calculated by using a pre-trained version of a model for the architecture presented in Section 6.1.2.

The input data is fed to a SoccerMap feature extraction block to obtain a single prediction surface. In this case, we must observe that the expected value of G should reside within the $[-1, 1]$ range, as described in Section 3.1.1. To do so, in the surface prediction block, we apply a sigmoid activation function to the SoccerMap predicted surface obtaining an output within $[0, 1]$. We then apply a linear transformation, so the final prediction surface consists of values in the $[-1, 1]$ range. Notably, our modeling approach does not assume that a successful pass must necessarily produce a positive reward or that missed passes must produce a negative reward.

The loss computation block computes the MSE between the predicted values and the reward assigned to each pass, defined in Equation 2.6. Note that the model design is independent of the reward choice for passes. In this work we choose a long-term reward associated with the observed outcome of the possession for the implementation of this model. The details about this implementation are described in Section 6.2.2.

6.1.4 Pass selection probability

Until now, we have models for estimating both the probability and expected value surfaces for both successful and missed passes. In order to produce a single-valued estimation of the expected value of the possession given a pass is selected, we model the pass selection probability $\mathbb{P}(A = \rho, D_t|T_t)$ as defined in Equation 3.1. The values of a pass selection probability surface must necessarily add up to 1, and will serve as a weighting matrix for obtaining the single estimate.

Both the input and feature extraction blocks of this architecture are equivalent to those designed for the pass success probability model (see Section 6.1.2). However, we use the softmax activation function presented in Equation 2.9 for the surface prediction block, instead of a sigmoid activation function. We then extract the predicted value at a given pass destination location and compute the log-loss between that predicted value and 1, since only observed passes are used. With the different models presented in Section 6.1.1, we can now provide a single estimate of the expected value given a pass action is selected, $\mathbb{E}[G|A = \rho, T_t]$.

6.1.5 Estimating ball drive probability

We will focus now on the components needed for estimating the expected value of ball drive actions. In this work's scope, a ball drive refers to actions where a player keeps control of the ball, following the definition presented in Section 3.1.5. For this implementation, ball drives lasting more than 1 second are split into a set of individual ball drives of 1-second duration. While keeping the ball, the player might sustain the ball-possession or lose the ball (either because of bad control, an opponent interception, or by driving the ball out of the field, among others). The probability of keeping control of the ball with these conditions is modeled by the expression $\mathbb{P}(O_\delta = 1|A = \delta, T_t)$.

We use a standard neural network architecture to learn a model for this probability, consisting of two fully-connected layers, each one followed by a layer of ReLu activation functions, and a single-neuron output preceded by a sigmoid activation function. We provide a state representation for observed ball drive actions that are composed of a set of spatial and contextual features, detailed in Appendix A. Among the spatial features, the level of pressure a player in possession of the ball receives from an opponent player is considered to be a critical piece of information to estimate whether the possession is maintained or lost. We model pressure through two additional features: the opponent's team density at the player's location and the overall team pitch control at that same location. Another factor that is considered to influence the ball drive probability is the player's contextual-relative location at the moment of the action. We include two features to provide this contextual information: the closest opponent's vertical pressure line and the closest possession team's vertical pressure line to the player. These two variables are expected to serve as a proxy for the opponent's pressing behavior and the player's relative risk of losing the ball. By adding features related to the spatial pressure, we can get a better insight into how pressed that player is within that context and then have better information to decide the probability of keeping the ball. We train this model by optimizing the loss between the estimated probability and observed ball drive actions that are labeled as successful or missed, depending on whether the ball carrier's team can keep the ball's possession during after the ball drive is attempted.

6.1.6 Estimating ball drive expectation

Finally, once we have an estimate of the ball drive probability, we still need to obtain an estimate of the expected value of ball drives, in order to model

the expression $\mathbb{E}[G|A = \delta, T_t]$, presented in Equation 3.4. While using a different architecture for feature extraction, we will model both $\mathbb{E}[G|A = \delta, O_\delta = 1, T_t]$ and $\mathbb{E}[A = \delta, O_\delta = 0, T_t]$, following an analogous approach of that used in Section 6.1.3.

Conceptually, by keeping the ball, players might choose to continue a progressive run or dribble to gain a better spatial advantage. However, they might also wait until a teammate moves and opens up a passing line of lower risk or higher quality. By learning a model for the expression $\mathbb{E}[G|A = \delta, T_t]$ we aim to capture the impact on the expected possession value of these possible situations, all encapsulated within the ball drive event. We use the same input data set and feature extractor architecture used in Section 6.1.5, with the addition of the ball drive probability estimation for each example. Similarly to the loss surface prediction block of the expected value of passes (see Section 6.1.3), we apply a sigmoid activation function to obtain a prediction in the $[0, 1]$ range, and then apply a linear transformation to produce a prediction value in the $[-1, 1]$ range. The loss computation block computes the MSE loss between the observed reward value assigned to the action and the model output.

6.1.7 Expected goals model

Once we have a model for the expected values of passes and ball drives, we only need to model the expected value of shots to obtain a full value state-value estimation for the action set A. We want to model the expectation of scoring a goal at time t given that a shot is attempted, defined as $\mathbb{E}[G|A = \varsigma]$. This expression is typically referred to as xG and is arguably one of the most popular metrics in soccer analytics (Eggels, 2016). For estimating this xG model we include spatial and contextual features related derived from the 22 players' and the ball's locations, to account for the nuances of shooting situations.

Intuitively, we can identify several spatial factors that influence the likelihood of scoring from shots, such as the level of defensive pressure imposed on the ball carrier, the interceptability of the shot by close opponents, or the goalkeeper's location. Specifically, we add the number of opponents that are closer than 3 meters to the ball-carrier to quantify the level of immediate pressure on the player. Additionally, we account for the interceptability of the shot (blockage count) by calculating the number of opponent players in the triangle formed by the ball-carrier location and the two posts. We include three additional features derived from the location of the goalkeeper.

The goalkeeper's location can be considered an important factor influencing the scoring probability, particularly since he has the considerable advantage of being the only player that can stop the ball with his hands. In addition to this spatial information, we add a contextual feature consisting of a boolean flag indicating whether the shot is taken with the foot or the head, the latter being considered more difficult. Additionally, we add a prior estimation of expected goal as an input feature to this spatial and contextual information, produced through the baseline xG model described in Section 4.3.2. The full set of features is detailed in Appendix A.

Having this feature set, we use a standard neural network architecture with the same characteristics as the one used for estimating the ball drive probability, explained in Section 6.1.5, and we optimize the MSE between the predicted outcome and the observed reward for shot actions. The long-term reward chosen for this work is detailed in Section 6.2.2.

6.1.8 Action selection probability

Finally, to obtain a single-valued estimation of EPV we weigh the expected value of each possible action with the respective probability of taking that action in a given state, as expressed in Equation 3.1. Specifically, we estimate the action selection probability $\mathbb{P}(A|T_t)$, where A is the discrete set of actions described in Section 3.1.1. We construct a feature set composed of both spatial and contextual features. Spatial features such as the ball location and the distance and angle to the goal provide information about the ball carrier's relative location in a given time instance. Additionally, we add spatial information related to the attacking team's pitch control and the degree of spatial influence of the opponent team near the ball. On the other hand, the location of both teams' dynamic lines relative to the ball location provides the contextual information to the state representation. We also include the baseline estimation of xG at that given time, which is expected to influence the action selection decision, especially regarding shot selection. The full set of features is described in Appendix A. We use a neural network architecture, analogous to those described in Section 6.1.5 and Section 6.1.6. This final layer of the feature extractor part of the network has size 3, to which a softmax activation function is applied to obtain the probabilities of each action. We model the observed outcome as a one-hot encoded vector of size 3, indicating the action type observed in the data, and optimize the categorical cross-entropy between this vector and the predicted probabilities, which is equivalent to the log-loss.

6.2 Experiments and results

In this section we describe the experimental setup for implementing the models described in Section 6.1, using a large spatiotemporal tracking data and events dataset. We first describe the characteristics of the dataset, and the approach followed for defining the estimands of each separate problem. Then we provide details about the experimental setup, indicating the way the dataset is split, the hyperparameters involved in the model selection process, and the metrics used to evaluate the results. Finally, we present the results where we show we can obtain calibrated probability estimates from each of the separated models, as well as from the joint estimation of glsepv.

6.2.1 Datasets

We build different datasets for each of the presented models based on optical tracking data and event data from 633 EPL matches from the 2013/2014 and 2014/2015 season, provided by *STATS LLC*. This tracking data source consists of every player's location and the ball at a $10Hz$ sampling rate, obtained through semi-automated player and ball tracking performed on match videos. The tracking data provided is integrated with event data consisting of human-labeled on-ball actions observed during the match, including the time and location of both the origin and destination of the action, the player who takes action, and the outcome of the event. Following our model design, we focus exclusively on the pass, ball drive, and shot events. Table 6.1 presents the total count for each of these events according to the dataset split presented below in Section 6.2.3. The definition of success varies from one event to another: a pass is successful if a player of the same team receives it, a ball drive is successful if the team does not lose control of the ball after the action occurs, and a shot is labeled as successful if a goal is scored from that shot. Given this data, we can extract the tracking data snapshot, defined in Section 3.1.1, for every instance where any of these events are observed. From there, we can build the input feature sets defined for each of the presented models. For the detailed list of features used, see Appendix A. For each sample, the players' and the ball locations are normalized so the team taking the action is attacking from left to right (i.e., scores goals in the rightmost goal, and concedes goals in the leftmost goal of the field).

Table 6.1: Total count of events included within the tracking data of 633 EPL matches from the 2013/2014 and 2014/2015 season

Data Type	# Total	# Training	# Validation	# Test	% Success
Match	633	379	127	127	-
Pass	480,670	288,619	96,500	95,551	79.64
Ball drive	413,123	284,759	82,271	82,093	90.60
Shot	13,735	8,240	2,800	2,695	8.54

6.2.2 Defining the estimands

Each of the components of the EPV structured model has different estimands or outcomes. For both the pass success and ball drive success probability models, we define a binomially distributed outcome, according to the definition of success provided in 6.2.1. These outcomes correspond to the short-term observed success of the actions. For the pass selection probability, we define the outcome as a binomially distributed random variable. A value of 1 is given for every observed pass in its corresponding destination location. We define the action selection model's estimand as a multinomially distributed random variable that can take one of three possible values, according to whether the selected action corresponds to a pass, a ball drive, or a shot.

For the EPV estimations of passes, ball drives, and shot actions, respectively, we define the estimand as a long-term reward, corresponding to the outcome of the possession where that event occurs. We follow the definition of possession presented in Section 3.1.4, where a possession starts with a kick-off event and ends when a goal is observed, or a match half ends. By doing this, we allow the ball to either go out of the field or change control between teams an undefined number of times until the next goal is observed. Once a goal is observed, all the actions between the goal and the previous one are assigned an outcome of 1 if the action is taken by the scoring team or -1 otherwise. If the match half ends before observing the next goal, the actions' outcome value is set to 0. Following this, each action gets assigned a long-term reward as an outcome.

Additionally, we will include the possession resetting state described in Section 3.1.4 to limit possessions' time extent. There is a low frequency of goals in matches (2.8 goals on average in our dataset) compared to the number of observed actions (1,433 on average). Given this, the definition of the time extent of possessions is expected to influence the balance between individual actions' short-term value and the long-term expected outcome

after that action is taken. Let ϵ be the constant representing the time in seconds between each action and the next goal; all the actions observed more than ϵ time from the observed goal received a reward of 0. For this work, we choose $\epsilon = 15s$, which corresponds to the average duration of standard soccer possessions in the available matches. Note this is equivalent to assuming that any given state of possession only has ϵ seconds impact.

For the implementation of this model we will use only passes, ball drives and shot actions that are observed within an open-play phase of the possession, and ignore the actions occurring during set-pieces. We will say that an action belongs to a set-piece if it is observed 5 seconds or less from the start of a direct or indirect free-kick, a corner kick, a throw-in or a penalty kick. All the other actions are considered to occur in open-play. It is important to remark that all the goals available in the dataset are used in this implementation, including those occurring within a set-piece time range. This means that if a goal is scored in a corner kick, all the actions preceding the goal will be labeled with -1, 1 or 0 (according to the definition of possession described above), except for those that are 5 seconds or less closer to the goal. By doing this, our implementation focuses on learning the expected value of open-play actions, and leaves for future work the modeling of set-pieces, since these involve different spatiotemporal dynamics.

6.2.3 Model setting

We randomly sample the available matches and split them into training (379), validation (127), and test sets (127). From each of these matches, we obtain the observed on-ball actions and the tracking data snapshots to construct the set of input features corresponding to each model, detailed in Appendix A. The events are randomly shuffled in the training dataset to avoid bias from the correlation between events that occur close in time. We use the validation set for model selection and leave the test set as a hold-out dataset for testing purposes. We train the models using the ADAM algorithm (Kingma and Ba, 2014), and set the β_1 and β_2 parameters to 0.9 and 0.999 respectively. For all the models we perform a grid search on the learning rate ($\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$), and batch size parameters ($\{16, 32\}$). We use early stopping with a delta of 10^{-3} for the pass success probability, ball drive success probability, and action selection probability models, and 10^{-5} for the rest of the models.

6.2.4 Model calibration

We include an after-training calibration procedure within the processing pipeline for the pass success probability and pass selection probability models, which presented slight calibration imbalances on the validation set. We use the temperature scaling calibration method (see Section 2.5.1) for both models, a useful approach for calibrating neural networks (Guo et al., 2017). We apply these post-calibration procedures exclusively on the validation set.

6.2.5 Evaluation Metrics

For the pass success probability, keep ball success probability, pass selection probability, and action selection models, we use the cross-entropy loss, defined in Equation 2.4. For the first three models, where the outcome is binary, we set the number of class as $M = 2$. We can directly observe that for this set-up, the cross-entropy is equivalent to the negative log-loss defined in Equation 2.5. For the action selection model, we set $M = 3$. For the rest of the models, corresponding to EPV estimations, we can observe the outcome takes continuous values in the $[-1, 1]$ range. For these cases, we use the MSE as a loss function, defined in Equation 2.6, by first normalizing both the estimated and observed outcomes into the $[0, 1]$ range.

We are interested in obtaining calibrated predictions for all of the models, as well as for the joint EPV estimation. Having the models calibrated allows us to perform a fine-grained interpretation of the variations of EPV within subsets of actions, as shown in Chapter 7. For doing this, we report the ECE metric and present a series of calibration reliability plots following a similar methodology as presented in Section 5.2.4.

6.2.6 Results

Table 6.2 presents the results obtained in the test set for each of the proposed models. The loss value corresponds to either the cross-entropy or the mean squared loss, as detailed in Section 6.2.5. The table includes the optimal values for the batch size and learning rate parameters, the number of parameters of each model, and the number of examples per second that each model can predict.

We can observe that the loss value reported for the final joint model is equivalent to the losses obtained for the EPV estimations of each of the three types of action types, showing stability in the model composition.

Table 6.2: The average loss and calibration value for each of the components of the EPV model, as well as for the joint EPV estimation, on the corresponding test datasets. Additionally, the table presents the optimal value of the hyper-parameters, total number of parameters, and the number of predicted examples by second, for each of the models

Model	Loss	ECE	Batch Size	Learning Rate	# Params.	Ex. (s)
Pass probability	0.1900	0.0047	32	1e−4	401,259	942
Ball drive probability	0.2803	0.0051	32	1e−3	128	67,230
Pass successful EPV	0.0075	0.0011	16	1e−6	403,659	899
Pass missed EPV	0.0085	0.0015	16	1e−6	403,659	899
Pass selection probability	5.7134	-	32	1e−5	401,259	984
Pass EPV * Pass selection	0.0067	0.0011	-	-	-	-
Ball drive successful EPV	0.0128	0.0022	16	1e−4	153	57,441
Ball drive missed EPV	0.0072	0.0025	16	1e−4	153	57,441
Shot EPV	0.2421	0.0095	16	1e−3	231	72,455
Action selection probability	0.6454	-	32	1e−3	171	23,709
EPV	0.0078	0.0023	-	-	-	-

The shot EPV loss is higher than the ball drive EPV and pass EPV losses, arguably due to the considerably lower amount of observed events available in comparison with the rest, as described in Section 6.2.1. While the number of examples per second is directly dependent on the models' complexity, we can observe that we can predict 899 examples per second in the worst case. This value is 89 times higher than the sampling rate of the available tracking data (10Hz), showing that this approach can be applied for the real-time estimation of EPV and its components.

Regarding the models' calibration, we can observe that the ECE metrics present consistently low values along with all the models. Figure 6.2 presents a fine-grained representation of the probability calibration of each of the models. The x-axis represents the mean predicted value for a set of $K = 10$ equal-sized bins, while the y-axis represents the mean observed outcome among the examples within each corresponding bin. The circle represents the percentage of examples in the bin relative to the total number of examples. In these plots, we can observe that the different models provide calibrated probability estimations along their full range of predictions, which is a critical factor for allowing a fine-grained inspection of the impact that specific actions have on the expected possession value estimation. Additionally, we can observe the different ranges of prediction values that each model produces. For example, ball drive success probabilities are distributed more often above 0.5, while pass success probabilities cover a

wide range between 0 and 1, showing that it is harder for a player to lose the ball when keeping the ball than it is to lose the ball by attempting a pass towards another location on the field. The action selection probability distribution is heavily influenced by each action type's frequency, showing a higher frequency and broader distribution on ball drive and pass actions compared with shots. The joint EPV model's calibration plot shows that the proposed approach of estimating the different components separately and then merging them back into a single EPV estimation provides calibrated estimations. We applied post-training calibration exclusively to the pass success probability and the pass selection probability models, obtaining temperature values of 0.82 and 0.5, respectively.

Having this, we have obtained a framework of analysis that provides accurate estimations of the long-term reward expectation of the possession, while also allowing for a fine-grained evaluation of the different components comprising the model.

6.3 Inspecting the EPV components

This section presents a deeper inspection into the characteristics of the different components of the EPV framework. First, we present how the probability surfaces predicted for each passing component provide a more fine-grained representation of game situations and a visual approach to game analysis. Then, we analyze the influence of the developed spatial and contextual features in predicting the expected value from shots and the probability of selecting an action, providing a detailed view of how these models are profiting from tracking data to produce more accurate predictions. Additionally, we present a novel description of the distribution of added value by a broader set of soccer actions to better understand the relative impact of these actions according to context.

6.3.1 Visually-interpretable passing components

The different passing components developed in this work allow coaches to perform a deeper inspection into actual game situations. Specifically, the prediction of full probability surfaces instead of single statistics offers the opportunity to analyze alternative passing options than those attempted during the matches and understand the spatial dynamics from one situation to another. Figure 6.3 presents the predicted surfaces for the pass expected

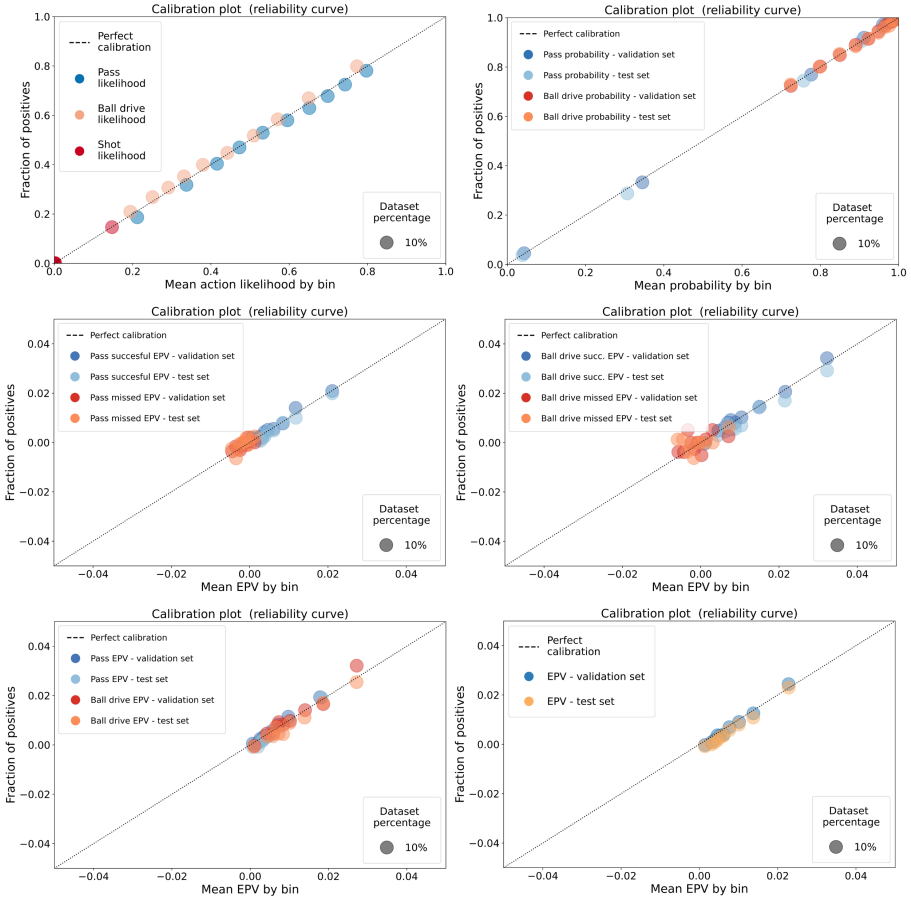


Figure 6.2: Probability calibration plots for the action selection (top-left), pass and ball drive probability (top-right), pass (successful and missed) EPV (mid-left), ball drive (successful and missed) EPV (mid-right), pass and ball drive EPV joint estimation (bottom-left), and the joint EPV estimation (bottom-right). Values in the x-axis represent the mean value by bin, among 10 equally-sized bins. The y-axis represents the mean observed outcome by bin. The circle size represents the percentage of examples in each bin relative to the total examples for each model

value (left), pass probability (center), and pass selection probability (center) components on the same game situation. First, we can observe that the SoccerMap architecture can produce very different probability surfaces from the same set of observed passes by learning the spatiotemporal features better adapted to each specific problem. In this game situation, a coach could observe that making a short pass to the defenders has a high probability of success and would provide a low but positive expected long-

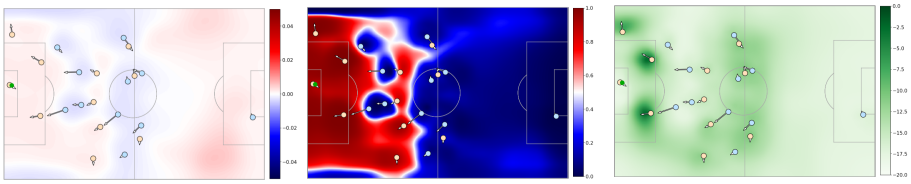


Figure 6.3: Three different passing surfaces calculated on the same game situation. On the left the pass expected value surface; on the center the pass probability surface; and on the right the pass selection probability.

term reward. Also, attempting a pass behind the first line of the opponent's pressure shows an increased risk of losing the ball and receiving a goal, despite the medium to high pass probability observed for the midfielders. Interestingly, the model captures that playing long balls to either of the wingers is expected to produce a positive reward, but that it also has a very low probability of success. Additionally, we can see that the likelihood of selecting a long-ball is low. The coach might select to instruct their team to play long balls despite the high likelihood of losing the ball to decrease the expectation of receiving a ball. Suppose the coach would prefer playing a short-passing playing style. In that case, he could indicate its defensive midfielder to move closer to the goalkeeper in these types of situations to decrease the risk of losing the ball near their goal (highlighted by the blue shaded area close to the three forward players of the blue team).

This highly informative representation of risk and reward observed in the pass expectation surface takes advantage of the surfaces produced for each component. Figure 6.4 shows how the pass probability and pass turnover surfaces are combined with the successful and missed pass expected value surfaces to produce this single estimation of the value of potential passes. In this game situation, we can observe that while the pass probability of the players in the back is high, the associated risk of losing the ball and receiving a goal lowers the total expected value. Also, we can observe that missed passes towards the box can produce a neutral and even positive expected reward. This coincides with the intuition that losing the ball near the opponent's goal increases the chance of scoring after a quick ball recovery (although it is not as high as making a successful pass). Additionally, we can observe that these surfaces also capture players' movement dynamics by showing an increased probability in farther away locations towards which a player is running.

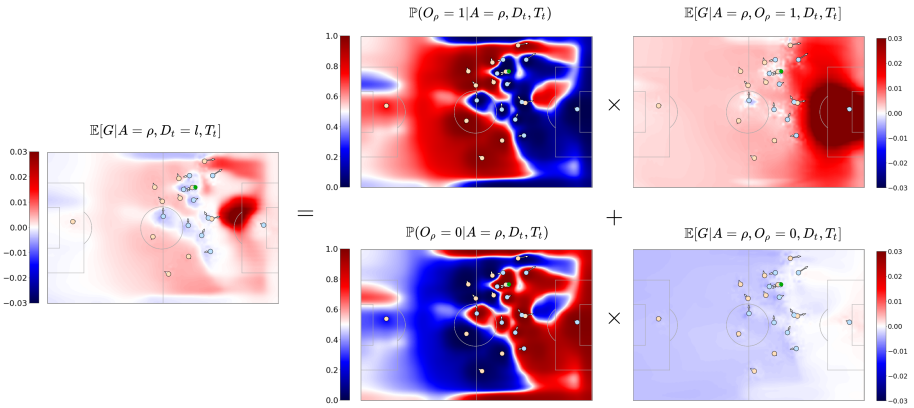


Figure 6.4: Surface of pass expected value, calculated as a combination of the predicted surfaces for pass probability and the pass expected value conditioned to the pass outcome.

6.3.2 An enhanced expected goals model

The shot component described in Section 6.1.7 introduces a series of spatial and contextual features derived from tracking data to predict the expectation of scoring from any location. These features are expected to provide a more fine-grained estimation of xG than those built with event data, such as those described in Section 2.2.4 and the baseline model presented in Section 4.3.2. To provide a deeper understanding, we calculated the importance and impact of these features in the developed model, using the SHAP framework approach (see Section 2.6). Figure 6.5 presents the mean importance (top) and the SHAP value for each of the features (bottom). We can observe that the tracking data-specific features such as the distance to the goalkeeper, the block count, the identification of the goalkeeper being surpassed, and the opponent density all provide an impact higher than 0.1 on average, showing to be influential features. Additionally, the event data features are still crucial for the model, especially the event data-based xG estimation, intended to provide a baseline estimation of the shot component. By taking a closer look at the impact of these features in the test set examples, we can see that while the lower the distance to the goalkeeper, the higher the goal expectation (unsurprisingly, give the correlation of this variable to the distance to the goal), when the goalkeeper is surpassed the goal expectation increases considerably. Similarly, when the goalkeeper is not surpassed, the impact of this feature is lower. On the other hand, the lower amount of players potentially blocking the shot (i.e., block count), the higher the goal expectation, and vice versa. We can also observe that, in general, a higher opponent den-

sity near the shooting location decreases the goal expectation considerably. The angle to goal shows to be an essential feature (also available in event data), indicating that a higher angle to the goal center decreases the goal expectation. In contrast, the more centered the shot location is, the higher the expectation of scoring goals.

6.3.3 Understanding action selection

The action selection model described in Section 6.1.8 produces an estimate of the probability of selecting a pass, ball drive, or a shot in any given game situation. These three components play a critical weighting role for the decomposed EPV model. In this section, we analyze the impact that the different spatial and contextual features have on the probability of selecting one action over the other. Figure 6.6 presents the average predicted probability of selecting a pass (left), ball drive (center) and a shot (right), for each location in a 104×68 representation of a soccer field. We can see that there are areas in the soccer field where there is a considerably high preference for selecting one of the actions instead of the others. The shot selection probability is concentrated in the opponent's box and presents a radial shape where the maximum values are found near the penalty kick location. When the ball is in the first quarter of the field, we can observe a higher tendency to attempt ball drives. This tendency decreases, in favor of selecting pass, when the ball approaches the second and third quarter. Interestingly, when the ball is on the sides of the field in the last quarter, there is a higher tendency for attempting ball drives instead of passes. This might be due to the players identifying a higher chance of scoring in the long-term by carrying the ball towards the box than attempting a cross (i.e., aerial pass to the box).

While the average probability provides an overall idea of where players tend to select each action, we will inspect further into the impact of each of the spatial and contextual features. In Figure 6.7 we compare the average SHAP value of each of the features for predicting each of action types, and the SHAP value for each of the examples in the test set to understand the feature importance in each case. In the average feature importance plot, we can observe that the opponent's density and the player's pitch control in control of the ball play a fundamental role in deciding between passes and ball drives. We can see in the feature impact plots that when the opponent's density increases and the player's pitch control decreases, the probability of attempting a pass increases considerably. In contrast, the opposite is observed for ball drives. This situation captures a known dynamic

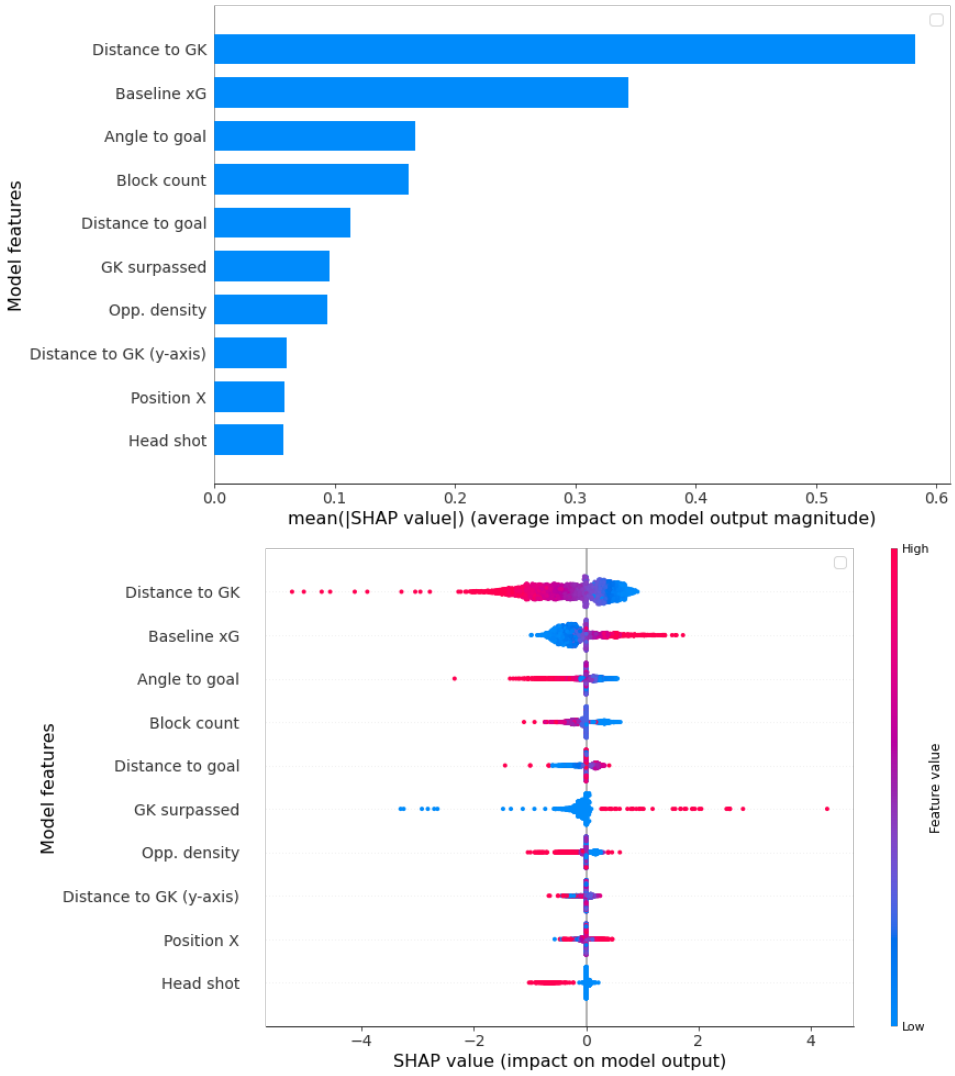


Figure 6.5: Two images showing the mean SHAP value for each of the features of the shot component of the EPV framework (top), and the SHAP value of each feature for the examples in the test set.

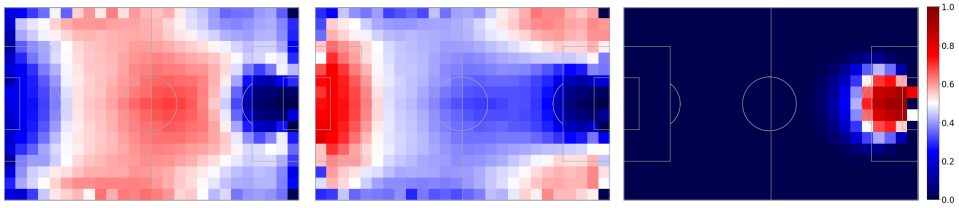


Figure 6.6: Three surfaces showing the average predicted probability of selecting a pass (left), ball drive (center) or shot (right), represented in a 104×68 grid.

in soccer, where players tend to keep the ball to attract the opponent players to their location and then make a pass when an opponent is close to avoid losing the ball. Also, we can argue that, in general, a player will tend to keep control of the ball when there are sufficient space and a lower risk of losing the ball and will tend to pass otherwise. Other features that influence the action selection probability are the closest team dynamic line and the opponent's pressure line, which provides information about the two teams' relative location. We can observe that when both the pressure line and team closest line is the lowest (i.e., ball in control of defenders and opponents' forwards pressing), there is a higher tendency to pass. The opposite is observed for ball drives, where players tend to carry the ball when they are closer to the opponent's defensive line. Additionally, this feature could be capturing the idea that forwards have a higher tendency to attempt ball drives than defenders. Regarding shot selection, the goal's distance and angle are shown to be the two most influential features. Additionally, we can observe that the higher the opponent's density, the lower the predicted probability of making a shot, showing the model can capture changes in decision making according to the spatial pressure.

6.3.4 Not all value is created (or lost) equal

There is a wide range of playing strategies that can be observed in modern professional soccer. There is no single best strategy found in successful teams, from Guardiola's creative and highly attacking FC Barcelona to Mourinho's defensive and counter-attacking Inter Milan. We could argue that a critical element for selecting a playing strategy lies in managing the risk and reward balance of actions, specifically which actions a team will prefer in each game situation. While professional coaches intuitively understand which actions are riskier and more valuable, there is no quantification of the actual distribution of the value of the most common actions in soccer.

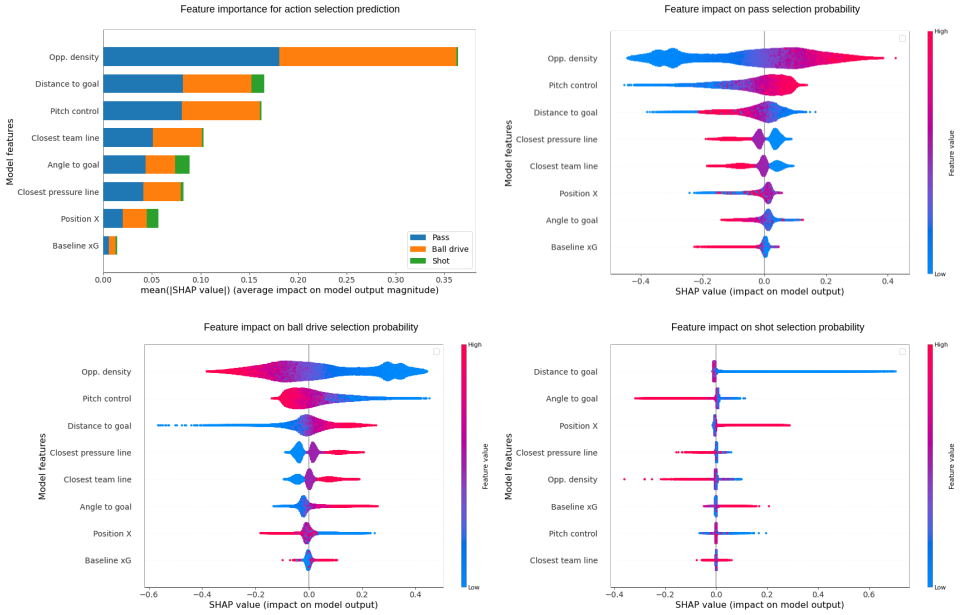


Figure 6.7: Four figures describing the importance of the action selection model's features. Top left figure presents the mean SHAP value for each of the features for predicting pass, ball drive and shot actions. The rest of the figures present the SHAP value of each feature for predicting pass (top right), ball drive (bottom left) and shot (bottom right) probability, for all the examples in the test set.

To assess the value created by individual actions we propose adapting the concept of EPV added, previously introduced in the first approach for EPV in basketball (Cervone et al., 2016b). Let ts and te be the start and ending time of an on-ball action a, and EPV_{ts} and EPV_{te} be the EPV of the possession at these time instances, respectively, we define the action's EPVA as in Equation 6.1.

$$\text{EPVA}(a) = \text{EPV}_{te} - \text{EPV}_{ts} \tag{6.1}$$

From all the passes and ball drive actions described in Section 6.2.1, and the spatial and contextual features described in Chapter 4 we derived a series of context-specific actions to compare their value distribution. We identify passes and ball drives that break the first, second, or third line from the concept of dynamic pressure lines. We define an action (pass or ball drive) to be under pressure if the player's pitch control value at the beginning of the action is below 0.4 and without pressure otherwise. A long pass is defined as a pass action that covers a distance above 30 meters. We define a pass back as passes where the destination location is closer to the team's goal than the ball's origin location. We count with manually labeled tags indicating when a pass is a cross and when the pass is missed from the available data. We identify lost balls as missed passes and ball drives ending in recovery by the opponent. For all of these action types, we calculate each observed action's added value (EPVA) based on the registered start and ending time of the action. We perform a kernel density estimation on the EPVA of each action type to obtain a probability density function. In Figure 6.8 we compare the density between all the action types. The density function value is normalized in the $[0, 1]$ range by dividing by the maximum density value to ease the distributions' visual comparison.

From Figure 6.8 we can gain a deeper understanding of the value distribution of different types of actions. From passes that break lines, we can observe that the higher the line, the broader the distribution, and the higher the extreme values. While passes breaking the first line are centered around 0 with most values ranging in $[-0.01, 0.015]$, the distribution of passes breaking the third line is centered around 0.005, and most passes fall in the interval $[-0.025, 0.05]$. Similarly, ball drives that break lines present a similar distribution as passes breaking the first line. Regarding the level of spatial pressure on actions, we can see that actions without pressure present an approximately zero-centered distribution, with most values falling in a $[-0.01, 0.01]$ range. On the other hand, actions under pressure

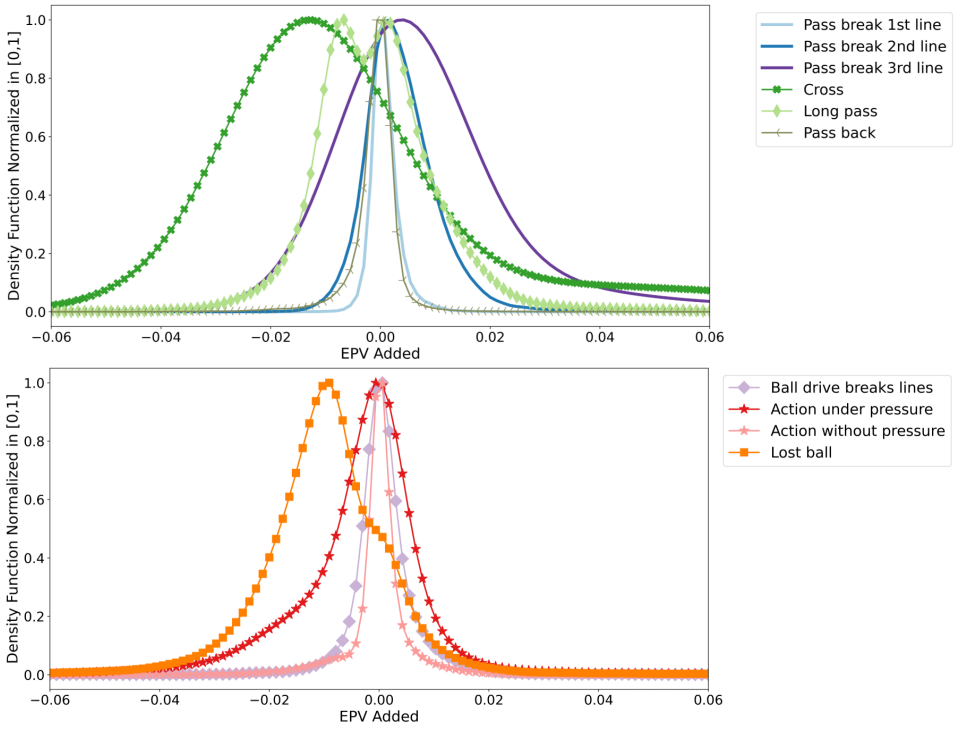


Figure 6.8: Comparison of the probability density function of the EPV added (EPVA) for ten different actions in soccer. The density function values are normalized into the $[0, 1]$ range. The normalization is obtained by dividing each density value by the maximum observed density value

present a broader distribution and a higher density on negative values. This shows both that there is more tendency to lose the ball under pressure, hence losing value, and a higher tendency to increase the value when the pressure is overcome with successful actions. Whether crosses are a successful way to reach the goal or not has been a long-term debate in soccer strategy. We can observe that crosses constitute the type of action with a higher tendency to lose significant amounts of value; however, it provides a higher probability of high-value increases in case of succeeding than other actions. Long passes share a similar situation, where they can add a high amount of value in case of success but have a higher tendency to produce high EPV losses. For years, soccer enthusiasts have argued about whether passing backward provides value or not. We can observe that, while the EPV added distribution of passing back is the narrowest, near half of the probability lies on the positive side of the x-axis, showing the potential value to

be obtained from this type of action. Finally, losing the ball often produces a loss of value. However, in situations such as being close to the opponent's box and pressure on the ball carrier, losing the ball with a pass to the box might provide an increment in the expected value of the possession, given the increased chance of a rebound.

6.4 Discussion

This chapter presents a comprehensive approach for estimating the instantaneous expected value of possessions in soccer. One of the main contributions of this work is showing that by deconstructing a single expectation into a series of lower-level statistical components and then estimating each of these components separately, we can gain greater interpretation insight into how these different elements impact the final joint estimation. Also, instead of depending on a single-model approach, we can make a more specialized selection of the models, learning approach, and input information that is better suited for learning the specific problem represented by each sub-component of the EPV decomposition. The deep learning architectures presented for the different passing components produce full probability surfaces, providing rich visual information for coaches that can be used to perform fine-grained analysis of player and team performances. We show that we can obtain calibrated estimations for all the decomposed model components, including the single-value estimation of the expected possession value of soccer possessions. We employ a broad set of novel spatial and contextual features for the different models presented, allowing rich state representations. In the next chapter, we present a series of practical applications showing how this framework could be used as a support tool for coaches, allowing them to solve new upcoming questions and accelerating the problem-solving necessities that arise daily in professional soccer.

We consider that this work provides a relevant contribution to improving the practitioners' interpretation of the complex dynamics of professional soccer. With this approach, soccer coaches gain more convenient access to detailed statistical estimations that are unusual in their practice and find a visual approach to analyze game situations and communicate tactics to players. Additionally, on top of this framework, there is a large set of novel research that can be derived, including on-ball and off-ball player performance analysis, team performance and tactical analysis for pre-match and post-match evaluation, player profile identification for scouting, young players evolution analysis, match highlights detection, and enriched visual

interpretation of game situations, among many others.

Chapter 7

Practical applications

The main purpose of soccer analytics is to provide practical applicability of advanced data analysis. Most of the design, modeling and implementation decisions presented in this work, were made taking into consideration that the produced models would provide the greatest possible practical applicability. In this section we present a series of novel practical applications that are achieved through exploiting the rich set of components provided by the proposed EPV framework, and the spatial and contextual features presented in this work. We structure these applications into three main groups: match and team analysis, off-ball performance, and on-ball performance. We show that this broad set of applications can be directly derived either from the overall EPV estimation, the estimations provided by single components, or the adaptation of any of these components to specific game situations.

7.1 Match and team analysis

7.1.1 A real-time control room

In most team sports, coaches make heavy use of video to analyze player performance, show players their correctly or incorrectly performed actions, and even point out other possible decisions the player may have taken in a given game situation. The presented structured modeling approach of the EPV provides the advantage of obtaining numerical estimations for a set of game-related components, allowing us to understand the impact that each of them has on the development of each possession. Based on this, we can build a control room-like tool like the one shown in Figure 7.1, to help coaches analyze game situations and communicate effectively with players.

Follow the link Control Room Video to watch a video showing the live usage of this tool.

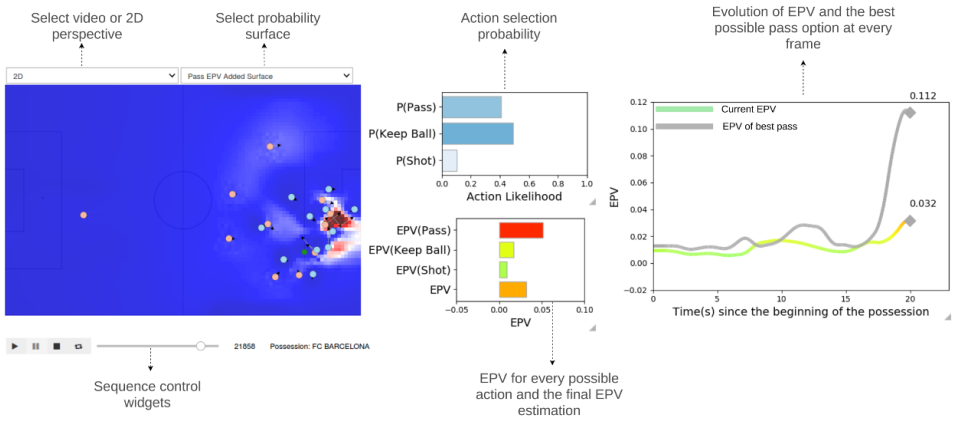


Figure 7.1: A visual control room tool based on the EPV components. On the left, a 2D representation of the game state at a given frame during the match, with an overlay of the pass EPV added surface and selection menus to change between 2D and video perspective, and to modify the surface overlay. On the bottom-left corner, a set of video sequence control widgets. On the center, the instantaneous value of selection probability of each on-ball action, and the expected value of each action, as well as the overall EPV value. On the right, the evolution of the EPV value during the possession and the expected EPV value of the optimal passing option at every frame. See Control Room Video for a video showing the live usage of the tool.

The control room tool presented in Figure 7.1 shows the frame-by-frame development of each of the EPV components. Coaches can observe the match's evolution in real-time and use a series of widgets to inspect into specific game situations. For instance, in this situation, coaches can see that passing the ball has a better overall expected value than keeping the ball or shooting. Additionally, they can visualize in which passing locations there is a higher expected value. The EPV evolution plot on the right shows that while the overall EPV is 0.032, the best possible passing option is expected to increase this value up to 0.112. The pass EPVA surface overlay shows that an increase of value can be expected by passing to the teammates inside the box or passing to the teammate outside the box. With this information and their knowledge on their team, coaches can decide whether to instruct the player to take immediate advantage of these kinds of passing opportunities or wait until better opportunities develop. Additionally, the player can gain a more visual understanding of the potential value of passing to specific locations in this situation instead of taking a shot. If the player tends to shoot in these kinds of situations, the coach could show that keeping the ball

or passing to an open teammate has a better goal expectancy than shooting from that location.

This visual approach could provide a smoother way to introduce advanced statistics into a coaching staff analysis process. Instead of evaluating actions beforehand or only delivering hard-to-digest numerical data, we provide a mechanism to enhance coaches' interpretation and player understanding of the game situations without interfering with the analysis process.

7.1.2 Team-based passing selection tendencies

The pass selection component presented in Section 6.1.4, provides a fine-grained evaluation of the passing likelihood in different situations. However, it is clear to observe that passing selection is likely to vary according to a team's player style and the specific game situation. While a league-wide model might be useful for grasping the expected behavior of a typical team in the league, a soccer coach will be more interested in understanding the fine-grained details that separate one team from the other. Once we train a SoccerMap network to obtain this league-wide model, we can fine-tune the network with passes from each team to grasp team-specific behavior. In this application example, we trained the pass selection model with passes from all the teams from EPL season 2014-2015. Afterward, we retrained the initial model with passes from two teams with different playing-styles: Liverpool and Burnley.

In Figure 7.2 we compare the pass selection tendencies between Liverpool (left column) and Burnley (right column). On the top left corner of both columns, we show a 2D plot with the difference between the league mean passing selection heatmap, and each team's mean passing selection heatmap, when the ball is within the green circle area. We can observe that Liverpool tends to play short passes, while Burnley has a higher tendency of playing long balls to the forwards or opening on the sides. However, this kind of information would not escape from the soccer coach's intuition, so we require a more fine-grained analysis of each team's tendencies in specific situations. In the two plots of Figure 7.2 we show over each players' location the percentage increase in passing likelihood compared with the league's mean value. In this situation, we can observe that when a left central defender has the ball during a buildup, Liverpool will tend to play short passes to the closest open player, while Burnley has a considerably higher tendency to play long balls to the forwards, especially if forwards are

starting a run behind the defender's backs, such as in this case. Through a straightforward fine-tuning of the SoccerMap-based model, we can provide detailed information to the coach for analyzing specific game situations.

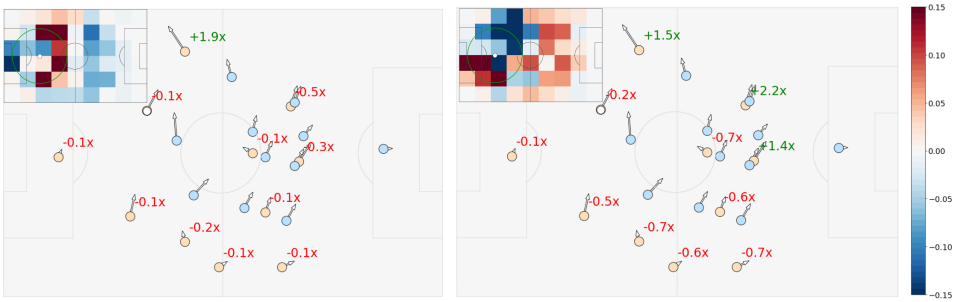


Figure 7.2: A game-state representation of a real game situation in soccer. Above each player (circles) we present the added percentage difference of pass likelihood in that given situation in comparison with the league for two teams: Liverpool (left column) and Burnley (right column). The heatmaps in both top left corners of each column represent the mean difference in pass selection likelihood with the league, when the ball is located within the green circle.

7.1.3 Optimizing lineup selection

Most teams in the best professional soccer leagues have at least one player who is the key playmaker. Often, coaches want to ensure that the team's strategy is aligned with maximizing the performance of these key players. In this section, we leverage tracking data and the passing components of the EPV model to analyze the relationship between the well known attacking midfielder David Silva and his teammates when playing at Manchester City in season 14/15. We analyze two different situations: when Silva has the ball and when any other player has the ball and Silva is on the field. We calculated the playing minutes each player shared with Silva and aggregated both the on-ball EPVA and expected off-ball EPVA of passes between each player pair for each match in the season. The on-ball EPV added is calculated following Equation 6.1 presented in Section 6.3.4, where we add the difference of the EPV at the end and start time of observed passes. On the other hand, then off-ball EPV added is calculated by calculating the difference between the predicted EPV of a pass taken to the location of the teammate (accounting for its velocity) and the EPV at the time the actual is evaluated. We also calculate the selection percentage, defined as the percentage of time Silva chooses to pass to that player when available (and vice

versa).

Figure 7.3 presents the sending and receiving maps involving David Silva and each of the two players with more minutes by position in the team. Every player is placed according to the most commonly used position in the league. Players represented by a circle with a solid contour have the higher sum of off-ball and on-ball EPV in each situation than the teammate assigned for the same position, presented with a dashed circle. The size of the circle represents the selection percentage of the player in each situation. We represent off-ball EPVA by the arrows' color, and on-ball EPVA of attempted passes by the arrow's size.

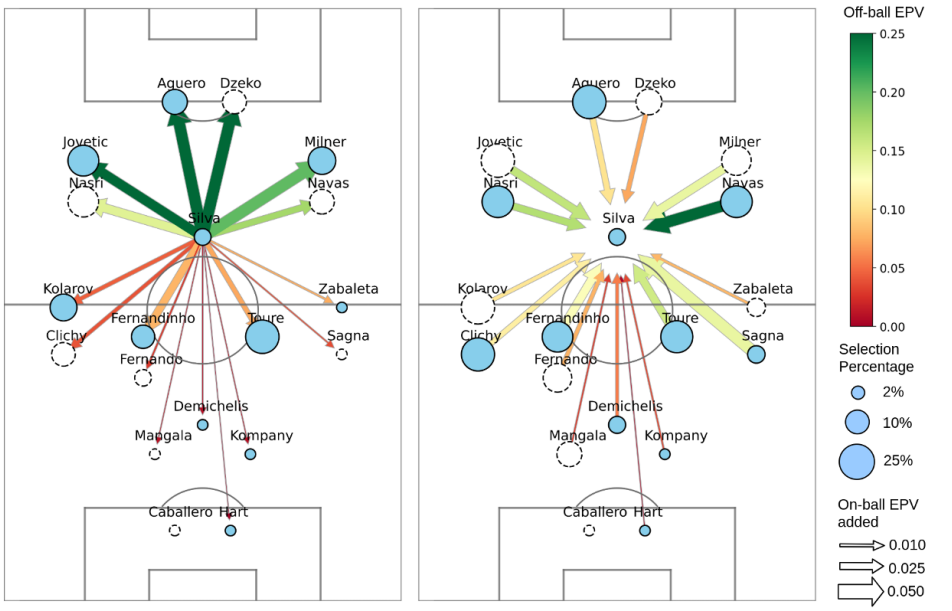


Figure 7.3: Two passing maps representing the relationship between David Silva and each of the two players with more minutes by position in the Manchester City team during season 14/15. The figure on the left represents passes attempted by Silva, while the figure on the right represents passes received by Silva. The color of the arrow represents the average expected off-ball EPVA of the passes. The size of the circle represents the selection percentage of the destination player of the pass. Circles present a solid contour when that player is considered better for Silva than the teammate in the same position. The size of the arrow represents the mean on-ball EPVA of attempted passes. Players are placed according to their highest used position on the field. All metrics are normalized by minutes played together and multiplied by 90 minutes

We can see that both the wingers and forwards generate space for Silva

and receive high added value from his passes. However, the most frequently selected player is the central midfielder Yaya Touré, who also looks for Silva often and is the midfielder providing the highest value to him. Regarding the other central midfielder, Fernandinho has a better relationship with Silva in terms of received and added value than Fernando. Silva shows a high tendency to play with the wingers; however, while Milner and Jovetic can create space and receive value from Silva, Navas and Nasri find Silva more often, with higher added value. Based on this, the coach can decide whether he prefers to lineup wingers that can benefit from Silva's passes or wingers, increasing Silva's participation in the game. A similar situation is presented with the right and left-backs. Additionally, we can observe that Silva tends to be a highly preferable passing option for most players. This information allows the coach to gain a deeper understanding of the effective off-ball and on-ball value relationship that is expected from every pair of players and can be useful for designing playing strategies before a match.

7.2 Off-ball performance

7.2.1 Deciding how to defend against buildups

A prevalent and challenging decision that coaches face in modern professional soccer is how to defend an organized buildup by the opponent. We consider an organized buildup as a game situation where a team has the ball behind the first pressure line. When deciding how to press, a coach needs to decide first in which zones they want to avoid the opponent receiving passes. Second, how to cluster their players in order to minimize the chances of the opponent moving forward. This section uses the EPV passing components and the dynamic pressure lines to analyze how to press Brendan Rodgers' Liverpool (season 14/15).

We identify the formation being used every time by counting the number of players in each pressure line. We assume there are only three pressure lines, so all formations are presented as the number of defenders followed by the number of midfielders and forwards. For every formation faced by Liverpool during buildups, we calculate both the mean off-ball and on-ball advantage in every location on the field. The on-ball advantage is calculated as the sum of the EPVA of passes with positive EPVA. On the other hand, the off-ball advantage is calculated as the sum of positive potential EPVA. We then say that a player has an off-ball advantage if he is located in a position where, in case of receiving a pass, the EPV would increase. Figure 7.4

presents two heatmaps for every of the top 5 formations used against Liverpool during buildups, showing the distribution where Liverpool obtained on-ball and off-ball advantages, respectively. The heatmaps are presented as the difference with the mean heatmap in all of Liverpool's buildups during the season.

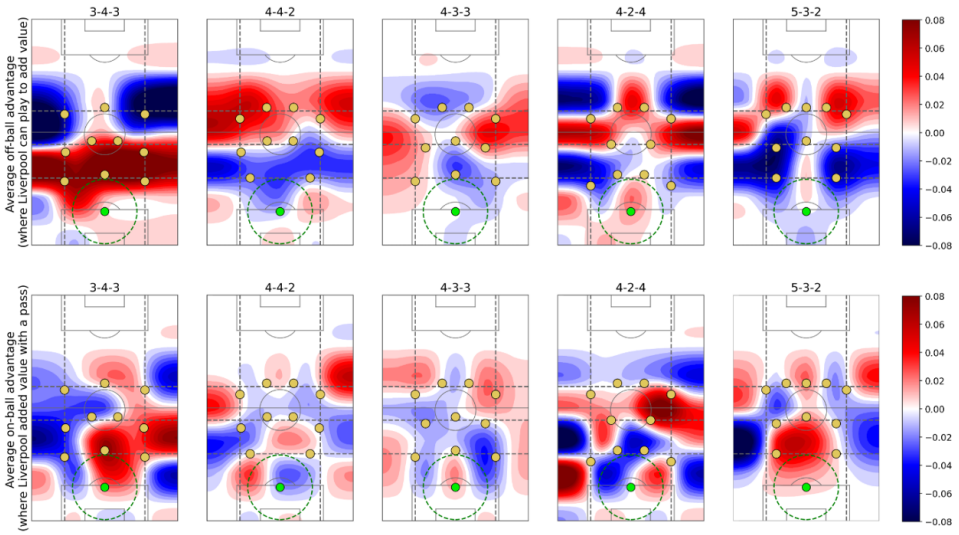


Figure 7.4: In the first row, one distribution for every formation Liverpool's opponents used during Liverpool's organized buildups, showing the difference between the distribution of off-ball advantages and the mean distribution. The second row is analogous to the first one, presenting the on-ball EPVA distributions. The green circle represents the ball location

We will assume that the coach wants to avoid Liverpool playing inside its team block during buildups. We can see that when facing a 3-4-3 formation, Liverpool can create higher off-ball advantages before the second pressure line and manages to break the first line of pressure by the inside successfully. Against the 4-4-2, Liverpool has more difficulties in breaking the first line but still manages to do it successfully while also generating spaces between the defenders and midfielders, facilitating long balls to the sides. If the coaches' team does not have a good aerial game, this would be a harmful way of pressing. We can see the 4-3-3 is an ideal pressing formation for avoiding Liverpool playing inside the pressing block. This pressing style pushes the team to create spaces on the outside, before the first pressure line and after the second pressure line. In the second row, we can observe

that Liverpool struggles to add value by the inside and is pushed towards the sides when passing. The 4-2-4 is the formation that avoids playing inside the block the most; however, it also allows more space on the sides of the midfielders. We can see that Liverpool can take advantage of this and create spaces and make valuable passes towards those locations. If the coach has fast wing-backs that could press receptions on long balls to the sides, this could be an adequate formation; otherwise, 4-3-3 is still preferable. Finally, the 5-3-2 provides significant advantages to Liverpool that can create spaces both by the inside above the first pressure line and behind the defenders back, while also playing towards those locations effectively.

This kind of information can be highly useful to a coach to decide tactical approaches for solving specific game situations. If we add the knowledge that the coach has of his players' qualities, he can make a fine-tuned design of the pressing he wants his team to develop.

7.2.2 Calculating player's optimal positioning

One of the most important skills in soccer is the player's ability to be located in spaces that increase the likelihood of receiving a pass successfully. Coaches often assess and correct players positioning in specific game situations when they observed a positional tendencies that are considered to be incorrect or detrimental for the team. By using the probability surfaces generated by pass probability component, we can detect the best possible location a player could occupy to increase the probability of receiving a pass directly. Additionally, to evaluate specific game situations selected by the coach, we could automatically identify other game situation in the past matches that may have gone unnoticed, and track player's evolution in this skill.

Given a game-state, where a player is in possession of the ball, we calculate the optimal location of each teammate. For doing so, we first generate a series of alternative game situations that are identical to current game state except for the location of one teammate. The teammate location is translated to any other possible location in 5×5 grid around its expected location in the next second (based on the player's current velocity). To obtain the optimal location of this player, we predict the pass probability surface for each of the calculated alternative situations. The location within that grid with the highest (positive) probability difference with the current player's location probability is set as the optimal passing location. Additionally, a set of sub-optimal passing locations are obtained by identifying

locations with positive probability difference and that are at least 5 meters away from the optimal location. We repeat this procedure for each player in the attacking team (i.e. in possession of the ball) to generate the optimal location of each player.

In Figure 7.5 we observe in green circles the expected pass probability added if the player would have been placed in that location instead. We can observe that if Semedo would move towards the right lane to increase the passing angle, the pass probability would increase 58%. Both Rakitic and De Jong would increase over 30% their probabilities of receiving a pass if they run towards the open space near them that separates them from the nearby opponents. For the two cases of Busquets we can observe the value of providing a set of alternative locations, instead of a single one. The optimal location for Busquets is placed behind the back of the opponent in a more advanced position, and provides an extra 21% pass probability. However, being this player a defensive midfielder, the coach might prefer he moves to the alternative location, where he offer a direct pass line to Pique, and still increase its probability in 7%. Also notice that the optimal position predicted for De Jong is close to the optimal position predicted for Busquets. Here, the coach can decide, based on the desired tactics and his knowledge of his players, what movements to demand to each player, and use this visual and numerical information to indicate to players the effect of this.

7.2.3 Quantifying space occupation and space generation

Based on the space quality concept introduced in Section 4.2.4, and both the pitch control model and pitch value models presented in Section 4.2.3 and Section 4.2.4, respectively, we will focus on quantifying two critical concepts in soccer: space occupation and space generation. Occupying space on the field is fundamentally about a player's act of continually positioning himself in an area of high value. Specifically, we identify two types: active occupation, when the player moves at running speed to earn the space, and passive occupation, when the player is below running speed (jogging or walking). Regarding space generation, we define it as the action of dragging opponents out of certain areas to create newly available space in previously covered areas. Specifically, we identify situations where a player drags an opponent away from another teammate whom the opponent was close to initially. The dragging concept is, at its simplest, creating space for a teammate by pulling their defender towards oneself.

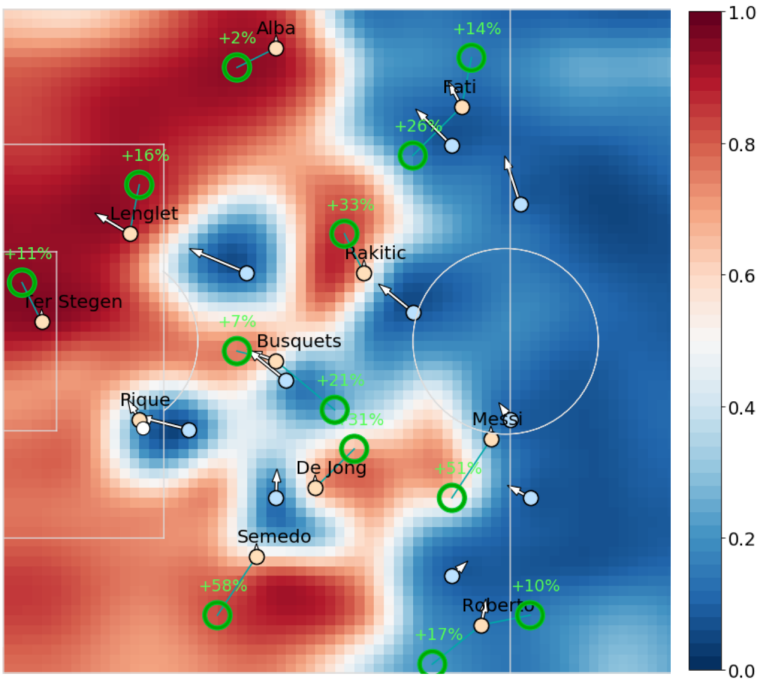


Figure 7.5: A game situation where yellow circles represent the players of the attacking team, and the blue circles the players in the defending team. The surface represents the pass probability for every location on the field. The green circles represent the optimal positioning of players increasing the expected pass probability if the players were placed in those locations at that time, and the number indicates the added probability.

From these definitions, we can derive two performance metrics: Space Occupation Gain (SOG) and Space Generation Gain (SSG), and also identify whether these are produced passively (while jogging or walking) or actively (above walking speed). Through the analysis of a first division Spanish league match, we show a handful of approaches to understand better a missing key factor for performance analysis in soccer: off-ball attacking dynamics. The quantification of space occupation gain and space generation allows us to observe Sergio Busquets' high relevance during positional attacks through his pivoting skills, the dragging power of Luis Suarez to generate spaces for his teammates, and the capacity of Lionel Messi to occupy spaces of value with smooth movements along the field, among other characteristics.

Space Occupation Gain

Employing the model that estimates the value of space ownership at any given time (represented in Equation 4.18), we can define a model for identifying gain in space occupation in time. As mentioned earlier in this section, we propose the Space Occupation Gain (SOG) concept as the relative amount of quality of owned space earned during a time window. An opposite concept is Space Occupation Loss (SOL), which relates to a negative gain during the time window. We first define the concept of gain in time G as the mean difference of quality of space occupation Q during a time window $[t+1, t+w+1]$, for a given player i. This is expressed in Equation 7.1.

$$G_i(t) = \frac{\sum_{t'=t+1}^{t+w+1} Q_i(t')}{w} \tag{7.1}$$

Given the dynamic nature of soccer, players are involved in a continuous process of winning and losing space. A small gain of space can happen when the nearby defenders follow the ball when it moves away from the player, leaving the player better control of space. However, the same can happen in a high-speed running situation between the attack and the defender, where the attacker moves slightly faster. In another case, a medium or high gain of space can happen when the player moves towards free space. Given this, it is necessary to define a level of space gain from which the earned space can be considered an actual occupational advantage and not a consequence of slower-moving contextual factors in a given situation. We set a constant ϵ as a threshold to account for space occupation gain only when the gain is above that threshold. We can do the equivalent for space occupation loss. Both expressions are defined in Equations 7.2 and 7.3.

$$SOG_i(t) = \begin{cases} G_i(t) & \text{if } G_i(t) \geq \epsilon \\ 0 & otherwise \end{cases} \tag{7.2}$$

$$SOL_i(t) = \begin{cases} -G_i(t) & \text{if } G_i(t) \leq -\epsilon \\ 0 & otherwise \end{cases} \tag{7.3}$$

Space is occupied actively when the player moves towards that space faster than a jogging pace. Otherwise, we consider that space to be occupied passively. Ric et al. (2017) define jogging pace as any speed lower than 1.5m/s.

Space Generation Gain

The generation of space for teammates is a concept that involves two or more teammates during a certain attacking situation. Two main types of actors are present: one generator and one or more receivers. The generator is a player that moves toward a certain space while dragging opponents during the process. This dragging behavior causes the freeing up of space previously occupied by the dragged opponent. When that opponent was previously close to one or more other teammates, we say those players are receiving space generated by the attracting player. In order to express this concept mathematically, we need to define a value for closeness. We will say that a player is close to another if the distance between them in a given time t is below a constant δ. Also, it is desirable to define another constant α for constraining the minimum attracting distance, which refers to the difference in distance between the starting and end position of the generator and the attracted opponent. This allows us to avoid inaccurate attractions when players are very close to each other initially. Given this, let $d_{i,j}(t)$ be the distance between players i and j, Equation 7.4 presents the necessary conditions for the concept of space generation SG between any pair of teammates (i, i') and any opponent j, for a time window $[t, t+w]$.

$$SG_{i,i'}(t) = \exists_j (d_{i',j}(t) \leq \delta) \wedge (d_{i,j}(t+w) \leq \delta) \wedge (d_{i',j}(t+w) > \delta) \wedge (d_{i,j}(t+w) - d_{i,j}(t) < \alpha) \quad (7.4)$$

Intuitively, Equation 7.4 expresses that player i is generating space for a teammate i', within a time frame $[t, t+w]$, if at time t there is an opponent j that is at least δ meters close from the teammate i', and that after $t+w$ seconds, it moves away from player i' and approaches player i closer than δ meters. Additionally, it provides the condition that players i and j got closer by at least α meters, within that time frame.

Once we can identify when a space generation behavior is occurring, we would like to focus on the cases in which we have a gain in space due to the dragging effect. Analogously to the SOG definition, we express the Space Generation Gain (SGG) as space generation situations where the gain is above a threshold ϵ, as presented in Equation 7.5.

$$SGG_{i_j}(t) = \begin{cases} G_j(t) & \text{if } SG_{i,j}(t) \wedge G_j(t) \geq \epsilon \\ 0 & otherwise \end{cases} \quad (7.5)$$

Essentially, we attribute space gain to a player when a defender leaves his mark and moves towards a teammate, subject to the conditions that the defender was close to the player and ended close to the teammate during a time window. It is essential to clarify that while SOG and SGG represent two frequent and relevant cases of space gain within soccer, other types of situations and movements might contribute to the total space created by a player during a match. An additional possible concept is that of potential space, referring to a space that the player is more likely to reach, within his positioning, but not in his immediate influence area. We will now focus on analyzing SOG and SGG within a match context.

Match analysis through space creation

The ability to create and occupy spaces are two commonly trained concepts in modern soccer. During training, coaches interrupt and reshape individual drills to teach players how to orient and move toward spaces and away from low-value local zones on the field. When analyzing off-ball performance, coaches appeal to video analysis. Although elite soccer analysis staff typically have a great capacity to understand complex concepts through match visualization, the dynamics of space creation are so frequent and happen in such short time windows that it becomes impractical for video analysts to grasp them all, even for a single match. However, it is important to note that there is no existence of ground truth data regarding the quantification of spaces in soccer. Hence we have performed an extensive validation of the developed concepts through video and studying individual situations within games, with the help of two expert soccer video analysts from F.C. Barcelona, to fine-tune our quantitative approach. The following videos are examples of the video-based validation tool we have used: http://www.lukebornn.com/sloan/space_occupation_1.mp4, http://www.lukebornn.com/sloan/space_occupation_2.mp4

Based on this, we provide a complete summary of off-ball movement statistics for a specific Spanish first division official match between F.C. Barcelona and Villareal F.C. in January 2017. Specifically, we provide an analysis focused on the concepts of space occupation and space generation, using Metrica Sports optical tracking data. This match ended with a 1-1 result, where Villareal F.C. scored the first goal at the 49th minute (second half), and the F.C. Barcelona equalizer came at the 90th minute by Lionel Messi. Situationally, this presents a game where F.C. Barcelona was required

to score during the final minutes and occupy and generate the most spaces possible to reach scoring chances. To identify space occupation and generation actions, we calculate for the attacking situations of F.C. Barcelona all the instances where a player had controlled possession of the ball with his feet. From each of those situations, and alongside expert football analysts from F.C. Barcelona, we define a window w of three seconds after each of these cases, reaching a total of 845 different situations. The closeness factor δ is set to 5 meters, based on the minimum distance an opponent is on average to a player in possession of the ball. We also set the minimum attraction distance for space generation α to 3 meters.

Table 7.1 and Table 7.2 present the space occupation statistics for F.C. Barcelona, sorted in descending order by the total amount of Space Occupation Gain (SOG). At first glance, it can be seen that over 41% of gain of space occupation was performed by Iniesta, Sergio Busquets, and Lionel Messi. Notably, these three players occupy different positions and have different roles within the team. Busquets is a pivot and has a specific role of helping to drive the ball with controlled possession during build-ups and accompanying the game creation during positional attacks. Iniesta is an attacking midfielder with significant control of the ball and exceptional skills in moving and finding spaces between lines. Messi is an attacker but not attached to a specific position and can cover broad areas of the pitch to find space and request the ball. However, the three players share a long-time tradition of possession-centered and off-ball movement quality during their careers. Suarez and Neymar, two highly mobile players, appear with a lower count of situations where space was gained. This can be associated with the high level of strictly closed marking these players suffered during the match.

It is interesting to observe that for most players, the active occupation of spaces is considerably more frequent than passive occupation. This is particularly noticeable on the left and right backs, Digne and Sergi Roberto, who need to cover more expansive spaces and show a high mean distance to the ball for SOG, a characteristic shared by central defenders Pique and Mascherano. A remarkable case is Lionel Messi, whose passive SOG is considerably higher than the active one. The passive characteristic of SOG does not mean the player is not occupying the space intentionally, but rather that he is not moving at running speed but slower. Much has been argued in recent years about several moments during matches where Messi walks through zones of the field. However, that walking behavior is not a detachment from the match but a conscious action to move through empty spaces of value,

Table 7.1: Statistics of space occupation for F.C. Barcelona in an official Spanish League match against Villareal F.C. Symbols #, $\sum$ and μ represent the total, sum, and mean of their associated variable. SOG refers to Space Occupation Gain, and Active (%) and Passive (%) the player percentage of times space was occupied through active or passive occupation.

Name	# SOG	$\sum$ SOG	μ SOG	Passive (%)	Mins
Iniesta	96 (14.8%)	15.77	0.16	43.75	94.86
S. Busquets	90 (13.9%)	14.85	0.16	52.22	94.86
Messi	81 (12.5 %)	14.72	0.18	66.67	94.86
A. Gomes	74 (11.4%)	12.58	0.17	31.08	68.61
Suarez	70 (10.8%)	12.27	0.18	42.86	94.86
Neymar	61 (9.4%)	9.46	0.16	40.98	94.86
S. Roberto	51 (7.8%)	7.34	0.14	21.57	94.86
Pique	29 (4.4%)	4.92	0.17	51.72	94.86
Mascherano	29 (4.4 %)	4.54	0.16	58.62	94.86
D. Suarez	22 (3.4%)	4.07	0.18	22.73	26.25
A. Turan	17 (2.6%)	3.51	0.21	47.06	23.32
Digne	26 (3.2%)	3.48	0.13	19.23	71.54

Table 7.2: Statistics of space occupation for F.C. Barcelona in an official Spanish League match against Villareal F.C. FRT and BEH indicate the amount of times SOG occurs in front or behind the ball. MBD represents the mean ball distance

Name	FRT	BEH	MBD	Mins
Iniesta	49	47	15.19	94.86
S. Busquets	44	46	16.65	94.86
Messi	58	23	17.50	94.86
A. Gomes	40	34	15.93	68.61
Suarez	57	13	13.46	94.86
Neymar	48	13	18.31	94.86
S. Roberto	25	26	25.10	94.86
Pique	6	23	21.05	94.86
Mascherano	2	27	22.03	94.86
D. Suarez	13	9	17.47	26.25
A. Turan	12	5	12.71	23.32
Digne	13	13	16.23	71.54

claim the control of valuable space, and ultimately the ball. Messi does this very effectively, placing him near the top of players in terms of space gained during the match, despite the lack of active gain. A relevant characteristic of this is that 71% of the time the gain in space is done in front of the ball rather than behind. The in front and behind the ball statistics show a clear tendency for central defenders to gain space behind the ball, while attackers show a higher rate of space gain in front of the ball. Noticeably Busquests, Iniesta, and the right and left backs (Digne and S. Roberto) have a balanced ratio of space gain behind and in front of the ball.

Table 7.3 and Table 7.4 presents the statistics Space Generation Gain (SGG) and for Space Occupation Loss (SOL), respectively. The SOL statistics show a clear tendency of higher space loss for players that are more often in possession of the ball, such as Iniesta, Messi, Neymar, and Suarez. The space loss can be directly associated with pressure by the opponent, who tends to increase density near attacking players to reduce their range of action, especially for highly skilled players. Regarding the generation of space, we obtain a different picture from the space occupation skills. Here, Neymar and Suarez appear to be, alongside Messi, the players who often drag opponents to create space. With a 4-3-3 system and high-quality players, a specific attacking strategy is to spread out attacking players to drag defenders out of position and provide wider spaces for attacking action. Busquets, a pivoting specialist, also appears at the top of the table, showing his value in supporting space creation. Notably, the left and right back, Digne and S. Roberto, do not generate much space. Given that they move towards the sides of the field, it is less likely that back defenders drag opponents.

A more detailed perspective of space generators and receivers is presented in Figure 7.6. Here we can observe the number of times generators produce space for receivers and discover some collaborative playing behavior. First to observe is that Busquets receives space from most of the players at least once, possibly showing his ability to stay at the center of the play. A renowned skill of FC Barcelona is the third-man pass, which consists of the following: if a player A wants to pass to player C but is marked, he passes to player B, dragging the opponents toward him, enabling C to receive the ball in more space. This plot might show a third-man behavior through Busquets. Notably, Suarez, Neymar, and Messi generate space commonly for each other, especially Suarez, who provides considerable space to both. A special connection between Suarez and Messi is also shown for this game, where both were able to generate a high amount of space for each other.

Table 7.3: Statistics of space generation for F.C. Barcelona in an official Spanish League Match against Villareal F.C. Symbols #, $\sum$ and μ represent the total, sum, and mean of their associated variable. # Generated and # Received indicate the total times a player generated or received generated space, accompanied by the team-relative percentage. SGG refers to Space Generation Gain

Name	# Generated	# Received	$\sum$ SGG	μ SGG	Mins
Neymar	28 (18.9%)	6 (4.1%)	5.97	0.21	94.86
Suarez	25 (16.9%)	18 (12.3%)	5.60	0.22	94.86
Messi	22 (14.9%)	24 (16.4%)	4.32	0.20	94.86
S. Busquets	15 (10.1%)	24 (16.4%)	3.83	0.26	94.86
Pique	14 (9.5%)	9 (6.2%)	3.66	0.26	94.86
Iniesta	13 (8.9%)	21 (14.4%)	2.62	0.20	94.86
A. Turan	8 (5.4%)	7 (4.8%)	2.26	0.28	23.32
S. Roberto	7 (4.7%)	2 (1.4%)	1.55	0.22	94.86
A. Gomes	9 (6%)	18 (1.2%)	1.49	0.17	68.61
Mascherano	5 (3.4%)	9 (6.2%)	0.80	0.16	94.86
D. Suarez	2 (1.4%)	8 (5.5%)	0.46	0.23	26.25

Table 7.4: Statistics of space occupation loss for F.C. Barcelona in an official Spanish League Match against Villareal F.C. Symbols #, $\sum$ and μ represent the total, sum and mean of their associated variable. SOL refers to Space Occupation Loss

Name	# SOL	$\sum$ SOL	μ SOL	Mins
Neymar	51	-8.53	-0.17	94.86
Suarez	52	-9.12	-0.18	94.86
Messi	68	-11.61	-0.17	94.86
S. Busquets	38	-6.16	-0.16	94.86
Pique	19	-2.77	-0.15	94.86
Iniesta	75	-11.79	-0.16	94.86
A. Turan	8	-1.29	-0.16	23.32
S. Roberto	31	-4.62	-0.15	94.86
A. Gomes	44	-6.25	-0.14	68.61
Mascherano	23	-3.39	-0.15	94.86
D. Suarez	16	-3.14	-0.20	26.25

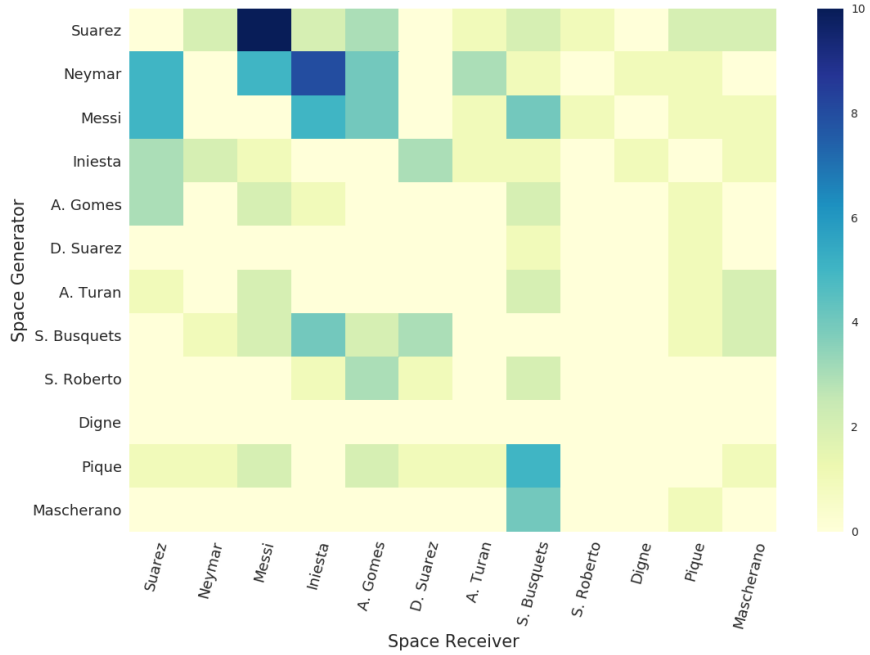


Figure 7.6: A heatmap showing the total times space was generated by generators (y-axis) for receivers (x-axis)

A different vision of space gain and generation can be grasped from Figure 7.7. Here we present the spatial heatmap for SOG and SGG situations. At first glance, we observe that the amount of space gained through occupation is considerably higher than through generation, a more complex process. Iniesta presents an interesting case where he can generate more space next to the left sideline of the field, while he is better at gaining spaces for himself at the interior of the field. Also, he produces a notable amount of space near the box. Busquets shows relevant collaborative behavior by generating space almost anywhere around the field. He also presents broad areas of SOG, but more intensively near the midfield, his natural habitat. Suarez presents a notable ability to generate space within the box, where he concentrates most of his generating contribution. Here he arises as a specialist in dragging defenders while making spaces for himself or while generating spaces for others. Messi also shows excellent ability in generating spaces around the attacking zones of the field, while Neymar concentrates on the left-wing, focused on high-speed diagonal runs towards the box. Defenders, as expected, show a minimal generation of space.

7.3 On-ball performance

7.3.1 Evolving passing networks

Passing network plots representing the passes between players and their average location in a given match, are arguably one of the most popular visualizations in soccer analytics (Knutson, 2018; McHale and Relton, 2018; Buldú et al., 2018). The usual passing network visualization consists of the average location of every player on the field at the time where they took the pass, a player circle size according to frequency of passes taken by the player, and lines between players with size and color related to frequency of passes. More sophisticated versions assign passes the xG value calculated at the end of the attack, with the limitation that all the passes within the same attack receive the same value (Knutson, 2018). While passing networks are useful tools to understand frequency of passes between pairs of players, they fail to recognize whether those passes added or subtracted value, and the distribution the EPVA of those passes. We introduce here an evolved version of passing networks that incorporates EPV-related metrics, in order to better evaluate passing quality.

Figure 7.8 presents a passing network for all passes departing from Rakitic in a single FC Barcelona match in La Liga season 2017-2018. Only

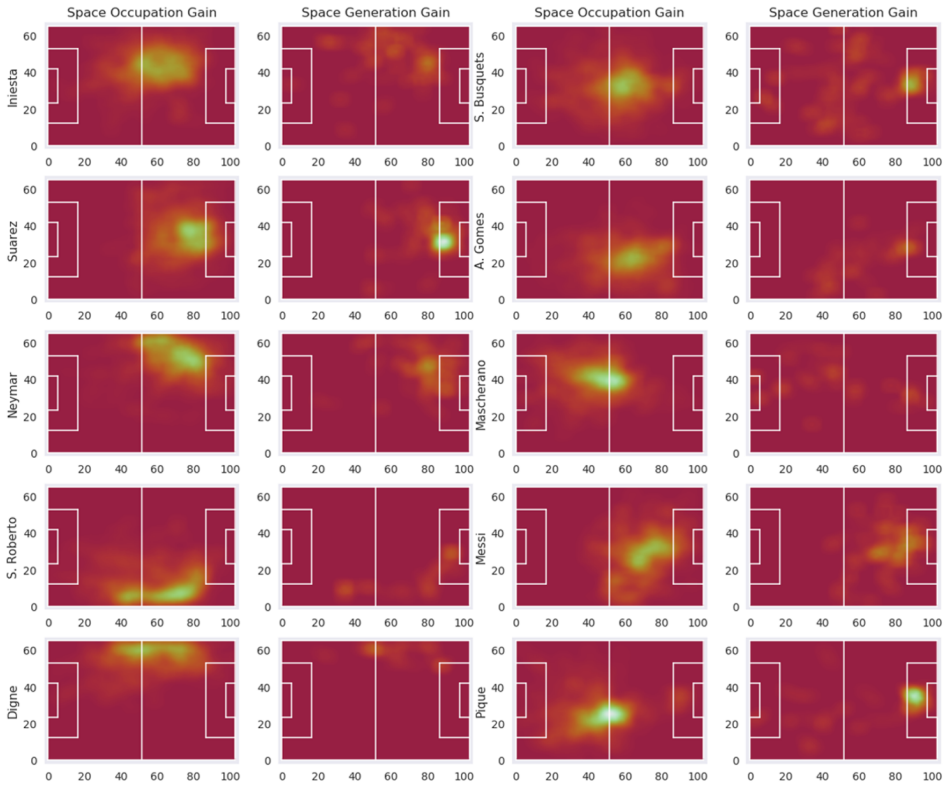


Figure 7.7: Space Occupation Gain and Space Generation heatmap for every field player playing over 60 minutes. The scaling factor is based on the maximum Space Occupation and maximum Space Generation among all the team, respectively

one player is shown for simplicity. This passing network introduces several concepts. Players are located at their mean location for all the passes received. The size of a player's circles is proportional to the mean pitch control they had when the pass was taken, where smaller sizes represent lower space control, thus high pressure on the player. Here we can see the central defenders are receiving balls from Rakitic with low pressure, while the wingers and forwards are under considerably higher pressure (Suarez in particular). The circles and lines are colored according to the EPVA of the passes between Rakitic and each corresponding teammate. Circles are colored with the mean EPVA by passes received by the player (except for Rakitic). The way the lines are colored deserves special attention. Since the contextual and spatial characteristics of the game state are expected to influence the value of passes, coloring based on a sole summary statistic

such as the mean or median EPVA can produce a considerable loss of information. In order to get a closer view into the passing distribution, arrows are divided into three equally sized blocks. Each block is colored according to the EPVA of the passes in the percentile .25, .50 and .75 respectively of each corresponding distribution of passes between players. Having this, we can observe that while passes to Suarez incorporate a wide range of results (from 0.01 to 0.12), the top 25% of passes where of great value. Also, we can observe that the top 25% of passes to Piqué provided over 0.05 EPVA. The plot also shows the average location of each pressure line when Rakitic attempted a pass, so the locations of the teammates also provide information about their average location relative to the opponents' formation lines. More detailed information can be obtained from the plot when filtering for specific situations, such as organized buildups phases (e.g. passes starting from behind the first pressure line) or creation phases (e.g. passes taken above the midfield and beyond the first pressure line).

7.3.2 Inspecting into passing tendencies relative to context

In Section 4.4 we explored the idea of contextualizing actions based on the relative location according to the mean position of pressure lines at a given time. Here, we leverage that concept to provide a contextualized view into the passing performance of a set of players in a given match. Figure 7.9 presents a comparison of the EPVA of passes between relative locations for two central defenders (Sergio Ramos and Gerard Piqué) and two midfielders (Luka Modric and Sergio Busquets) for a FC Barcelona vs Real Madrid match, in La Liga season 2018-2019. Each column groups passes taken from the red colored area shown on the first row. Each of the four areas (that we call Z1, Z2, Z3 and Z4) represent the space between the own goal and the first pressure line (Z1), the space between the first pressure line and the second (Z2), the space between the second and the third pressure line (Z3) and the space between the third pressure line and the opponents goal (Z4). Notice that these zones are not predefined but they move dynamically according to the opponents location at every time step. Each plot shows a stacked bar chart representing the same concept as arrows in Section 7.3.1. That is, the distribution of passes is split into three equally sized groups (favoring the top in case of not exact division by 3) and colored according to the EPV added after the pass. The x-axis locations represent the destination of the pass, and indicates whether the ball stays in the current relative zone or it goes back or forward to any other zone.

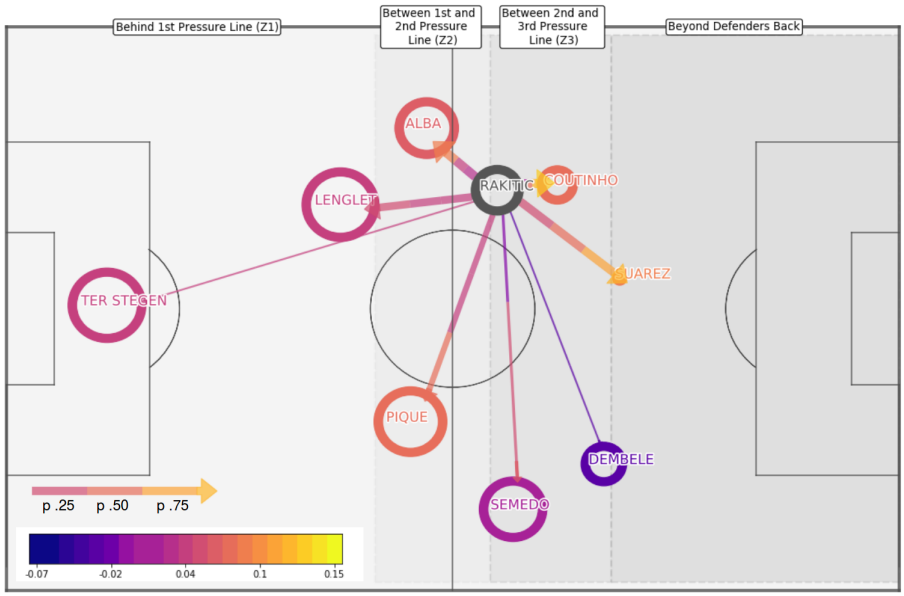


Figure 7.8: Passing network for all passes of Rakitic in a 2017-18 FC Barcelona match. Circles are located in the mean pass destination location for every other player, while the circle representing Rakitic's location is placed at the mean location where passes were taken. Circle size is related to the pitch control the player had when making the pass, where smaller means less space (higher pressure). The color of the circles represent the mean added value of those passes. The lines are split into three equally sized blocks. Each block (starting from Rakitic) is colored by the added value associated with the percentile .25, .50 and .75 in the distribution of passes between those two players. The gray areas represent the space between the mean pressure lines of the opponent.

Let's analyze this figure column by column. For actions behind the first pressure line (Z1) we observe two different pass tendencies from central defenders (Ramos and Piqué). Ramos had a higher tendency to overcome the first line of pressure towards the second but this resulted most of the time in a loss of value for the possession. Piqué showed a higher tendency to keep possession value by passing behind the first pressure line, but he also attempted to overcome the first and the second pressure line. His three attempts to overcome the first line of pressure successfully added value to the possession. The second column represents passes starting between the first and second pressure line, typically during the progression phase of the possession. For the central defenders, we observe again two different tendencies. Piqué passed back to Z1 twice as much as Ramos, both with the tendency of losing an average of 0.01 EPV in their possession when passing behind this line. When keeping the ball in Z2 Piqué was able to increase the EPV of the possession while values for Ramos in that zone were considerably lower, providing a hint for different types of pressure received by each team. Regarding the midfielders, we can see the need for analyzing the distribution of passes instead of jumping directly to summary statistics. For both midfielders, two thirds of the passes subtracted or added little EPV to the possession, however, the last third of passes where able to add value to the possession. Remarkably, Busquets was able to attempt successfully two passes beyond the defenders back adding a considerable amount of value. For passes starting between the first and second line (Z3) we have different situations. Here, the contribution of central defenders was very low for Ramos and non-existent for Piqué. Ramos was not able to add value through passing while remaining in this zone, however his presence in this zone adds more information to the match analysis regarding the tendency of the defender to contribute with the attack. For both midfielders passes within the same zone presented a consistent loss of value, which shows the difficulty of adding value in the relative zone in soccer where the highest pressure is found. Remarkably, passing back to Z2 found added value in the third of the cases for both Modric and Busquets, but subtracted value when they went back to Z1, showing again the changing dynamics of soccer according to context.

7.3.3 Calculating optimal passing locations

In Section 3.1.7 we argue that one of the main concepts to be considered when designing and EPV framework is capturing the idea that, in soccer, passes con go anywhere on the field. Both the development of the Soccer-

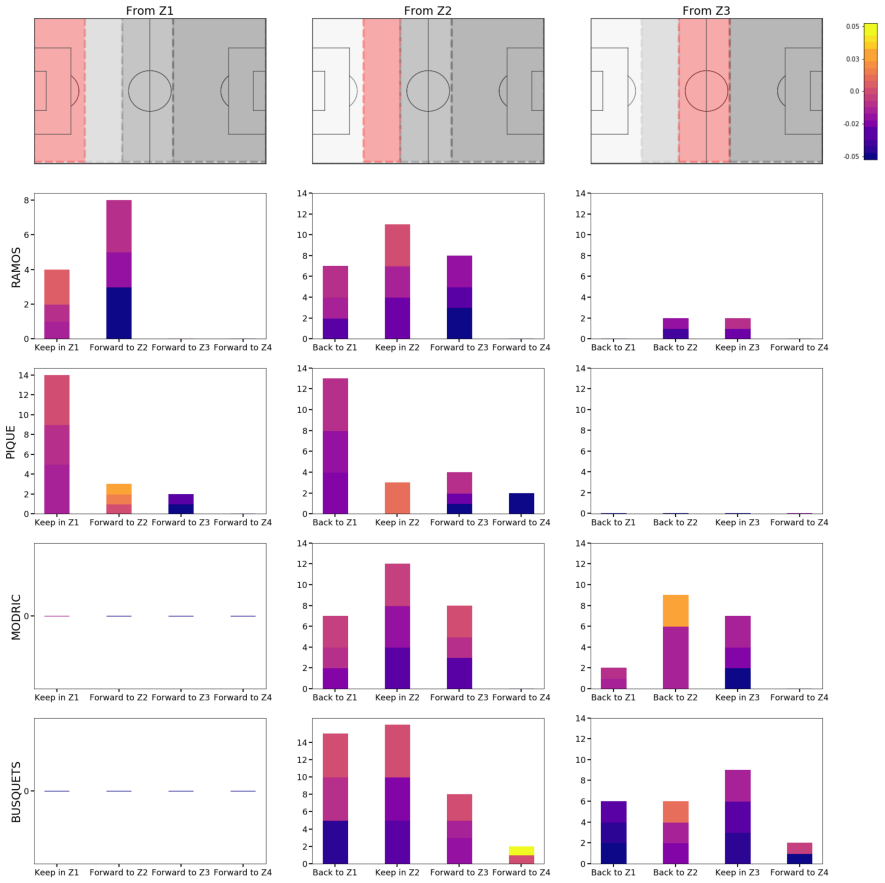


Figure 7.9: The image presents the distribution of EPVA for passes between relative zones for four different players in a FC Barcelona vs Real Madrid match of La Liga season 2018-2019. Columns group passes taken from three relative zones: behind the first pressure line (Z1), between the first and second pressure line (Z2) and between the second and third pressure line (Z3). The stacked bar charts represent the frequency of passes and are split into three equally sized groups. From bottom to top the color of the bars correspond to the EPV added of the actions at percentile .25, .50 and .75 respectively. X-axis in the bar charts represent the destination of the pass, including a fourth zone (Z4) corresponding to the space between the third pressure line and the opponents goal. The direction of the attack goes from left to right.

Map architecture, as well as the implementation of the decomposed EPV model (presented in Chapters 5 and 6, respectively) provides the necessary components to evaluate which passing destination locations passes provide greater impact. Following a similar approach to that presented in Section 7.2.2, we can exploit the prediction of passing probability surfaces to assess the optimal destination location for any potential pass, to ensure the pass is completed successfully.

Given a game-state, where a player is in possession of the ball, we define the optimal and sub-optimal pass destinations as the locations near the teammates than provide a higher pass probability than their current location. To obtain the optimal passing locations we follow a similar procedure to the one presented in Section7.2.2, where a series of alternative game situations are obtained by shifting each player's position within a 5×5 grid, adjusted to player's velocity. For each of these situations we obtain the expected success probability of each potential pass, from which we select the optimal and a set of sub-optimal passing locations. Figure 7.10 presents in red circles the set of best passing locations for each of the possession team players for a given game state. This kind of visualization provides a coach the ability to perform a direct visual inspection of passing options and allows her to provide direct feedback to players about specific game situations, improving the coach's effective communication options. We can observe that the optimal passing location do not always coincide with the optimal location for the player presented in Section 7.2.2. Here we can observe that the optimal passing locations from Piqué to Semedo are both in a straight line between both, and consider either passing in short to bring Semedo to Piqué's location (71%) or making a longer pass behind Semedo's back, to avoid the pressure of the nearest opponent. In the cases of both Lenget and Busquets, we can see the optimal locations follow the player's velocity, specially considering the higher speed of these compared to the other teammates. We can also observe that the optimal passing location for Rakitic is influenced by the velocity vector of two closest opponents; one that is quickly moving away, opening a space behind his back, and the other that is quickly approaching Rakitic to press him and reduce his passing success probability. Having this information we could automatically locate game situations where players select passing locations that are considerably below the optimal, and provide the coach with specific information about alternative passing locations. If a certain behavior pattern is detected (for example, consistently passing back or being unable to find the playmaker) the coach could use this visual representation and the associated video to

instruct to the players.

Note that for all of the passes (with the exception of the pass to the goalkeeper) the optimal location is rarely found in the player's current location. Also, the location and velocity of the opponents is shown to be determinant factor for the success of passes. Both observations, provide an idea of the complexity of the spatial dynamics of soccer, and the necessity of considering the full extent of the soccer field, as well as the dynamics of the 22 players and the ball.

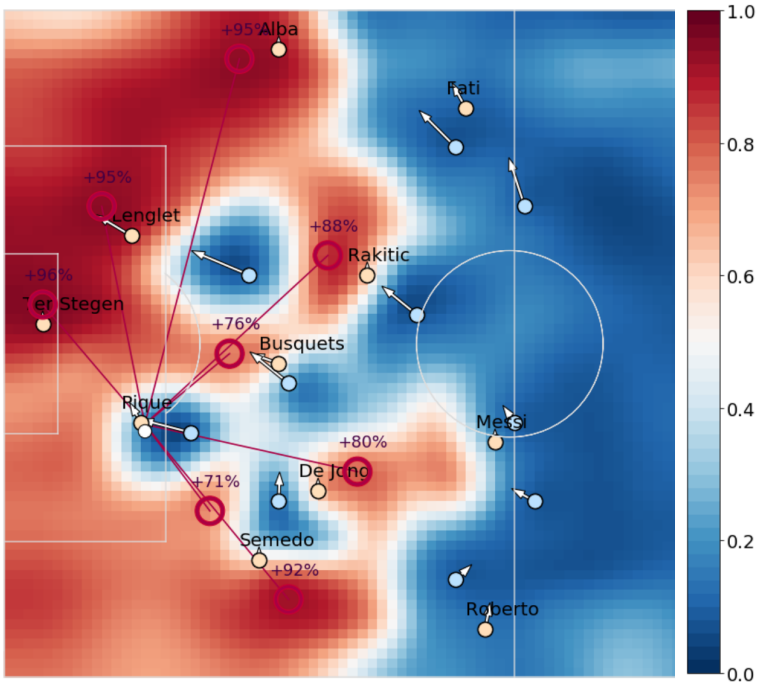


Figure 7.10: A game situation where yellow circles represent the players of the attacking team, and the blue circles the players in the defending team. The surface represents the pass probability for every location on the field. The red circles represent the optimal and sub-optimal (when available) locations maximizing pass probability of potential passes to each of the players. The number above the red circles indicates the calculated pass probability.

Assessing individual passing skill

Following the identification of optimal passing location presented in Section 7.3.3, we propose a new metric pass completion added (PPA) to quantify a player's passing skill. For each observed pass, we calculate the player's passing accuracy, weighted by the difference between the probability of the optimal pass and the probability of the selected pass. This metric is formally defined in Equation 7.6, where S and M are the set of successful and missed passes, respectively, $\hat{y}$ is the optimal pass probability, and y is the selected pass probability. Intuitively, a player reward is discounted if the selected pass was not optimal. In the case of the pass being unsuccessful, the player is only penalized in proportion to the probability difference with the optimal location, rewarding the player's pass selection.

$$PPA = \sum_{s=1}^{S}(1-\hat{y}^s)(1-(\hat{y}^s-y^s)) - \sum_{m=1}^{M}(\hat{y}^s)(\hat{y}^s-y^s) \tag{7.6}$$

In table 7.5 we present the best ten players in pass completion added for the 2014-2015 season of the EPL, where The cumulative PPA of a player is normalized by 90 minutes played. The table includes the players' estimated market value in 2014, provided by www.transfermarkt.com. We can observe that the list contains a set of the best players in recent times in this league, including creative midfielders such as Oezil,Silva, Hazard and Fabregas, deep creative wingers such as Navas and Valencia, and Rosicky, a historical player.

Table 7.5: Ranking of the best ten players in pass completion added for the season 2014-2015 of the EPL.

Team	Player Name	PPA/90m	Age	Market value
Arsenal	Mesut Oezil	0.0578	24	€45M
Manchester City	David Silva	0.0549	28	€40M
Chelsea	Eden Hazard	0.0529	23	€48M
Manchester United	Antonio Valencia	0.0502	29	€13M
Arsenal	Tomas Rosicky	0.0500	33	€2M
Chelsea	Cesc Fabregas	0.0484	27	€40M
Arsenal	Santi Cazorla	0.0470	29	€30M
Manchester City	Jesus Navas	0.0469	28	€20M
Manchester City	Yaya Toure	0.0466	30	€30M
Manchester City	Samir Nasri	0.0447	26	€22M

Chapter 8

Conclusions and future work

This chapter presents the conclusions of this work, emphasizing the most relevant contributions concerning the research questions posed. Additionally, we detail the main limitations of this approach and propose a series of ideas for future work.

8.1 Conclusions

The primary motivation of this work is focused on the current difficulty of integrating data analytics within professional soccer clubs, especially given the complexity of soccer's spatiotemporal relationships and the challenge of translating findings to practitioners. While applied research in soccer has covered a broad set of topics such as evaluating the impact of individual actions, measuring spatial dominance, or estimating player's motion dynamics, we identify that most of these correspond to isolated aspects of the game. Given this, we devote this work to designing and developing an analysis framework that could capture relevant spatiotemporal relationships and provide interpretable and visual explanations of the outcomes.

By developing the concept of EPV, we provide a comprehensive analysis framework capable of producing calibrated estimates of the long-term expected value of any game situation. We propose a decomposed approach from a modeling standpoint where a series of foundational soccer components can be estimated separately and then merged to provide a single EPV estimation, providing flexibility to this integrated model. From an inference and implementation standpoint, we leverage several function approximation approaches within the field of machine learning and statistical analysis

to exploit complex relationships in spatiotemporal tracking data. We design deep learning architectures to exploit both low-level spatiotemporal data and soccer-specific features for producing calibrated predictions in a wide range of components, from action-selection probability to action success probability and expected value. An essential contribution of this work is the proposal of SoccerMap, a flexible deep learning architecture capable of producing accurate probability surfaces and accurate estimates in a broad range of problems. Based on the large set of spatial and contextual features developed in this work, and the availability of low-level spatiotemporal tracking data from professional matches, we show we are able to model each of the components of the decomposed EPV approach successfully and provide accurate EPV estimates at any time instance. The flexibility and interpretation capabilities of the proposed model allows one to produce a broad set of examples of practical applications in three different fields of interest within soccer analytics: on-ball performance, off-ball performance, and match analysis. Additionally, we provide a series of practical applications derived directly from the developed spatial and contextual features. Through this modeling process, we achieve a functional framework for addressing a comprehensive set of problems of interest within the non-stop flow of questions that usually arise in the relationship between data analysts and coaching staffs. Based on this model, future developments can incorporate a set of improvements for performing more sophisticated analysis of unexplored soccer problems. Some of these improvements include learning from more recent and larger datasets, incorporating match-specific and team-specific information, adding kinematic information such as pose estimation derived from video footage, considering richer physical data, or adding ball height, among many others.

In the following sections, we address each of the research questions presented in Chapter 1.2, which constitute the focus for the design and development of this work, highlighting the main conclusions related to the findings and contributions of this work.

8.1.1 Expected possession value

The fundamental research question of this thesis is the following: "can we estimate the expectation of a team scoring or conceding the next goal, at any time in the game (EPV)?" Through the results of the joint EPV model implementation, we are able to verify that it is possible to estimate the expected value of any given ball-possession in the long term in a way that it

produces calibrated estimates and that the visual representation of results matches the understanding of professional coaches. One of the differential aspects of this approach is the definition of soccer possessions, where we consider the fluid nature of the game, allowing the ball to change control between teams an indefinite number of times until a goal, an objective metric of success is observed. An additional critical decision is the joint modeling of the probability of conceding and scoring goals. This allows the model to capture non-intuitive characteristics of the game such that risking to lose the ball near the opponent's box produces a positive expected reward on average. In a broader perspective, considering both goal scoring and conceding probabilities provides the ability to perform a more direct and fine-grained evaluation of the risk and reward of actions, especially for potential or non-observed actions. A limiting factor is the consideration of a maximum time threshold before observing a goal. Naturally, the sparsity of goals, compared with the long duration of soccer matches, provides a time limitation for the influence that any given individual action has on future outcomes. The instantaneous estimation of EPV itself can be used directly in a broad set of practical applications, including the evaluation of the frame-by-frame expected goal probability during possessions, identifying the most valuable actions, the evaluation of a player's on-ball and off-ball performance according to the added value by observed actions, quantifying a team's attacking and defensive production (e.g., the sum of last actions' EPV in attacks and defenses), and even the generation of either match-specific or cross-season highlight reels.

8.1.2 A decomposed EPV approach

In this work, we argue that while the EPV estimates can be applied directly to solve a broad set of practical needs, single-model approaches lack flexibility and interpretability, two fundamental elements for successfully implementing such a model within a coaching staff's day-to-day work. Considering the possibility of modeling EPV in a more granular way, we pose the question: "can we express this expectation (EPV) in terms of a series of smaller components, akin to coaches' language so that they can be estimated and interpreted separately?" We present a new approach for modeling EPV consisting of decomposing the primary expression into a series of subcomponents by following the law of total expectation. To apply this decomposition, we decided to define a series of essential building blocks representing concepts familiar to coaches, modeled and calculated separately, and then joined to produce a single EPV expectation. Specifically, we consider a reduced

set of three actions: passes, ball drives, and shots, to encompass the most significant number of types of on-ball actions in soccer. This approach provides a comprehensive analysis framework, where each of the components can be inspected, adapted, and extended separately. Additionally, this allows one to conduct a deep exploration into the factors that influence the variations of EPV in any given situation and develop a considerably more extensive set of practical applications. We consider that the idea of decomposing and estimating an expression in this way is a valuable contribution to the development of more interpretable machine learning models.

8.1.3 Obtaining calibrated estimations

To provide the ability to perform fine-grained analysis on each component's impact, we must grasp the uncertainty related to the occurrence and success of the different events. An example of this is evaluating the risk and reward of actions, which would benefit from a deeper understanding of the actions' probability and expected value. From a practical standpoint, we are more interested in providing a sense of the likelihood of an event rather than classifying it into discriminant classes. For example, the probability of scoring a goal becomes more useful and granular information from a practical standpoint than strictly predicting if the shot will produce a goal or not. Following this, we express the components so that their outcomes provide estimates for either the probability or the expected value (depending on the component) of the concept they represent. Given that these probabilities constitute the pieces of information to be delivered and used in practice, it becomes necessary to ensure that each of the models provides calibrated estimates. This reasoning lead to the following research question: "can the models built for these components produce calibrated probability estimates?"

In the implementation of the EPV framework, presented in Chapter 6, we show the models are able to produce calibrated estimations. These are validated in two ways. First, calibration reliability curves are provided, showing that the average observed outcome coincides with the average probability by bin for a set of ten equally-sized bins of the model's predictions. Additionally, we report a quantitative calibration value for each model based on the ECE metric, allowing us to compare these results with future research. A noticeable result of this work is that calibrated probabilities could be obtained in a set of different learning scenarios, from fully convolutional network-based architectures, which have a high number of parameters, to

boosting regression models. Remarkably, the SoccerMap architecture produces calibrated results in three different problems, despite the challenging learning set-up posed by the single-location label restriction, described in Chapter 5. Most interestingly, we show that ensuring calibrated outcomes on each of the separated components is sufficient for producing a calibrated joint EPV model.

8.1.4 Development of spatial and contextual features

An essential part of this work is the development a broad set of spatial and contextual features aimed to provide a richer representation of game states. The development of these features was supported by a broad set of professional coaches, whose experience in-game analysis produced a series of suggestions that influenced the design and implementation of these features. While the experience of these coaches is considered a valuable aspect for the mentioned features' design, an important question is whether these features would influence the models developed. Specifically, we propose the following research question: "can soccer-specific features developed with experts contribute to the models' estimations?"

The influence and usefulness of these features is presented in three ways: by evaluating feature importance through the SHAP methodology, applying some of these features for providing contextual information within EPV-derived practical applications, and showing practical applications directly derived from these features. Through the SHAP value analysis, we can see how concepts such as pass, ball drive, and shot selection probability are highly influenced by the pitch control and opponent density features. Additionally, spatial features such as block count and goalkeeper surpassed influence the estimation of the expected value of shots. The detection of dynamic pressure lines play a significant role in this thesis, both as a feature within the models and as a useful variable for deriving practical applications directly. The dynamic lines of both the attacking team and opponent are used in the expected value models for passes and ball drives, providing location-relative contextual information, and enriching the game state representation.

Specifically, the dynamic formation lines provide the models with information absent in standard low-level spatiotemporal data and comprehending a series of complex spatiotemporal dynamics. For example, the value of a potential pass now considers if that pass breaks the last pressure line of the

opponent and if the pass is being taken from the first formation line (typically from a player in the center-back position) or the outside or inside of the opponent's block. In Section 4.4.4, we show that the expected value of passes varies considerably according to this contextual information. Moreover, we present a series of comprehensive practical applications directly obtained by conditioning both the quantitative and visual information to the dynamic formation lines information. An example of this is detecting and comparing players' passing tendencies relative to the opponents' positioning. We show how the passing selection and reception dynamics interpretation is enriched when adding this contextual information. Another practical application is identifying game phases, which provides higher granularity for understanding the changing dynamics within a single ball possession, and can serve as a sophisticated method for structuring the analysis of matches.

We also show how the pitch control and pitch influence models could be exploited to develop useful practical applications directly. We present two novel concepts in soccer analytics, the quantification of space occupation and space creation, two off-ball qualities that are omnipresent in coaches' language. From these concepts, we derive a series of metrics for providing a more in-depth evaluation of the spatial performance of players and the value they provide to teammates. This approach itself provides a new way of evaluating players' off-ball performance, a little-studied concept in soccer analytics, and that becomes particularly valuable given the vast amount of time that players spend without ball control. Beyond these specialized features, the series of more natural spatial features such as distance, angle, and velocity, constitute a fundamental piece of information in all the models developed.

8.1.5 SoccerMap: producing visually-interpretable outputs

While most existing research to date focused on developing models on top of carefully designed features, we consider exploring if rich features could be learned autonomously from low-level spatiotemporal data by employing approaches that could exploit the spatial relationship of nearby locations. Additionally, if such a model could produce visual representations of the predictions, we could have a more granular and intuitive way to translate the results of complex models to coaches in a more digestible way. Based on this, we propose the following research questions: "can we develop a model capable of ingesting raw tracking data and producing probability surfaces in a way that is easily adaptable to other problems? Can this model be

developed through a spatial-aware deep learning architecture?"

In this work, we introduce SoccerMap, a fully convolutional neural network-based architecture capable of receiving low-level tracking data and producing probability surfaces representing predictions at every location on the field. We show that SoccerMap can be easily adapted to produce accurate estimates of different challenging problems such as pass probability, pass expected value, and pass selection likelihood. The architecture can receive an arbitrary number of layers, from sparse matrices with low-level information (e.g. location of the defending players), to more carefully designed dense matrices (e.g., defensive pressure lines corresponding to each location on the field). Additionally, we show we can successfully train this architecture with single-location labels and produce entire probability surfaces with accurate and calibrated estimates. The remarkable performance of the presented SoccerMap-based models, highlights the capacity of the architecture to exploit low-level spatiotemporal data to produce representative features for each problem. In other words, the results of these models suggest that SoccerMap is capable of making sense of the most relevant complex spatiotemporal dynamics to produce accurate predictions.

The flexibility of SoccerMap and the visual interpretability of the outcome it produces make this architecture a useful tool for quickly learning complex problems with little specification and exploring its results with the aid of practitioners with a non-scientific background. The production of probability surfaces is still a little explored area in sports analytics. With SoccerMap, we show that these surfaces are usually more easily interpreted than standard machine learning approaches but that these open the door for a deeper exploration of off-ball performance and the analysis of the impact of potential (non-observed) actions, a critical aspect of performance analysis. This architecture could be directly applied in other team sports that count with the availability of tracking data, such as basketball, American football, handball, among many others, by performing slight modifications to its configuration.

8.1.6 A large set of novel practical applications

One of the main objectives of the proposed comprehensive EPV framework is allowing data analysts to quickly solve the wide variety of performance-related questions that arise daily within professional coaching staff. Based on this, a critical part of this work is devoted to exploring the practical

applicability that such a framework would allow, which we synthesize in the following research question: "can we produce practical applications from the developed models so the set of EPV components can be understood as an analysis framework?"

We present over ten practical applications directly derived from the EPV framework and the different spatial and contextual features developed in this work. In particular, we show how the separated components of the EPV model, the capacity of producing entire probability surfaces, and in general, the application of the framework for assessing the risk and reward balance of both observed and potential actions, provides great flexibility for approaching different applied areas of interest. The majority of applications are structured into three main topics: match and team-level analysis, off-ball performance, and on-ball performance.

For the match and team-level analysis topic, we present a real-time control room where the impact of the different components can be accessed on a frame-by-frame basis for assessing the evolution of attacks. Through this application, coaches can perform a deep inspection into the expected possession value and understand how different decisions might have impacted the outcome of the possession at any given time. Interestingly, we provide both quantitative information and visual representations of the different situations.

In a more specific application, we show how a coach could exploit their opponent teams' different pass-selection tendencies to understand better what to expect in the next match. These passing tendencies are learned directly from the data by adapting the pass-selection component to condition it to the information available for each team. Finally, we show how the process of selecting an optimal initial squad could be enriched by examining the on-ball and off-ball relationship between pairs of players. Based on the evaluation of the added and received EPV, and the pass selection tendencies between teammates, we explore the different team squad configurations that Manchester City's 2014 coach could select to optimize the performance of their playmaker David Silva.

A considerably novel topic explored from the applied standpoint is assessing off-ball performance, a yet-little explored area in soccer analytics. The capacity of the EPV framework of evaluating the risk and reward associated with any potential action provides a new dimension of analysis.

One of these applications consists of exploring the defensive system that could be used to suppress the organized buildup capabilities of Brendan Rodgers's Liverpool. Here, we analyze the possibilities of creating danger in different field zones relative to the defending team's positioning. Instead of directly recommending a specific system, this approach allows a coach to decide the optimal configuration that fits better with his optimal strategy (e.g., progressing through the center lane and breaking the first pressure line). Another novel contribution from the application standpoint is the evaluation of the optimal positioning of players in specific game situations. Given the high amount of time that players spend without being in contact with the ball, positioning in space in time plays a critical role in a player's performance, so coaches spend much time evaluating and correcting off-ball behaviors. The optimal positioning analysis provides a way of exploiting the predicted pass probability surfaces for producing a quantitative assessment of the optimal spaces to occupy. Additionally, the availability of these surfaces eases the communication of these findings to players. Finally, we provide the first known quantification of two popular off-ball concepts, space occupation and space generation (or space creation), both universal concepts in coaches' jargon. We showed how players' spatial performance could be evaluated from the space creation standpoint, identifying the way space is generated (i.e., jogging or running pace), where this space is generated, and the spatial contribution between pairs of players.

We provide three new practical applications regarding on-ball performance, exploiting the fine-grained evaluation of EPV and the spatial and contextual features here developed. First, we propose a new approach for visualizing passing networks that consider the value added with passes between teammates, going beyond the simple representation of pass frequency. In this plot, we combine players' spatial dominance when they are about to receive a pass, the players' average location, and the distribution of added value to gain a more detailed picture of the intricacies of passing dynamics. Second, we present a fine-grained analysis of four different players' added value and passing tendencies relative to the origin and destination location of passes and the teammates' pressure lines. Through this visualization we get information about the players' risk-taking profiles and the obtained reward from those taken risks, contextualized according to the positioning of the defensive team. Specifically, with this information, we can understand how frequently a player risks breaking formation lines through passes and how much value is added from those actions. For example, in the match being analyzed we show how Sergio Ramos tended more frequently to break the

first pressure line, but he added considerably less value than Gerard Piqué, who, in change, attempted to break lines less often. The third on-ball performance application introduces the idea of quantifying the optimal passing location in any game situation. We exploit the pass probability surfaces to identify the destination locations that would maximize the probability of successfully passing to each teammate. This approach provides a clear perspective on the value added by predicting the entire probability surface, where we can explore the impact of passes played into space, rather than only considering the teammate's exact location, offering a more realistic evaluation of passing options. In addition to this, we introduce a new metric to evaluate individual passing skill, PPA. While the usual approach compares the success of passes with the predicted pass probability, we moved a step forward and considered the probability of the optimal passing location. In other words, we are identifying how well did a player identify the optimal passing location and weighting its ability according to the observed success, thus providing a sense of players passing intelligence.

While each of these practical applications can be interesting and useful themselves, the main takeaway is the flexibility and adaptability that the EPV framework provides to develop new applied ideas as quickly as possible. Either by exploiting a subset of components or by directly employing the joint EPV estimate, this approach presents itself as a fully functional framework for addressing challenging applied problems while considering complex spatiotemporal relationships. This EPV framework can be seen as a versatile and flexible toolbox for enhancing soccer analysis with rich quantitative metrics and the capacity of performing visual inspection for better translating the results to practitioners.

8.1.7 Collaboration with professional soccer coaches

The practical focus of this work is catalyzed and enriched by the close collaboration with professional soccer coaches from a variety of FC Barcelona teams. The support of these coaches played a critical role in designing and validating the broad set of spatial and contextual features here developed. The followed methodology, involved periodical meetings over four years, the selection of specific topics of discussion, and the development of a web-based support tool for validating the results along with the video, providing a successful incremental approach for enriching the developed models with the expert considerations of these coaches. This collaboration results especially valuable in developing the different practical applications, ensuring that both

the factors considered in the design of the models and the presentation of results could be integrated into professional coaching staffs.

8.2 Limitations and future work

In this section we present the main limitations of the SoccerMap architecture and the full EPV framework presented. We accompany these limitations with suggestions for further work that could improve the developed models, emphasizing the game state representation, the availability of broader datasets, the learning and optimization considerations, and the availability of new types of data sources.

Coarse field representation The use of a coarse representation of the field in the SoccerMap architecture limits the output to a discrete matrix of values. In this work, we prove we can produce accurate estimations of pass success probability, pass selection probability, and pass expected value estimates with a field representation of size 104×68. Although the use of convolutional neural networks limits the prediction to a discrete space, it has the advantage of considering the spatial relationships of nearby locations directly. Other approaches could address the impact of selecting different configurations for the discrete output matrix. Additionally, other layers consisting of precomputed surfaces of more sophisticated or problem-specific features that consider continuous location space could be introduced to enhance the SoccerMap-based models. An example of this could be to introduce a surface layer representing the degree of interceptability of passes based on the physics-based motion model presented in Spearman et al. (2017).

Model size and number of parameters The number of parameters of the SoccerMap-based model for pass probability is considerably large compared to the benchmark models presented in Chapter 5. We show that SoccerMap could provide real-time estimations (near 200Hz) if GPU computation is available. However, if GPU processing is unavailable, the inference time of the model is expected to increase considerably.

Broader and larger datasets A sufficiently large dataset is required in order to train a model with this high number of parameters. While the access of tracking data for full seasons has been historically limited to public access, several high-level competitions such as the EPL and the German Bundesliga provide tracking data of all the matches to every team

in the league. Recently, private companies are starting to offer tracking data generated from broadcast videos or enhanced versions of event data, including players' locations at every event. This increasing availability of data should facilitate the development of models derived from this approach. Additionally, training these models with more recent tracking data might help capture up-to-date soccer characteristics such as improved physical conditions of players or a tendency to prefer short passes over long passes.

Player and team-specific features This kind of features, such as player skills or team-level playing tendencies, could provide enriched information for producing more accurate estimates in specific game situations. Some examples of these features are: a player-passing skill feature (e.g., average pass completion) for the pass probability model, an action-selection feature indicating either the player's or the team's tendency to pass, keep the ball, or shoot, or a player-level shooting skill. Alternatively, a series of components could be added to the different neural network architectures proposed in this work to learn team-specific features within the same learning process. For example, a group of neurons encoding team passing-style could be added to produce pass-likelihood surfaces adjusted to each team in the dataset. However, it is important to notice that considering player-level or team-level features could make the attribution of value more challenging.

Meta-information of the game and team state The overall estimation of EPV could benefit from including information such as the time when actions occur, the current score, an estimation of the match importance, the known rivalry between the two teams, or an estimation of a team's mental pressure at any given time (Bransen et al., 2019).

Training set-pieces separately The components related to passes and shots could be improved by differentiating between open-play and set-pieces at an implementation level.

Incorporating player orientation and kinematics Computer vision methods could be exploited to derive a new series of features providing information derived from the pose estimation and other robust object recognition technologies. During this work, we collaborated with the development of a model for identifying the upper-torso orientation of players at any time and incorporated that feature within a pass probability model (Arbués-Sangüesa

et al., 2020a,b). These new sources of information could provide new perspectives to incorporate kinematics and motion analysis of players.

Incorporating ball height The standard format of available tracking data provides players locations in a 2D coordinate system, ignoring the z-axis or ball height. While this kind of data is still not usually available in soccer, its incorporation might provide richer information of the game state for most developed components, emphasizing the passing models. The z-axis would allow us to differentiate directly between an aerial and a ground-level pass instead of extracting this information as a latent factor related to the pass distance and its origin and destination location. This information could enrich the decision-making evaluation, for example, in suggesting optimal passing locations by providing a fine-grained assessment of the expected outcome of the different types of passes.

A more nuanced evaluation of short-term value attribution The presented framework focuses on the estimation of the long-term expected outcome of a possession. While the estimation of the EPVA of actions provides a signal for understanding a player's on-ball contribution, we could develop more sophisticated approaches for understanding how to attribute this value in short term. For example, a high-value pass from a midfielder might have been influenced by two previous passes with lower added value, which created the opportunity for the later high-value pass. Through methods related to the reinforcement learning research area, we could learn how to better attribute the high-value pass to the group of players participating in actions close in time to provide a more refined evaluation of a player's actual contribution to the increase of goal-scoring probability.

Different approaches for learning inter-player relationships A recent approach explores graph convolutional neural networks to evaluate players' defensive contribution (Stöckl et al., 2021). The exploration of different neural network architectures and optimization approaches, such as the graph networks learning approach, might provide a different way of learning to estimate EPV, by conditioning the state representation to the relationships between adjacent nodes. While we would lose the rich information provided by estimating probability surfaces, we could obtain a more direct way to understand the expected value of passing to a teammate without considering multiple locations, although risking a too steep simplification of players' expected behavior.

Appendix A

List of spatial and contextual features

Table A.1,A.2, A.3, and A.4 describe the complete set of features used as input for each presented model. The concept type column refers to the general feature grouping described in Chapter 4, including a prefix indicating whether the feature is a spatial feature (SP), a contextual feature (CX), or other types (OT). Model names are presented with acronyms, including: pass success probability (PP), pass selection probability (PS), pass success and missed EPV (PE), ball drive probability (DP), ball drive success and missed EPV (DE), action selection probability (AS), and shot EPV (SE). For PP, PS, and PE models, the input features are either sparse or full matrix of 104×68. When the feature description indicates the value is set of every location, this input will correspond to a full matrix; otherwise, it corresponds to a sparse matrix. For the rest of the models, each feature is provided as a single variable. We refer to the team in control of the ball as the *attacking team*, and its players as the *attacking players*. We refer to the other team as the *defending team*, and its players as the *defending players*. All the features are normalized, assuming a left to right attacking direction of the team in control of the ball (attacking team).

Table A.1: First part of the set of spatial features used as input for each presented model. The concept type column includes a prefix indicating the feature belongs to the spatial feature type (SP). For the rest of the columns a checkmark indicates the models where the feature is used, including: pass success probability (PP), pass selection probability (PS), pass success and missed EPV (PE), ball drive probability (DP), ball drive success and missed EPV (DE), action selection probability (AS), and shot EPV (SE)

Concept Type	Feature	PP	PS	PE	DP	DE	AS	SE
SP - (x,y) location	1 on attacking players' location (x,y).	✓	✓	✓				
SP - (x,y) location	Ball location (x).				✓	✓	✓	✓
SP - (x,y) location	1 on defending players' location (x,y).	✓	✓	✓				
SP - (x,y) location	1 if the ball is closer to the goal than the opponent's goalkeeper.							✓
SP - Velocity	Attacking team players' speed (m/s) (x).	✓	✓	✓				
SP - Velocity	Attacking team players' speed (m/s) (y).	✓	✓	✓				
SP - Velocity	Defending team players' speed (m/s) (y).	✓	✓	✓				
SP - Velocity	Defending team players' speed (m/s) (y).	✓	✓	✓				
SP - Angle	Angle between every location and the goal	✓	✓	✓				
SP - Angle	Angle between the ball and the goal.				✓	✓	✓	✓
SP - Angle	Sine of the angle between every location and the ball location.	✓	✓					
SP - Angle	Cosine of the angle between every location and the ball location.	✓	✓					
SP - Angle	Sine of the angle between the ball carrier velocity vector and every other location.	✓	✓					
SP - Angle	Cosine of the angle between the ball carrier velocity vector and every other location.	✓	✓					

Table A.2: Second part of the set of spatial features used as input for each presented model. The concept type column includes a prefix indicating the feature belongs to the spatial feature type (SP). For the rest of the columns a checkmark indicates the models where the feature is used, including: pass success probability (PP), pass selection probability (PS), pass success and missed EPV (PE), ball drive probability (DP), ball drive success and missed EPV (DE), action selection probability (AS), and shot EPV (SE)

Concept Type	Feature	PP	PS	PE	DP	DE	AS	SE
SP - Distance	Distance between every location and the goal.	✓	✓	✓				
SP - Distance	Distance between every location and the ball.	✓	✓	✓				
SP - Distance	Distance between the ball and the goal.				✓	✓	✓	✓
SP - Distance	Distance between the ball and the goalkeeper in y-axis.							✓
SP - Distance	Distance between the ball and the goalkeeper.							✓
SP - Pitch control	Pitch control of the attacking team at the ball location				✓	✓	✓	
SP - Pitch influence	Pitch influence of the defending team at the ball location.				✓	✓	✓	

Table A.3: First part of the set of contextual features and other feature types used as input for each presented model. The concept type column includes a prefix indicating whether the feature is a contextual feature (CX) or other types (OT). For the rest of the columns a checkmark indicates the models where the feature is used, including pass success probability (PP), pass selection probability (PS), pass success and missed EPV (PE), ball drive probability (KP), ball drive success and missed EPV (KE), action selection probability (AS), and shot EPV (SE)

Concept Type	Feature	PP	PS	PE	DP	DE	AS	SE
CX - Dynamic pressure lines	Index of the closest attacking team line to every location.			✓				
CX - Dynamic pressure lines	Index of the closest attacking team line to the ball location.				✓	✓	✓	
CX - Dynamic pressure lines	Index of the closest defending team line to every location.			✓				
CX - Dynamic pressure lines	Index of the closest defending team line to the ball location.				✓	✓	✓	
CX - Outplayed players	Number of attacking team's players between the ball and every other location.			✓				
CX - Outplayed players	Number of defending players between the ball and every other location.			✓				
CX - Outplayed players	Number of attacking players between the opponent's goal and every other location.			✓				

Table A.4: Second part of the set of contextual features and other feature types used as input for each presented model. The concept type column includes a prefix indicating whether the feature is a contextual feature (CX) or other types (OT). For the rest of the columns a checkmark indicates the models where the feature is used, including pass success probability (PP), pass selection probability (PS), pass success and missed EPV (PE), ball drive probability (KP), ball drive success and missed EPV (KE), action selection probability (AS), and shot EPV (SE)

Concept Type	Feature	PP	PS	PE	DP	DE	AS	SE
CX - Dynamic pressure lines	Index of the closest attacking team line to every location.			✓				
CX - Outplayed players	Number of players of the defending team between the opponent's goal and every other location.			✓				
CX - Interceptability	Number of defending players inside the triangle formed between the ball location and the posts of the opponent's goal.							✓
CX - Interceptability	Number of defending players located less than 3 meters away from the ball location.							✓
CX - Event-based xG	Expected goals based on the action location and the angle to the goal.						✓	✓
OT - Type	1 of action is attempted with the head.							✓
OT - Probability	Pass probability surface.		✓					
OT - Probability	Ball drive probability.					✓		

Bibliography

11Tegen11 (2015) A close look at my new expected goals model. http://11tegen11.net/2015/08/14/a-close-look-at-my-new-expected-goals-model/, [Online; accessed 1-July-2019]

Analysis AS (2017) What are expected Goals? https://www.americansocceranalysis.com/explanation/, [Online; accessed 1-July-2019]

Arbués-Sangüesa A, Martin A, Fernández J, Ballester C, Haro G (2020a) Using player's body-orientation to model pass feasibility in soccer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 886–887

Arbués-Sangüesa A, Martín A, Fernández J, Rodríguez C, Haro G, Ballester C (2020b) Always look on the bright side of the field: Merging pose and contextual data to estimate orientation of soccer players. In: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, pp 1506–1510

Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence 39(12):2481–2495

Bornn L, Cervone D, Fernández J (2018) Soccer analytics: Unravelling the complexity of "the beautiful game". Significance 15(3):26–29

Bransen L, Van Haaren J (2018) Measuring football players' on-the-ball contributions from passes during games. In: International workshop on machine learning and data mining for sports analytics, Springer, pp 3–15

Bransen L, Robberechts P, Van Haaren J, Davis J (2019) Choke or shine? quantifying soccer players' abilities to perform under mental pressure. In: 13 th Annual MIT Sloan Sports Analytics Conference

Buldú JM, Busquets J, Martínez JH, Herrera-Diestra JL, Echegoyen I, Galeano J, Luque J (2018) Using network science to analyse football passing networks: Dynamics, space, time, and the multilayer nature of the game. Frontiers in psychology 9:1900

Caley M (2015a) Expected goals and uncertainty. http://pena.lt/y/2016/04/29/expected-goals-and-uncertainty/, [Online; accessed 1-July-2019]

Caley M (2015b) Premier league projections and new expected goals. https://cartilagefreecaptain.sbnation.com/2015/10/19/9295905/premier-league-projections-and-new-expected-goals, [Online; accessed 1-July-2019]

Cervone D, Bornn L, Goldsberry K (2016a) Nba court realty. In: 10th MIT Sloan Sports Analytics Conference

Cervone D, D'Amour A, Bornn L, Goldsberry K (2016b) A multiresolution stochastic process model for predicting basketball possession outcomes. Journal of the American Statistical Association 111(514):585–599

Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence 40(4):834–848

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM, pp 785–794

Dahl GE, Sainath TN, Hinton GE (2013) Improving deep neural networks for lvcsr using rectified linear units and dropout. In: 2013 IEEE international conference on acoustics, speech and signal processing, IEEE, pp 8609–8613

Decroos T, Bransen L, Van Haaren J, Davis J (2019) Actions speak louder than goals: Valuing player actions in soccer. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 1851–1861

Dick U, Brefeld U (2019) Learning to rate player positioning in soccer. Big data 7(1):71–82

Dozat T (2016) Incorporating nesterov momentum into adam

Duchi J, Hazan E, Singer Y (2011) Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research 12(Jul):2121–2159

Dumoulin V, Visin F (2016) A guide to convolution arithmetic for deep learning. arXiv preprint arXiv:160307285

Eggels H (2016) Expected goals in soccer: Explaining match results using predictive analytics. In: The Machine Learning and Data Mining for Sports Analytics workshop, p 16

Fonseca S, Milho J, Travassos B, Araújo D (2012) Spatial dynamics of team sports exposed by voronoi diagrams. Human movement science 31(6):1652–1659

Friedman J, Hastie T, Tibshirani R, et al. (2000) Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Annals of statistics 28(2):337–407

Fujimura A, Sugihara K (2005) Geometric analysis and quantitative evaluation of sport teamwork. Systems and Computers in Japan 36(6):49–58

Goldsberry K (2019) Sprawlball: A Visual Tour of the New Era of the NBA, Houghton Miffling Publishing Company, chap The Investor

Goodfellow I, Bengio Y, Courville A, Bengio Y (2016) Deep learning, MIT press Cambridge, chap Deep Feedforward Networks

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

Guo C, Pleiss G, Sun Y, Weinberger KQ (2017) On calibration of modern neural networks. In: International Conference on Machine Learning, PMLR, pp 1321–1330

Gyarmati L, Stanojevic R (2016) Qpass: a merit-based evaluation of soccer passes. arXiv preprint arXiv:160803532

Haase J, Brefeld U (2013) Finding similar movements in positional data streams. In: MLSA@ PKDD/ECML, pp 49–57

Hacker R (1962) Certification of algorithm 112: position of point relative to polygon. Communications of the ACM 5(12):606

Haines E (1994) Point in polygon strategies. Graphics Gems 4:24–46

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

Hubáček O, Šourek G, Železnỳ F (2018) Deep learning from spatial relations for soccer pass prediction. In: International Workshop on Machine Learning and Data Mining for Sports Analytics, Springer, pp 159–166

Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:150203167

Kaden S, Mellmann H, Scheunemann M, Burkhard HD (2013) Voronoi based strategic positioning for robot soccer. In: Proceedings of the 22nd International Workshop on Concurrency, Specification and Programming (CS&P), vol 1032, pp 271–282

Kim S (2004) Voronoi analysis of a soccer game. Nonlinear Analysis: Modelling and Control 9(3):233–240

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980

Knutson T (2018) Explaining xGChain Passing Networks. https://statsbomb.com/2018/08/explaining-xgchain-passing-networks/, [Online; accessed 12-March-2021]

Law J (2005) Analysis of multi-robot cooperation using voronoi diagrams. In: 3rd International RCL/VNIItransmash Workshop on Planetary Rovers, Space Robotics and Earth-Based Robots

Le HM, Carr P, Yue Y, Lucey P (2017) Data-driven ghosting using deep imitation learning. In: The 11th Annual MIT Sloan Sports Analytics Conference

LeCun Y, Bengio Y, et al. (1995) Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks 3361(10):1995

Link D, Lang S, Seidenschwarz P (2016) Real time quantification of dangerousity in football using spatiotemporal tracking data. PloS one 11(12):e0168768

Liu G, Schulte O (2018) Deep reinforcement learning in ice hockey for context-aware player evaluation. arXiv preprint arXiv:180511088

Liu G, Luo Y, Schulte O, Kharrat T (2020) Deep soccer analytics: learning an action-value function for evaluating soccer players. Data Mining and Knowledge Discovery 34(5):1531–1559

Long J, Zhang N, Darrell T (2014) Do convnets learn correspondence? In: Advances in Neural Information Processing Systems, pp 1601–1609

Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440

Lucey P, Bialkowski A, Monfort M, Carr P, Matthews I (2014) Quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. In: Proc. 8th annual mit sloan sports analytics conference, pp 1–9

Lundberg S (2020) shap. https://github.com/slundberg/shap

Lundberg S, Lee SI (2017) A unified approach to interpreting model predictions. arXiv preprint arXiv:170507874

Masheswaran R, Chang Y, Su J, Kwok S, Levy T, Wexler A, Hollingsworth N (2014) The three dimensions of rebounding. MIT SSAC

McHale IG, Relton SD (2018) Identifying key players in soccer teams using network analysis and pass difficulty. European Journal of Operational Research 268(1):339–347

Memmert D, Lemmink KA, Sampaio J (2017) Current approaches to tactical performance analyses in soccer using position data. Sports Medicine 47(1):1–10

Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M (2013) Playing atari with deep reinforcement learning. arXiv preprint arXiv:13125602

Naeini MP, Cooper G, Hauskrecht M (2015) Obtaining well calibrated probabilities using bayesian binning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 29

Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: Icml

Niculescu-Mizil A, Caruana R (2005) Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning, pp 625–632

Nixon J, Dusenberry MW, Zhang L, Jerfel G, Tran D (2019) Measuring calibration in deep learning. In: CVPR Workshops, vol 2

Nwankpa C, Ijomah W, Gachagan A, Marshall S (2018) Activation functions: Comparison of trends in practice and research for deep learning. arXiv preprint arXiv:181103378

Odena A, Dumoulin V, Olah C (2016) Deconvolution and checkerboard artifacts. Distill 1(10):e3

Papandreou G, Chen LC, Murphy KP, Yuille AL (2015) Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 1742–1750

Pathak D, Krahenbuhl P, Darrell T (2015) Constrained convolutional neural networks for weakly supervised segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 1796–1804

Perl J, Memmert D (2016) Soccer analyses by means of artificial neural networks, automatic pass recognition and voronoi-cells: An approach of measuring tactical success. In: Proceedings of the 10th International Symposium on Computer Science in Sports (ISCSS), Springer, pp 77–84

Platt J, et al. (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 10(3):61–74

Power P, Ruiz H, Wei X, 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, ACM, pp 1605–1613

Prokopenko M, Wang P, Obst O (2014) Gliders2014: Dynamic tactics with voronoi diagrams. In: RoboCup 2014 symposium and competitions: Team description papers, Joao Pessoa, Brazil, Citeseer

Rein R, Memmert D (2016) Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. SpringerPlus 5(1):1410

Rein R, Raabe D, Memmert D (2017) "which pass is better?" novel approaches to assess passing effectiveness in elite soccer. Human movement science 55:172–181

Ric A, Torrents C, Gonçalves B, Torres-Ronda L, Sampaio J, Hristovski R (2017) Dynamics of tactical behaviour in association football when manipulating players' space of interaction. PloS one 12(7):e0180773

Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer, pp 234–241

Roth V, Steinhage V (2000) Nonlinear discriminant analysis using kernel functions. In: Advances in neural information processing systems, pp 568–574

Rudd S (2011) A framework for tactical analysis and individual offensive production assessment in soccer using markov chains. In: New England Symposium on Statistics in Sports. http://nessis.org/nessis11/rudd.pdf

Rüping S (2006) Robust probabilistic calibration. In: European Conference on Machine Learning, Springer, pp 743–750

Schaper S (2021) "packing" and "impect". https://support.scisports.com/en/articles/3161884-packing-and-impect, [Online; accessed 28-March-2021]

Scherer D, Müller A, Behnke S (2010) Evaluation of pooling operations in convolutional architectures for object recognition. In: International conference on artificial neural networks, Springer, pp 92–101

Schiffer S, Ferrein A, Lakemeyer G (2006) Qualitative world models for soccer robots. In: Qualitative Constraint Calculi, Workshop at KI, vol 2006, pp 3–14

Schölkopf B (2001) The kernel trick for distances. In: Advances in neural information processing systems, pp 301–307

Schulte O, Khademi M, Gholami S, Zhao Z, Javan M, Desaulniers P (2017) A markov game model for valuing actions, locations, and team performance in ice hockey. Data Mining and Knowledge Discovery 31(6):1735–1757

Seidl T, Cherukumudi A, Hartnett A, Carr P, Lucey P (2018) Bhostgusters: Realtime interactive play sketching with synthesized nba defenses. In: 12 th Annual MIT Sloan Sports Analytics Conference

Shamir O (2018) Distribution-specific hardness of learning neural networks. The Journal of Machine Learning Research 19(1):1135–1163

Shapley LS (2016) 17. A value for n-person games. Princeton University Press

Shi W, Caballero J, Theis L, Huszar F, Aitken A, Ledig C, Wang Z (2016) Is the deconvolution layer the same as a convolutional layer? arXiv preprint arXiv:160907009

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, et al. (2016) Mastering the game of go with deep neural networks and tree search. nature 529(7587):484

Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, Lanctot M, Sifre L, Kumaran D, Graepel T, et al. (2017) Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv preprint arXiv:171201815

Singh K (2019) Introducing expected threat (xt). https://karun.in/blog/expected-threat.html, accessed: 2020-10-16

Spearman W (2018) Beyond expected goals. In: Proceeding of the 12th MIT Sloan Sports Analytics Conference

Spearman W, Basye A, Dick G, Hotovy R, Pop P (2017) Physics—based modeling of pass probabilities in soccer. In: Proceeding of the 11th MIT Sloan Sports Analytics Conference

Spruyt V (2014) Geometric interpretation of the covariance matrix. http://www.visiondummy.com/2014/04/geometric-interpretation-covariance-matrix, [Online; accessed 1-March-2021]

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15(1):1929–1958

Stein M, Janetzko H, Seebacher D, Jäger A, Nagel M, Hölsch J, Kosub S, Schreck T, Keim D, Grossniklaus M (2017) How to make sense of team sport data: From acquisition to data modeling and research aspects. Data 2(1):2

Stöckl M, Seidl T, Marley D, Power P (2021) Making offensive play predictable-using a graph convolutional network to understand defensive performance in soccer. In: Proceeding of the 14th MIT Sloan Sports Analytics Conference

Sutton RS, Barto AG, et al. (1998) Introduction to reinforcement learning, vol 135. MIT press Cambridge

Taki T, Hasegawa Ji (2000) Visualization of dominant region in team games and its application to teamwork analysis. In: Proceedings Computer Graphics International 2000, IEEE, pp 227–235

Vercruyssen V, De Raedt L, Davis J (2016) Qualitative spatial reasoning for soccer pass prediction. In: Machine learning and data mining for sports analytics, ECML/PKDD workshop, Riva del Garda, vol 19

Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X (2017) Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3156–3164

Wilks DS (1990) On the combination of forecast probabilities for consecutive precipitation periods. Weather and forecasting 5(4):640–650

Yurko R, Matano F, Richardson LF, Granered N, Pospisil T, Pelechrinis K, Ventura SL (2020) Going deep: models for continuous-time within-play valuation of game outcomes in american football with tracking data. Journal of Quantitative Analysis in Sports 1(ahead-of-print)

Zeiler MD (2012) Adadelta: an adaptive learning rate method. arXiv preprint arXiv:12125701