Leiden University

Thesis submitted for the degree

Master of Science

Statistics: Data Science

Increasing the Predictive Power of the Possession Metric in Football by Adding Spatio-temporal Context

by

Tim Lukas Schwarz

Supervisors:

Dr. Odysseas Kanavetas
Dr. Arno Knobbe
Dr. Rens Meerhoff

November 2021

Abstract

Increasing the Predictive Power of the Possession Metric in Football by Adding Spatio-temporal Context

In recent years, statistics play an increasing role in professional football. A controversial topic inside the emerging field of football data science is the effect of ball possession on match outcomes. We contribute to this discussion by analyzing the effect of possession on match outcomes while controlling for match status and match-up balance. We examine the importance of the position of possession by comparing the kernel density estimate of winning and losing teams. Based on these findings we split the football pitch into distinct zones using Voronio cells based on the centroids of a k-means clustering. We fit a multiple linear regression model that regresses a matches final goal difference on possession per match status per zone using a 5x5-fold nested cross-validation. The resulting model splits the football pitch into 11 zones. Our metric holds higher predictive power than the traditional metric. To demonstrate the potential of this work for both analysts and journalists, we analyze a teams performance over a whole season as well as individual match performances using the metric.

Keywords: football, soccer, possession, event data, spatio-temporal data, match status, match-up balance, kernel density estimation, k-means clustering, voronoi cells

Contents

Abstract **ii**

Contents **iii**

List of Figures **vii**

List of Tables **x**

Introduction **1**

1 Existing Literature **3**

1.1 How is Possession Currently Measured? 3

1.2 'Possession - an Empty Metric?' The Debate on the Effect of Possession on Match Outcome 4

1.3 Recent Developments 5

2 Methods **6**

2.1 Kernel Density Estimation 6

2.2 K-means Clustering 7

2.3 Multiple Linear Regression 8

2.4 Nested Cross-Validation 9

3 Data **12**

3.1 The Event Data Set 12

3.2 The Bookmaker Odds Data Set 12

3.3 Features Relevant for this Project 13

3.4 Event Type Pass and its sub-types 16

4 Exploratory Data Analysis 19

4.1 Effect of Raw Possession 19

4.2 Controls 20

4.2.1 Match Status 20

4.2.2 Match-up Balance 22

4.2.3 Combining Match Status and Match-up Balance 23

4.3 Kernel Density Estimation of Passes 25

5 Zone Definition 27

5.1 Mirroring Passes along the Horizontal Axis of the Pitch 27

5.2 Rectangles 28

5.3 Voronoi Cells 29

5.4 Possession per Zone 31

6 The Model 33

6.1 Model Design 33

6.1.1 Controlling for Match Status 35

6.1.2 Dealing with Missing Values 36

7 Results 37

7.1 Results of Nested Cross-Validation 37

7.1.1 Results Inner Loops 37

7.1.2 Results Outer Loop. 40

7.1.3 Results on Subset of Matches with Match Status Even. 41

7.2 Final Model 41

8 Applications 44

8.1 Transformation 44

8.2 Application on Aggregated Team Level 46

8.2.1 Atlético Madrid 47

8.2.2 Las Palmas 50

8.3 Application on Individual Match Level 51

8.3.1 Real Betis - Real Madrid 53

Conclusion 58

Bibliography 63

A Significance Tests 66

A.1 Possession per Outcome 67

A.2 Possession per Outcome Match Status 67

A.3 Possession per Outcome - Match Status 68

A.4 Possession per Outcome - Match-up Balance 68

A.5 Possession per Outcome - Match Status and Match-up Balance. 69

B KDE of Sub-events of Event Type Pass 70

B.1 Simple Pass 70

B.2 High pass 71

B.3 Head pass 71

B.4 Cross 72

B.5 Launch 72

B.6 Smart Pass 73

B.7 Hand Pass 73

C Nested Cross Validation Results 74

C.1 Full Results Voronoi Zones 75

C.2 Full Results Rectangular Zones 76

List of Figures

2.1 The Outer Loop of a Nested Cross-Validation with $l = 5$ Outer Folds Visualized. . . . 9

2.2 An Inner Loop of a Nested Cross-Validation with $m = 5$ Inner Folds Visualized. . . . 10

3.1 5 Events from the Match Eintracht Frankfurt (2462) vs. SV Werder Bremen (2443) on Matchday 11 of the 2017/2018 Bundesliga Season Visualized. . . 16

4.1 The Subset Used to Control for Match Status Visualized. . . . 21

4.2 The Subset Used to Control for Match-up Balance Visualized 22

4.3 The Subset Used to Control for the Combination of Match-up Balance and Match Status Visualized. . . . 24

4.4 Density Estimates of Winners (a) and Losers (b) in the Controlled Subset Compared. . . . 25

4.5 Difference Between the Normalized Density Estimates of Winners and Losers in the Controlled Subset. . . . 26

5.1 The Mirroring Process Visualized. . . . 28

5.2 Rectangular Zones Visualized 29

5.3 Voronoi Zones Visualized. . . . 30

5.4 The Mean Possession Per Zone in the Controlled Subset. . . . 31

6.1 The Composition of X_i for the Basic Version of the Model with $k = 10$ Voronoi Zones Visualized. 34

6.2 The Composition of X_i for the Match Status Controlled Version of the Model with $k = 10$ Voronoi Zones Visualized. 36

7.1 Mean Score per k During the $l = 5$ Folds of each Inner Loop. 38

7.2 The Weights per Zone of the Final Model Visualized. 42

8.1 Comparison of Cumulative Distribution Functions of Raw Possession and our Transformed Model Prediction 45

8.2 Our Final Models Prediction for Atlético Madrid's Performance for Status Drawing in the 2017-2018 La Liga Season. 48

8.3 The Two Components of Atlético Madrid's Prediction for Status Drawing in the 2017-2018 La Liga Season. 49

8.4 Our Final Models Prediction for Las Palmas' Performance for Status Drawing in the 2017-2018 La Liga Season. 50

8.5 The Two Components of Las Palmas' Prediction for Status Drawing in the 2017-2018 La Liga Season. 51

8.6 Timeline of the Match Real Betis Sevilla - Real Madrid on Matchday 24 of the 2017-2018 La Liga Season. 54

8.7 Model Prediction for Match Status Trailing for the Match Real Betis Sevilla - Real Madrid on Matchday 24 of the 2017-2018 La Liga Season. 55

8.8 Model Prediction for Match Status Drawing for the Match Real Betis Sevilla - Real Madrid on Matchday 24 of the 2017-2018 La Liga Season. . . 56

8.9 Model Prediction for Match Status Leading for the Match Real Betis Sevilla - Real Madrid on Matchday 24 of the 2017-2018 La Liga Season. 57

A.1 Possession per Outcome: Histogram of the Mean Possession of Winners under H_0. 67

A.2 Possession per Match Status: Histogram of the Mean Possession of Trailing Teams under H_0. 67

A.3 Possession on Outcome for Match Status drawing: Histogram of the Mean Possession of Winners under H_0. 68

A.4 Possession on Outcome for Match-up Balance even: Histogram of the Mean Possession of Winners under H_0. 68

A.5 Possession on Outcome for Match Status drawing and Match-up Balance even: Histogram of the Mean Possession of Winners under H_0. 69

B.1 KDE of Pass Sub-type Simple Pass. 70

B.2 KDE of Pass Sub-type High Pass. 71

B.3 KDE of Pass Sub-type Head Pass. 71

B.4 KDE of Pass Sub-type Cross. 72

B.5 KDE of Pass Sub-type Launch. 72

B.6 KDE of Pass Sub-type Smart Pass. 73

B.7 KDE of Pass Sub-type Hand Pass. 73

List of Tables

3.1 The columns in our data set containing an events identifiers. 13

3.2 The columns in our data set containing match and event information. . . . 14

3.3 The columns in our data set containing the spatio-temporal context. 15

3.4 Frequency Counts of Event Types. 17

3.5 Frequency Counts of Sub-types of Event Pass. 18

4.1 Number of Passes per Match Outcome. 20

4.2 Number of Passes per Match Status. 21

4.3 Number of Passes per Outcome for Match Status Drawing. 22

4.4 Number of Passes per Outcome for Match-up Balance Even. 23

4.5 Number of Passes per Outcome for Match-up Balance Even and Match Status Drawing. 24

7.1 Test Scores and Chosen Hyper-parameter k for Inner Loops. 39

7.2 Aggregated Results Inner Loops. 39

7.3 Results Outer Loop 40

7.4 Results Outer Loop on Even Match-up Balance Subset. 41

8.1 Final Table of La Liga for the 2017-2018 Season 46

8.2 The 20 Matches during the 2017-2018 Season across the Top 5 Leagues with the Biggest Absolute Difference Between Raw Possession and our Models Transformed Prediction. 52

C.1 Full Results of all 25 Inner Folds of the Nested Cross-Validation for Zone Type Voronoi 75

C.2 Full Results of all 25 Inner Folds of the Nested Cross-Validation for Zone Type Rectangular 76

Introduction

Football is the most popular sport in the world. It is also a billion dollar industry where small edges can make the difference between winning and losing. Therefore, it does not come as a surprise that with the emergency of more and better data, the field of football data science has been rapidly growing. While for sports like Basketball and Baseball so called 'advanced analytics' to analyze and report on players and teams performances are well established, football journalism still relies on metrics based exclusively on the counting of events, so called 'counting stats'. Examples of these stats are the amount of shots taken, corner kicks taken or possession per team, without any further context. While there have been recent improvements to parts of these metrics by providing them with context, most prominently the expected goals approach to shots taken (see [1]), the possession metric has been reported in the same manner for the last 25 years, when modern football data science took of, enabled by the founding of data provider Opta Sports.

The current way of reporting on possession, expressed as a percentage share per team, which will be referred to as raw possession from here on out, has not been without criticism. Players, coaches and analysts alike have voiced their dissatisfaction with the metric, going as far as calling it 'useless'. In the academic field of football data science, this debate continues. Existing literature trying to shed more light on the effect of possession on match outcomes shows contradicting results, partly due to the limitations imposed by the data sets used.

The goal of this thesis is two-fold: First, we aim to contribute to the debate on the effect of possession on match outcome by making use of the more granular event data available. This will be done by splitting up the event data set into subsets, allowing to control for *Match Status* and *Match-up Balance* and its combination. Second, we aim to improve up on raw possession, by adding spatio-temporal context to the passes played. The spatial context will be included into the metric by splitting up the football pitch into a number of distinct zones and weighting these zones according to their effect on the outcome of football matches. The temporal context will be included by taking into account the *Match Status* at the time a pass is played.

This thesis starts off by giving an overview on the existing literature. Then, the methods used in this project, as well as their notation are established. Next, the two data sets and the features relevant for this project are described. Once the data is introduced, the effect of possession on match outcomes will be explored. To explore the temporal context, we examine the effect of different levels of *Match Status* and *Match-up Balance* on the relationship of possession and match outcome. Based on these finding, a controlled subset of the data is defined. On this subset, the effect of the spatial context of passes will be explored using kernel density estimation. Once the importance of position is identified, ways of segmenting the football pitch into distinct zones to capture these differences are introduced. Based on these zones, a multiple linear regression, regressing goal difference on possession per zone, is formulated. The process of fitting the model and its hyper-parameters using nested cross-validation is explained and the resulting model and its coefficients are being presented. Once we obtained our final model, possible applications for analysts and journalists are presented. In the discussion section the results of my project will be reflected upon and the results will be put into the context of the existing literature. The current model's limitations as well as possible further work will be discussed.

Chapter 1

Existing Literature

In this chapter, an overview over the current way of measuring possession as well as the academic discussion surrounding it is given. Section 1.3 introduces recently developed approaches of valuing possessions sequences using machine learning techniques.

1.1 How is Possession Currently Measured?

To improve upon anything, it is vital to understand the way that it is done currently. Right now, there is not a uniform way that possession is measured. There are different approaches to tracking raw possession that are being employed by the data providers in football. Two of them will be discussed in this chapter.

The first one is to let an observer decide about the change of possession and clock the time for which a team was in control of the ball. This method is analogous to a chess clock used in tournament chess, just that instead of the time needed to think about a move per player the time spent with the ball gets summed up for each team. The total time spent with the ball per team then gets turned into a ratio by dividing it through the sum of both clocks. This method is an intuitive way of calculating raw possession. However, the implementation of the 'chess clock method' requires us to make subjective choices about the start and end to a possession. For example, it is not clear how to treat cases where the ball is temporarily out of play or cases where possession is currently being fought over.

The second one, which has been used by the prominent football data provider Opta Sports, is to base possession on the number of passes played per team. This method disregards the time per possession, assuming that the average time per played pass is equal among teams over the course of a whole match.

Both approaches have their advantages and drawbacks. In this project, possession will be defined as the number of passes played per team. The method was chosen as it allows to measure possession without the need for additional subjective definitions.

1.2 'Possession - an Empty Metric?' The Debate on the Effect of Possession on Match Outcome

There is an ongoing debate about the impact of possession on the outcome of football matches. Analysts and commentators are often very critical about the metric, some going as far as calling it an 'empty metric', implying that it holds no explanatory power at all. The existing literature shows that these claims are not without reason.

Collet et al. [2] argue that while there is a positive effect of possession on outcomes, this changes once you control for teams strength. They find that in a balanced match-up, so a match between teams of equal strength, the effect of possession on outcome turns negative. Possibly due to limitations of their data, Collet et al. [2] omit another important factor in their analysis: Lagos et al. [3] found that the match status has a significant effect on possession. Possession tends to lean towards the trailing team once the game is not tied anymore, with the team in the lead showing lower possession. As time spent trailing naturally correlates with losing a match, omitting this factor will lead to a negative bias in the effect of possession on match outcomes.

As our data set allows controlling for both team strength and match status, both the individual impact of *Match-up Balance* and *Match Status* as well as their combined effect will be analyzed and taken into account while building and validating our model.

1.3 Recent Developments

With the increasing availability of granular data, new approaches to evaluate passes and possessions have emerged. One of them is the metric Expected Possession Value (EPV) by Fernandez et al. [3]. EPV uses machine learning techniques to evaluate the likelihood of a goal occurring from the current game state and to calculate the effect of events like passes on that likelihood.

Multiple approaches to evaluate individual passes where developed by Bransen et al. [4]. Zone-oriented Pass Value (ZPV), Pass-oriented Pass Value (PPV) and Sequence-oriented Pass Value (SPV) use an underlying expected goals model to assign value to individual passes. ZPV uses this expected goals model to value zones of equal size on the pitch and uses the passes origin and destination to assign value to a pass. ZPV is the approach most similar to what will be done during this thesis, but its focus is on individual passes and weights zones based on possession outcomes, while the focus of our metric will be on aggregated team possession values and match outcomes. PPV and SPV compare a pass to similar passes (PPV) and possession to similar possessions (SPV) played in the past and base their evaluation on this comparison.

While approaches like EPV, ZPV, PPV and SPV are of great use for experts in the field (both papers are highly recommended reads for anyone interested in football data science), they have not found acceptance in mainstream football culture yet. We argue that this is in part due to their high complexity. Therefore, we strive to develop a metric that extends on the familiar format of raw possession and is easy to interpret for football fans and players alike.

Chapter 2

Methods

In this chapter, the statistical methods used during our project are introduced. For each method, we give a brief mathematical background and introduce the notation used during this thesis.

2.1 Kernel Density Estimation

Kernel density estimation (KDE) is a common density estimation method. It allows to get an estimate of the probability distribution function of a given sample. In this section, a brief introduction of the method and its notation is provided. For a more detailed explanation see Weglarczyk [5].

Suppose we have an i.i.d. sample $x_1, x_2, ..., x_n$ of size n.
KDE tries to infer the density function f that generated this sample. KDE assigns each data point x_i in the sample a kernel function k.
This estimate $\hat{f}$ is defined as:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{\|x - x_i\|}{h} \right),$$

where h is a bandwidth parameter, and the kernel is commonly a Gaussian,

$$k(t) = \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}t^2).$$

Bandwidth parameter h is central for the results of a KDE. It has a similar function to the bin size of histrograms. If h is too large, the density estimate will be oversmoothed, missing to pick up on variation in the data. If h is too small, the density estimate becomes sensitive to noise in the data. There are multiple ways to tune parameter h, including rule of thumbs like the 'Silverman's rule' or 'Scott's rule', as well as resampling methods like bootstrapping.

2.2 K-means Clustering

The k-means algorithm is a method used to divide a given sample into a predefined amount of k clusters. This section will give a brief introduction to the standard version of the algorithm. For a more detailed explanation see Lloyd [6].

Suppose we have an i.i.d. sample $x_1, x_2, ..., x_n$ of size n. The algorithm starts with a random assignment of k center-points $(\mu_1, \mu_2, ..., \mu_n)$. Now, all observations x are assigned to the center point with the smallest distance:

$$S_i^{(t)} = \big\{x_p : \big\|x_p - \mu_i^{(t)}\big\|^2 \leq \big\|x_p - \mu_j^{(t)}\big\|^2 \;\forall j, 1 \leq j \leq k\big\} \tag{2.1}$$

If there are multiple centers with the same distance, the center is chosen at random. Afterwards, the center-points are re-assigned by calculating the mean of the observations clustered to their respective center-points:

$$\mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \tag{2.2}$$

The steps described in Eq. 2.1 and Eq. 2.2 are repeated until all observations remain at the assigned center-points. As k-means is prone to get stuck in a local minimum, multiple random starts can be used to increase the probability of finding the global minimum as a solution.

2.3 Multiple Linear Regression

Multiple Linear Regression is a standard statistical method used across a wide range of fields in science. It is an extension of Simple Linear Regression, allowing for more than one predictor at the same time. In this section, a short explanation of Multiple Linear Regression and its notation is given. For a more detailed explanation see Aiken et al. [7].

Assume we want to predict outcome y based on k predictors. Our sample contains an outcome vector y, as well as the k predictors for all n observations in our sample in the form of a data matrix X. The regression equation for a single observation i in our sample is then:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_k x_{i,k} + \epsilon_i,$$

with β_0 being the intercept, $\beta_1, \beta_2, ..., \beta_k$ being the coefficients for our parameters and ϵ_i being a random error term following a distribution $N \sim (0, \sigma^2)$, with a constant error variance σ^2, identical to Simple Linear Regression. Multiple Linear Regression searches for a combination of the $k+1$ coefficients $\beta_0, \beta_1, ..., \beta_k$ that minimizes the sum of squared errors over our whole sample. The vector B containing the least square estimate of our coefficients $\beta_0, \beta_1, ..., \beta_k$ is obtained by calculating

$$B = (X^{'}X)^{-1}X^{'}Y.$$

2.4 Nested Cross-Validation

In this section the nested cross-validation (N-CV) technique is explained. The notation deviates from the symbols k, l commonly used for the inner and outer folds in N-CV. This comes from the fact that k is already used in the notation of the k-means algorithm. To avoid confusion, the letters l, m are used instead. Cross-validation (CV) is a model validation technique used for estimating a models performance on unseen data (see [8]). Regular CV splits the available data into l folds and assigns one of the folds as the test sets and the remaining $l - 1$ folds as the training set. The model is now trained on the data in the training set and tested on the fold containing the test set. The procedure is repeated for all $\binom{l}{l-1} = l$ combinations of train and test sets (referred to as the Outer Loop from here on) and the mean test score calculated. The resulting mean score gives an estimate of the performance of the model on unseen data.

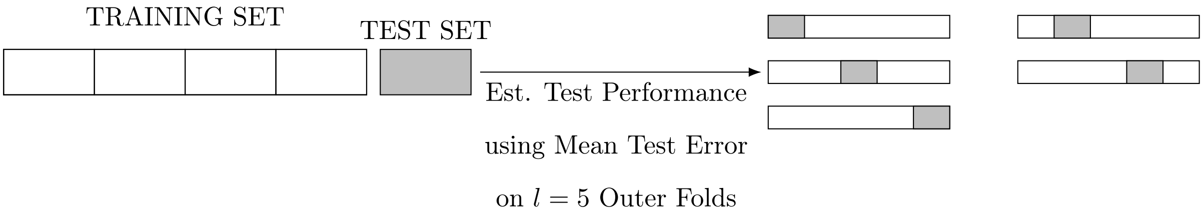


Figure 2.1: The Outer Loop of a Nested Cross-Validation with $l = 5$ Outer Folds Visualized.

Figure 2.1 visualizes this train and test split. If the goal is to test a single model, CV is a sufficient method to obtain unbiased estimates of the models performance on unseen data. But what if there are choices to be made about the models parameters (so called hyperparameters) to choose the combination of hyperparameters that maximizes test performance? Simply comparing the results of a CV of multiple hyperparameter combinations and picking the one that results in the highest mean test score would lead to biased estimates, overestimating the models performance on unseen data (see [9]).

Fortunately, with N-CV there exists a method to obtain an unbiased test performance estimate while simultaneously tuning the hyperparameters of a model. N-CV achieves this by 'nesting' a second CV procedure into the standard CV method. For each of the l train and test combinations of the Outer Loop, the train data of this combination will be split into another m folds. Now, $m-1$ folds are used as training data and one fold is held back as the Validation Set. Analogous to the Outer Loop we cycle over all $\binom{m}{m-1} = m$ combinations per outer fold. A full cycle over all m inner folds is referred to as an Inner Loop. Given the l folds in the Outer Loop, a N-CV has l Inner Loops with a total of $l \times m$ inner folds.

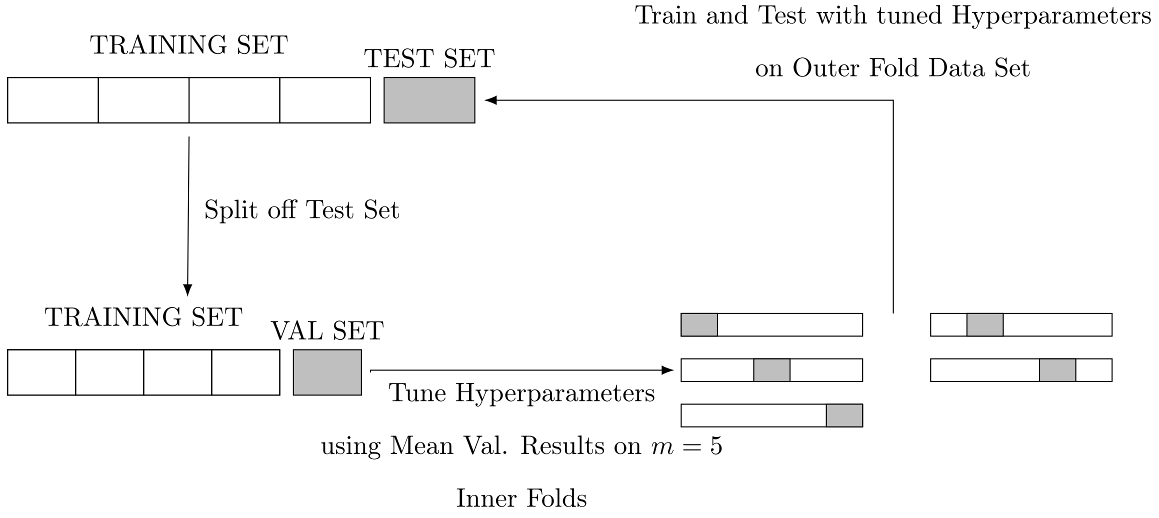


Figure 2.2: An Inner Loop of a Nested Cross-Validation with $m = 5$ Inner Folds Visualized.

Figure 2.2 shows the Inner Loop of a N-CV setup used during model selection and fitting for $m = 5$. During that Inner Loop, the model with different combinations of hyperparameters will be fit on the training set. For each combination, the mean score will be calculated and the combination of hyperparameters with the highest mean score will be selected.

Now, a model with these hyperparameters will be fit on the training set of the Outer Fold and the test score on this model will be the unbiased estimate of the models performance on unseen data. In addition to the estimate of test performance, N-CV also gives an estimate of the models sensitivity to the training data by comparing the hyperparameters of the m different models resulting from the Inner Loops.

Chapter 3

Data

Data sets from two sources were used in this project. In this chapter, the two data sets and their sources will be introduced. Additionally, we will give a description of the features relevant for this project and the additional features constructed based on them.

3.1 The Event Data Set

The *event data set* contains all events of all matches of the 2017-2018 season of the top 5 leagues in Europe (Bundesliga, La Liga, Premier League, Serie A, Ligue 1), as well as the tournaments European Championship 2016 and the World Cup 2018. As tournament football has different rules than league play only the events of league competitions are considered for this project. The data set was made available by Pappalardo et al. [10] under the Creative Commons License [11].

3.2 The Bookmaker Odds Data Set

The second data set used in this project is the collection of betting odds provided by football-data.co.uk [12]. The website collects betting odds at the point of kickoff from various bookmakers and makes them publicly available. The *bookmaker odds data set* contains pre-match odds for every match played in the top 5 leagues during the 2017-2018 season. The betting odds will be used to define *Match-up Balance.*

3.3 Features Relevant for this Project

The two data sets were merged during preprocessing. In this section all features of the merged data set that were used during this project will be introduced and explained. This will be done on the example of the match between Eintracht Frankfurt and SV Werder Bremen on Matchday 11 of the 2017-2018 Bundesliga Season. This match will be referred to as the *sample match* for the rest of this chapter. The data in the following tables show all features of 5 selected events from the *sample match.*

Table 3.1: The columns in our data set containing an events identifiers. The 5 rows are 5 events of the Match between Eintracht Frankfurt and SV Werder Bremen on Matchday 11 of the 2017-2018 Bundesliga Season.

Event ID	Match ID	Team ID	Player ID	Home ID
158037	2516834	2443	16025	2462
158365	2516834	2462	69616	2462
158430	2516834	2462	110	2462
159079	2516834	2443	55990	2462
159797	2516834	2443	82340	2462

Table 3.1 shows the columns containing an events identifiers. *Event ID* is a primary key uniquely identifying every event in the data set. *Match ID*, *Team ID*, *Player ID* and *Home ID* contain the identifiers belonging to an event's Match, Team, player and the ID of the home team for a given match respectively.

Table 3.2: The columns in our data set containing match and event information. The 5 rows are 5 events of the Match between Eintracht Frankfurt and SV Werder Bremen on Matchday 11 of the 2017-2018 Bundesliga Season.

Event ID	Match-up Balance	Match Outcome	Event Type	Event Sub-Type
158037	uneven	lost	Pass	Simple pass
158365	uneven	won	Shot	Shot
158430	uneven	won	Free Kick	Throw in
159079	uneven	lost	Pass	Simple pass
159797	uneven	lost	Pass	Simple pass

Table 3.2 shows the columns containing information about the match in which an event happened, as well as the classification into event type and sub-types. *Match-up Balance* is a categorical variable with levels $[even, uneven]$. A Match-up is classified as even if it belongs to the 50% of matches with the smallest skill difference δ_S, which is defined as

$$\delta_S = |((P_{home} - \mu_{HFA}) - P_{away})|,$$

with P_{home} and P_{away} being the chance of winning implied by the betting odds for the home and away team of a match and $\mu_{HFA} = 0.15$ being the mean home field advantage across the whole data set. The correction for home field advantage was taken as we are trying to control for teams skill level and home field advantage involves other factors such as fan support, additional travel for the away team or referee home bias (see [13]). Additionally, it allows both subsets (*even* and *uneven* matches) to have a balanced amount of home and away matches in them.

Match outcome is a categorical variable representing eventual outcome of the match from the perspective of the active team. As Frankfurt won the game, it is won for all of Eintracht Frankfurt's events, and lost for all of Werder Bremen's.

Events are classified into *event types*, with each *event types* having multiple *event sub-types*. They are given in the respective columns. *event types* and their *sub-types* will be introduced in further detail in Section 3.4.

Table 3.3: The columns in our data set containing the spatio-temporal context. The 5 rows are 5 events of the Match between Eintracht Frankfurt and SV Werder Bremen on Matchday 11 of the 2017-2018 Bundesliga Season.

Event ID	Period	Event Time	Own Score	Opp. Score	Match Status	X	Y
158037	1H	2.25	0	0	drawing	49	51
158365	1H	976.74	0	0	drawing	86	25
158430	1H	1205.76	1	0	leading	26	0
159079	2H	386.18	1	1	drawing	62	80
159797	2H	2893.81	1	2	trailing	59	94

Table 3.2 shows the columns containing an events temporal and spatial context. *Period* and *Time* contain the temporal information about an event. *Own Score* and *Opp. Score* represent the goals scored by each them at the time of the event. *Match Status* is extracted from the score at the time of the event from the perspective of the active team. In line 3 for instance, Eintracht Frankfurt takes a throw in while they are leading 1:0. Therefore, the *Match Status* for that event is leading. *X* and *Y* give the position of an event in a range of [1,2,3...,100].

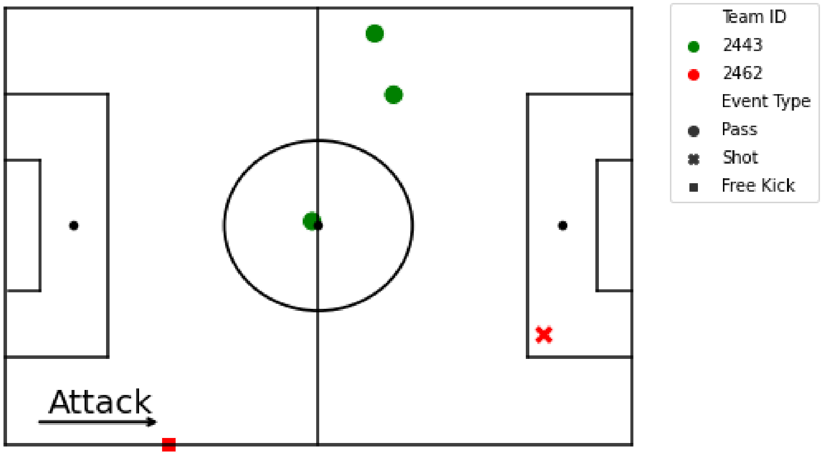


Figure 3.1: 5 Events from the Match Eintracht Frankfurt (2462) vs. SV Werder Bremen (2443) on Matchday 11 of the 2017/2018 Bundesliga Season Visualized.

Figure 3.1 is a visualization of the five selected events from the *sample match*. For instance, the event in line 2, Eintracht Frankfurt's shot that lead to the leading goal, is visualized as the red cross on the pitch. This visualization of the football pitch is used repeatedly during this thesis. While speaking about certain areas of the pitch, this will always be done from the perspective of the attacking team. This means that X describes the position on the axis going from goal to goal and Y the position from left to right. For Eintracht Frankfurt's shot that means that it is in an area with high vertical progression, close to the opponents goal and on the right side of the pitch.

3.4 Event Type Pass and its sub-types

The events in the data set are classified into different event types. Each of this event types contains multiple sub-types.

Table 3.4: Frequency Counts of Event Types.

Event Type	Count	Freq (in %)
Pass	1.565.356	50.97
Duel	832.055	27.09
Others on the ball	242.837	7.91
Free Kick	182.468	5.94
Interruption	130.096	4.24
Foul	47.955	1.56
Shot	40.461	1.32
Save attempt	16.567	0.54
Offside	7.821	0.25
Goalkeeper leaving line	5.779	0.19

Table 3.4 shows the frequency of the different event types. As the focus of this project will be on possession which in turn will be measured on the amount of passes played, events of the type *Pass* will be of central importance. Type *Pass* make up just over half of all events in the data set with type *Duel* making up a little more than another quarter and the rest of share split up over the remaining events.

Table 3.5: Frequency Counts of Sub-types of Event Pass.

Pass Type	Count	Freq. (in %)
Simple pass	1.207.448	77.14
High pass	123.214	7.87
Head pass	91.194	5.83
Cross	58.634	3.75
Launch	43.303	2.77
Smart pass	28.428	1.82
Hand pass	13.135	0.84

Event *Pass* itself comes in seven different sub-types, as can be seen in Table 3.5. Over three fourths of passes are of type *Simple pass*, with 6 other sub-types making up the remaining share. Due to limitations imposed by sample size constraints, differentiation between pass sub-types is not included in our final model. Additional information about the distribution of pass sub-types on the pitch is provided in Appendix B.

Chapter 4

Exploratory Data Analysis

When dealing with a large granular data set like the event data used for this thesis, it is integral to explore the available data before doing any kind of modelling. This chapter starts off by examining the effect of raw possession on match outcomes. Then, a way of controlling for *Match Status* and *Match-up Balance* by looking at different subgroups of the data, is introduced and applied to the data set. All effects are tested on their significance using random label assignment. For readability purposes, only the resulting estimated p-values $\hat{p}$ are reported in this chapter. A detailed explanation of the empirical significance tests including histograms can be found in Appendix A. In Section 4.3, the importance of the position of passes will be explored with the help of kernel density estimation (KDE).

4.1 Effect of Raw Possession

As our aim for this project is to improve upon raw possession, we start off by looking at the amount of explanatory power the raw statistic holds on the data set. This is done by comparing the number of passes played grouped by the categorical variable outcome.

Table 4.1: Number of Passes per Match Outcome.

Outcome	Passes played	Mean Possession
Lost	559,815	0.472
Drew	379,909	0.5
Won	625,632	0.528

Table 4.1 shows that winners have a mean possession of 52.8% in their matches. This number is in line with the findings of Colett et al. [2]. Without controlling for *Match Status* or *Matchup Balance*, raw possession has a small, but significant positive effect on match outcomes ($\hat{p} = 0.018$).

4.2 Controls

Previous studies on possession were limited in their ability to control for *Match Status* due to their data being aggregated on the match level. Additionally, they had to rely on a subjective definition of *Match-up Balance* due to a lack of information on team strength to base *Match-up Balance* on. In this section, new ways to control for both factors and their combination are introduced. This is done by splitting the data set into different subgroups.

4.2.1 Match Status

The work of Lago et al. [3] suggests that aggregated possession numbers can not be interpreted without having additional information about the *Match Status.* This stems from the fact that while trailing teams tend to have a higher amount of possession. And as matches in which a team is trailing are more likely to be lost by that team, high possession on the aggregated match level correlates with a long amount of time spent trailing and consequently with a higher chance of losing the match.

Table 4.2: Number of Passes per Match Status.

Status	Passes played	Mean Possession
Trailing	419,137	0.521
Drawing	625,632	0.5
Leading	385,920	0.479

Table 4.2 confirms the need for such a control. Trailing teams have significantly higher possession than leading teams ($\hat{p} < 0.001$). As the match status may switch multiple times during the same match, it is not straightforward to control for it on an aggregated match level. With the help of the event data set, the subset of passes played while *drawing* can be analyzed in isolation. As every match starts in status *drawing*, it is guaranteed that every match is in *Match Status drawing* at least some amount of time. As drawing can be interpreted as the 'neutral' *Match Status*, this subset is not affected by the effect of *Match Status* on possession.

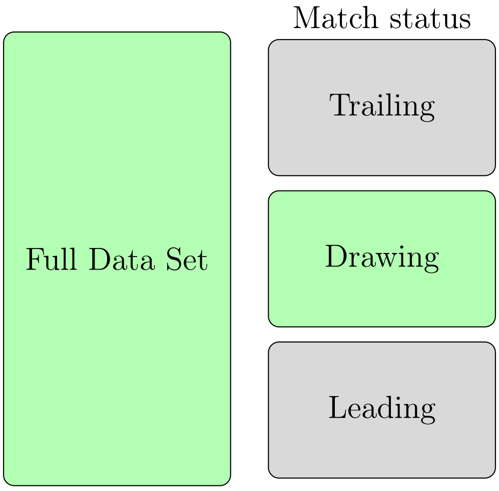


Figure 4.1: The Subset Used to Control for Match Status Visualized.

Figure 4.1 shows the subset resulting from such a restriction. Now, that *Match Status* is controlled for, the comparison of the number of passes played grouped by the categorical variable outcome can be repeated on the *Match Status* controlled subset.

Table 4.3: Number of Passes per Outcome for Match Status Drawing.

Outcome	Passes played	Mean Possession
Lost	213,679	0.439
Drew	272,762	0.5
Won	273,858	0.561

Table 4.3 shows that with 0.561 possession while *drawing*, winning teams hold a significant ($\hat{p} < 0.001$) surplus of possession.

4.2.2 Match-up Balance

Collet et al. [2] suggest that team strength is another important factor to control for while looking at the relation of possession to match outcomes. In their work, they control for *Match-up Balance* by assigning different tiers to clubs and fitting individual linear regressions for the different tiers. They find that the effect of possession turns negative for match-ups between teams of even tiers.

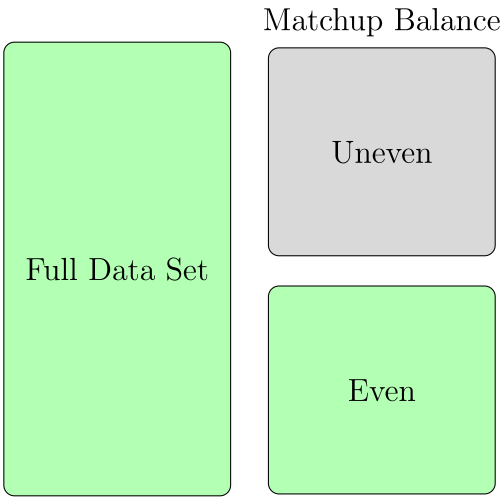


Figure 4.2: The Subset Used to Control for Match-up Balance Visualized.

Instead of a subjective classification into tiers, we will make use of the definition based on the pre-match betting odds introduced in Chapter 3 to define *even* matches. Figure 4.2

visualizes the resulting subset. Once again, the number of passes played grouped by the categorical variable outcome are compared for the *Match-up Balance* controlled subset.

Table 4.4: Number of Passes per Outcome for Match-up Balance Even.

Outcome	Passes played	Mean Possession
Lost	272,064	0.521
Drew	223,619	0.5
Won	249,602	0.479

Table 4.4 confirms the findings from Lago et al.[3]. In matches between opponents of *even* strength, without simultaneously controlling for *Match Status*, winners hold less possession than losers. However, the effect is not significant at level $a = 0.05$ ($\hat{p} = 0.134$).

4.2.3 Combining Match Status and Match-up Balance

The results from Section 4.2.1 and Section 4.2.2 help to understand the ongoing debate about the effect of possession on match outcomes in the field of football data science. As the effects of controlling for one of the two controls moves the effect of possession in opposite directions, the result will be different depending on the control applied. In this section, both controls will be combined.

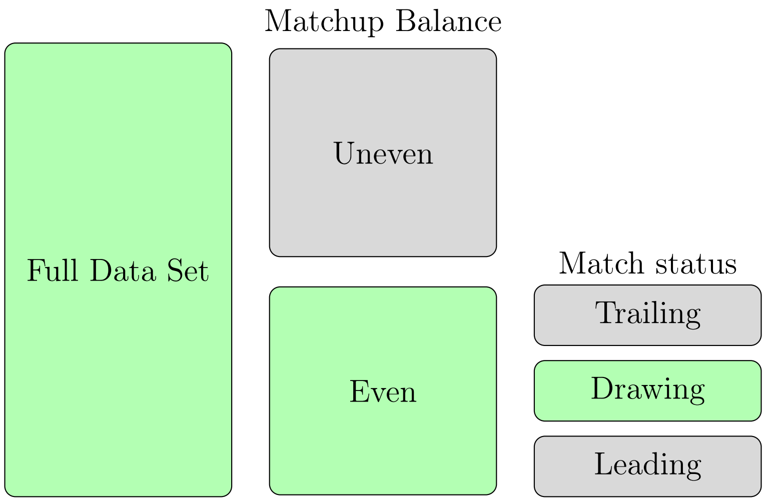


Figure 4.3: The Subset Used to Control for the Combination of Match-up Balance and Match Status Visualized.

Figure 4.3 shows the subset of passes played during *Match Status* drawing in matches with even *Match-up Balance*. This subset combining both controls, referred to simply as the **controlled subset** from here on out, will be used to once more compare the number of passes played grouped by the categorical variable outcome.

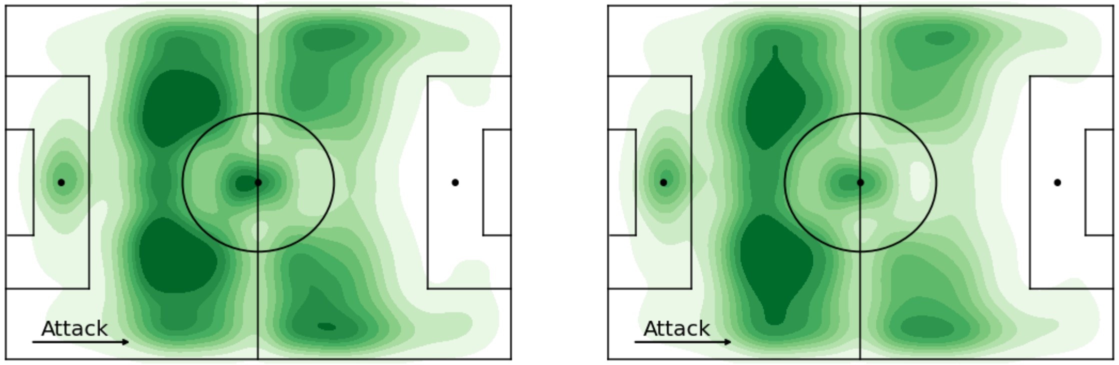
Table 4.5: Number of Passes per Outcome for Match-up Balance Even and Match Status Drawing.

Outcome	Passes played	Mean Possession
Lost	110,847	0.473
Drew	162,108	0.5
Won	123,979	0.527

Table 4.5 shows the possession shares of winners and losers on the *controlled subset*. On this subset of the data, winners do hold more possession than losers. However, the effect is not significant at level $a = 0.05$ with $\hat{p} = 0.114$.

4.3 Kernel Density Estimation of Passes

The goal of this section is to identify in which parts of the pitch possession has the biggest effect on the outcomes of football matches. A way to do this is by comparing the distributions of the passes of winners and losers of these matches. A method that provides an intuitive visual representation of these densities is KDE. We applied a 2-dimensional KDE with a Gaussian Kernel and Scott's rule for tuning bandwidth parameter h to all the passes of winners and losers in the *controlled subset* and visualized the density estimates on a football pitch. Additionally, the normalized densities were subtracted from each other, to help visualize these differences. The same steps were taken for the different sub-types of passes. The results can be found in Appendix A.



(a) Winners

(b) Losers

Figure 4.4: Density Estimates of Winners (a) and Losers (b) in the Controlled Subset Compared. Darker shades of green correspond to higher density estimates.

Figure 4.4 shows the KDEs of all passes played of Winners and Losers in the *controlled subset*. The density estimates differ in multiple ways. Winners play a higher share of their passes in the opponents half. They also play more passes in the centre of the pitch, especially during build up. Losers have a higher density in front of their own goal and on the left and right ends of the pitch. To further highlight the differences between the two distributions, the density estimates were normalized and the estimate of the losers was subtracted from the one of the winners.

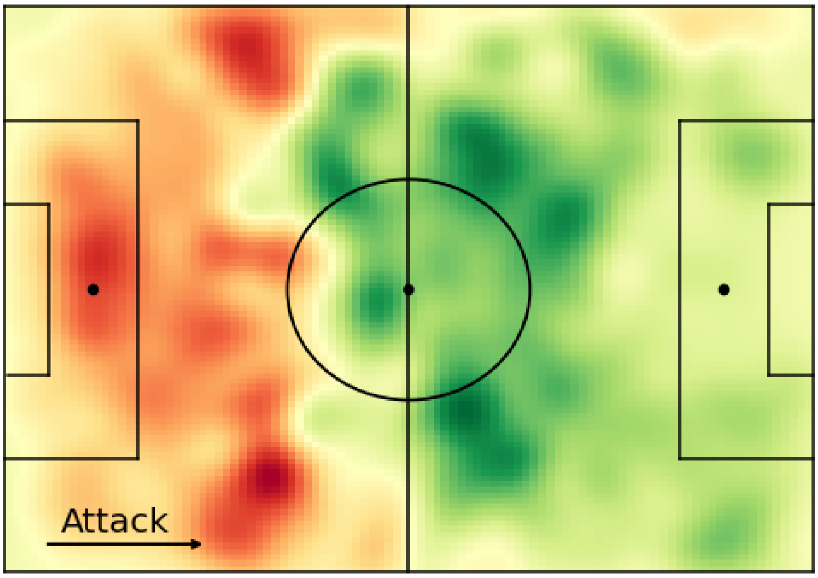


Figure 4.5: Difference Between the Normalized Density Estimates of Winners and Losers in the Controlled Subset. The color scale is anchored at the color yellow for equal densities. Green represents higher density for winners, red for losers. Darker shades of each color correspond to larger density differences.

The picture becomes even more clear in Figure 4.5, which shows the difference in density estimates between winners and losers. In the opponents half, winners play a higher share of their passes in all areas except the ones on the very left and right of the pitch. This surplus of density extends into their own half, for central areas close to the kick off point. The biggest density surplus of winners can be found around the centre of the pitch. Losers play a higher share of their passes in front of their own goal and in the areas close to the left and right edges of the pitch. The biggest surplus of density for Losers can be found in the areas deep in their own half.

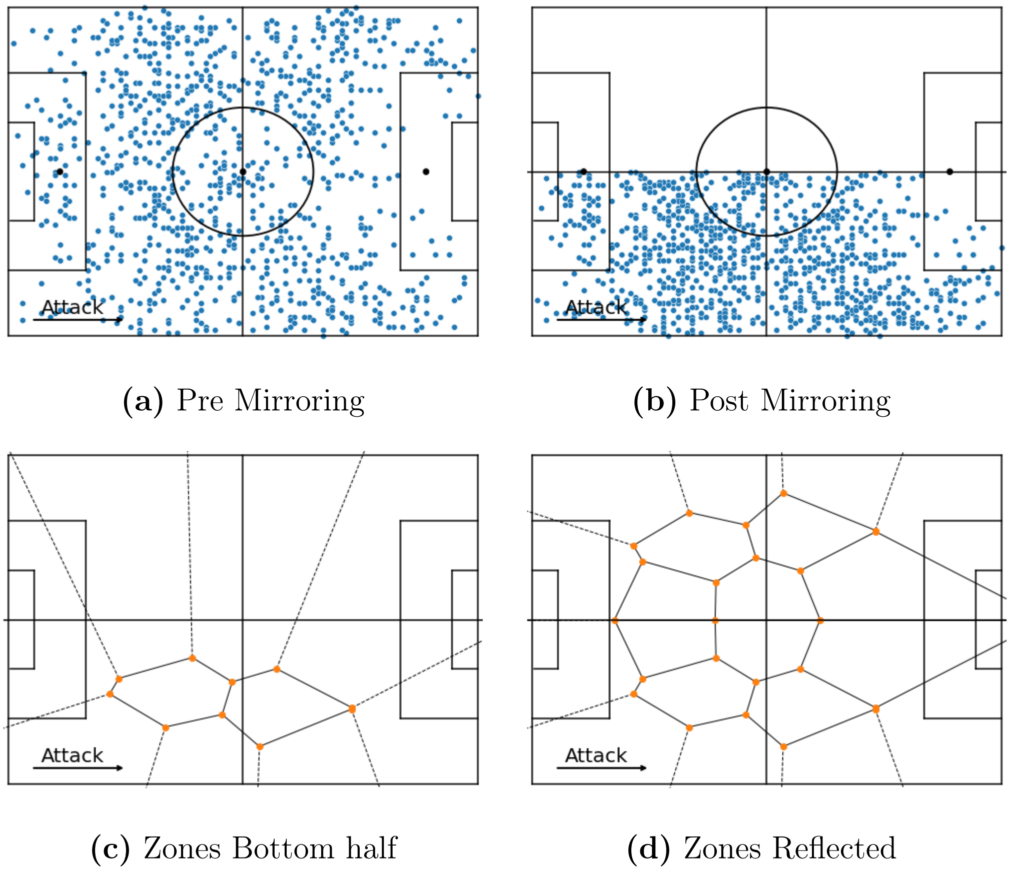
Chapter 5

Zone Definition

Section 4.3 revealed that there are differences in the effect of outcome between passes given their position on the pitch. From this follows a need for differentiation of passes given their position. A way of achieving such a differentiation is by dividing the football pitch into distinct zones. Two approaches for defining such zones and the necessary preprocessing steps are described in this chapter. Once the reader knows how the zones are defined, the mean possession per zone on the *controlled subset* will be given in Section 5.4.

5.1 Mirroring Passes along the Horizontal Axis of the Pitch

There are various formations in football and they differ in many ways. Famous examples are 4-4-2, 4-2-3-1, 3-5-2 or 4-3-3. What all of the above mentioned formations have in common is that they are symmetric along the horizontal axis of the football pitch. Very rarely one will find a team that decides to play with a left-back but no right back, or a right winger but no left winger. This symmetry is exploited to reduce the number of zones to define, while simultaneously increasing the number of passes available per zone.



(a) Pre Mirroring

(b) Post Mirroring

(c) Zones Bottom half

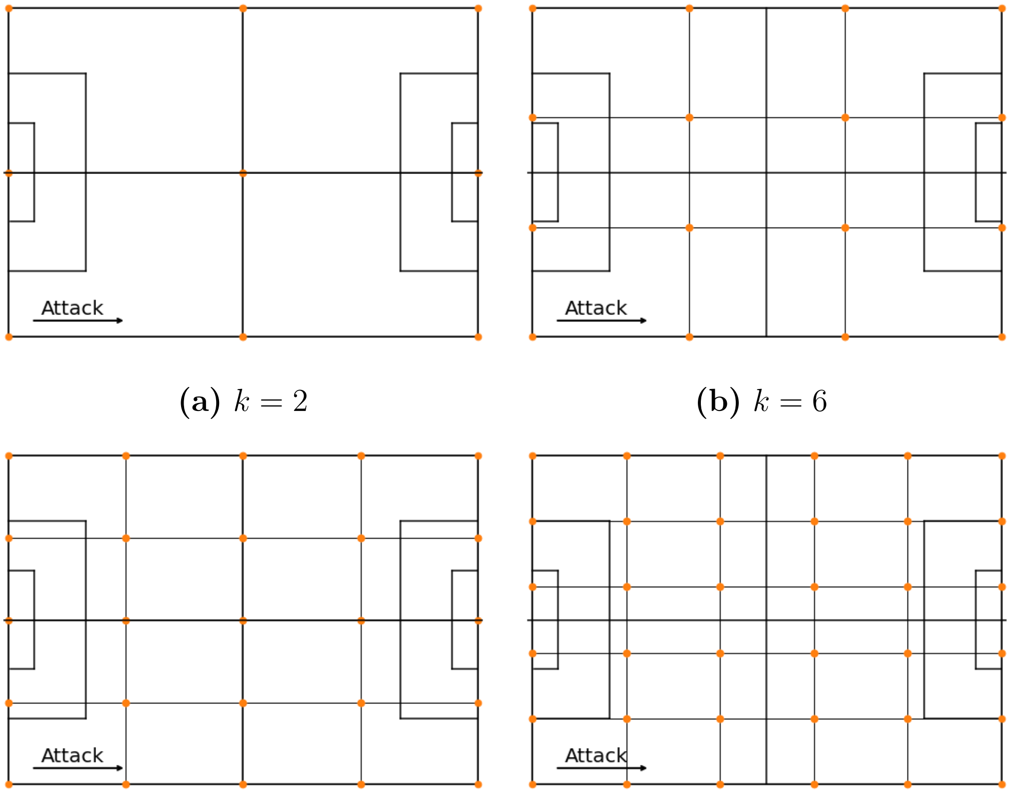
(d) Zones Reflected

Figure 5.1: The Mirroring Process Visualized. (a) shows a sample of n = 1000 passes in the data set plotted on the football pitch. (b) shows the same passes after mirroring along the horizontal axis. (c) shows the Voronoi zones resulting from a k-means clustering with k=10 on those sample passes. (d) shows the final zones after reflecting the zones back along the horizontal axis.

Figure 5.1 visualizes this process. Through the exploitation of the symmetry of the passes the number of zones to defines is halved while the available sample size per zone is doubled. This way, stability of the solutions is greatly improved, both for the zone definition based on k-means in Section 5.3 as well as the linear regression in Chapter 6.

5.2 Rectangles

A straightforward way of segmenting a football pitch into zones is by dividing it into rectangles of equal size. This method allows for differentiation between different combinations of vertical and horizontal progression. The resulting zones of such a segmentation are referred to as *rectangular zones* from here on.



(a) $k = 2$ **(b)** $k = 6$

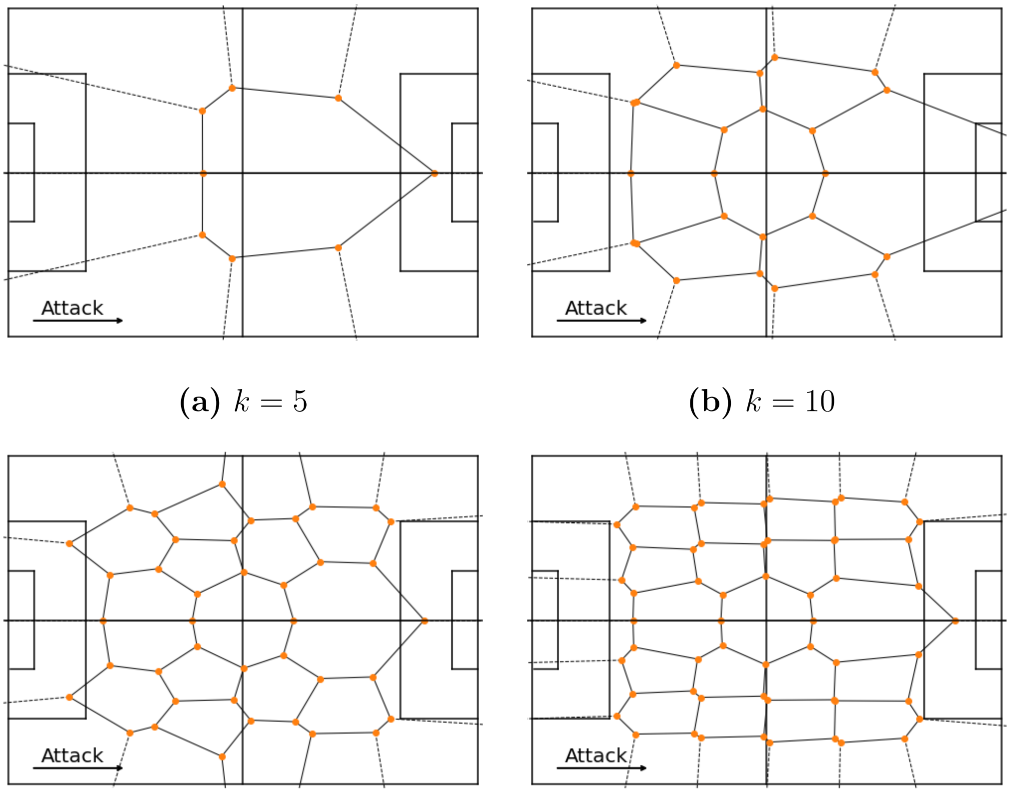
(c) $k = 8$ **(d)** $k = 15$

Figure 5.2: Rectangular Zones Visualized. (a), (b), (c) and (d) show the resulting zones after reflection on the x-axis for $k = 2$, $k = 6$, $k = 8$, and $k = 15$ respectively.

Figure 5.2 shows the *rectangular zones* resulting for different levels of k. One limitation of this kind of zone definition is that it does not allow for differentiation of zone size.

5.3 Voronoi Cells

A popular approach to segment the football pitch into zones is by using Voronoi cells based on the centroids of a K-means clustering. For k clusters, the Voronoi cell of each cluster k is defined as the area for which the euclidean distance to k is smaller than to any other cluster. This approach has the advantage of creating zones of different shapes and sizes, grouping similar passes together.



(a) $k = 5$

(b) $k = 10$

(c) $k = 15$

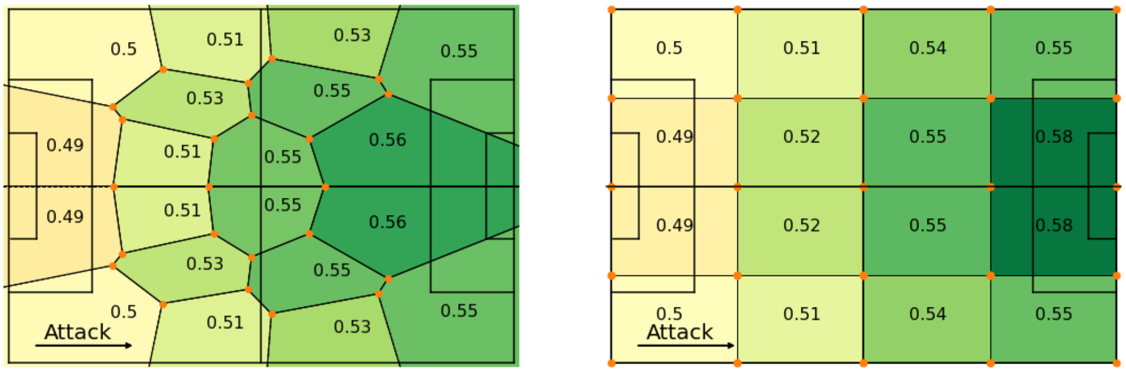
(d) $k = 20$

Figure 5.3: Voronoi Zones Visualized. (a), (b), (c) and (d) show the resulting zones after reflection on the x-axis for $k = 5$, $k = 10$, $k = 15$, and $k = 20$ respectively.

Figure 5.3 shows the zones based on the Voronoi cells resulting from the cluster centers of a k-means algorithm for different levels of k. From here on, these zones will be referred to as the *Voronoi zones*. Contrary to the rectangular zones, they vary in shape and size for each level of k. For lower levels of k, the *Voronoi zones* differ significantly from the rectangular ones. For higher levels of k, *Voronoi zones* in the middle of the pitch are close to the rectangular zones, but the zones close to both goals are still significantly different between both methods.

5.4 Possession per Zone

Now that two methods of segmenting the pitch into zones are established, in this section, these two methods are applied to explore how the possession values per outcome from Section 4.2.3 are distributed among the resulting zones.



(a) Voronoi, $k = 10$

(b) Rectangular, $k = 8$

Figure 5.4: The Mean Possession Per Zone in the Controlled Subset. (a) shows a Voronoi zone definition with 10 zones and (b) a rectangular zone definition with 8 zones. The color scale is anchored at the color yellow for a weight of 0. Green represents a zone with higher possession for winners, red a zone with higher possession for losers. Darker shades of each color correspond to larger possession differences.

In Figure 5.4 the possession of winners for the *controlled subset* is visualized. The figure confirms the findings from Figure 4.5, showing that winners play a higher share of their passes in the zones in the opponents half. Note that while Figure 5.4 and Figure 4.5 allow for the same conclusions, they represent different things. Figure 4.5 visualized the difference in densities between the passes played per outcome. As winners hold higher possession in general (0.527), they might play more passes in a given zones than losers, even if they play a lower share of their passes in this zone. This can be observed in most of the zones in the attacking teams own half. While winners have possession less than 0.527, they still hold higher or equal possession than losers. An exception to this are the zones right in front of their own goal. Here, winners hold lower possession than losers.

Comparing the two approaches for zone definition, small differences can be observed, especially in the opponent half. During the hyper-parameter tuning in the inner loops of N-CV in Chapter 7.1, both approaches will be compared and the best performing one will be selected for our final model.

Chapter 6

The Model

Now, that the need for spatial and temporal context has been established and tools for incorporating such context have been developed, it is time to formulate a model based on these findings. In this chapter, the multiple linear regression model underlying our new metric and its notation will be introduced.

6.1 Model Design

Our data set contains information about n football matches between two teams. Each row $i = 1, 2, ..., n$ contains information about home team h and away team a.

Let y_i be defined as the goal difference from the perspective of the home team,

$$y_i = G_{\text{i,h}} - G_{\text{i,a}},$$

with $G_{\text{i,h}}$ and $G_{\text{i,a}}$ being the amount of goals scored per team.

Let X_i be a vector of size k, with k being a number of separated zones on the football pitch and $NP_{\text{i,h,k}}$ and $NP_{\text{i,a,k}}$ be the number of passes played per team in each of these k zones. The possession for match i of home team h per zone k is then defined as:

$$P_{\text{i,h,k}} = \frac{NP_{\text{i,h,k}}}{NP_{\text{i,h,k}} + NP_{\text{i,a,k}}} - \mu(P_{\text{h,k}}).$$

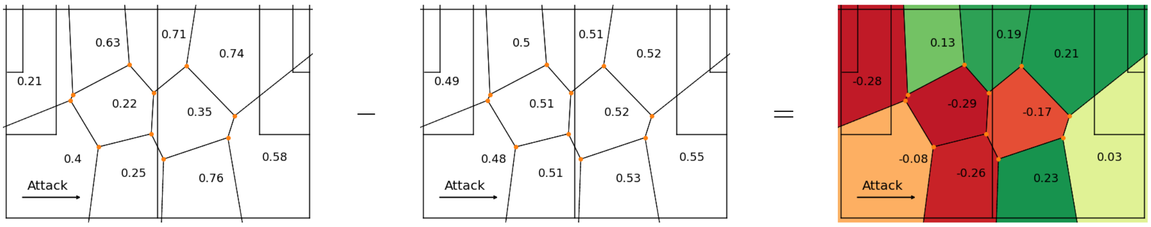
Then $X_{\mathrm{i,k}}$ will contain the mean centred possession surplus per zone of home team h, defined as:

$$X_{\mathrm{i,k}} = P_{\mathrm{i,h,k}} - \mu(P_{\mathrm{h,k}}), \tag{6.1}$$

with $\mu(P_{\mathrm{h,k}})$ being the mean possession of home teams in that zone over all n matches. Our linear regression equation for one match i now is:

$$y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_k X_{i,k} + \epsilon_i,$$

with $\beta_0, \beta_1, ..., \beta_k$ being the intercept and the coefficients for our parameters and ϵ_i being the random error term with distribution $N \sim (0, \sigma^2)$.



$P_{\mathrm{i,h}}$ $\mu(P_{\mathrm{h}})$ $X_{\mathrm{i,h}}$

Figure 6.1: The Composition of X_i for the Basic Version of the Model with $k = 10$ Voronoi Zones Visualized. The color scale is anchored at the color yellow for a a mean centred possession value of the home team of 0. Green represents a zone with a possession surplus for the home team, red a zone with a surplus for the away team. Darker shades of each color correspond with a larger absolute size of possession differences.

Figure 6.1 visualizes Eq. 6.1 for a sample match for a model with $k = 10$ and *Voronoi zone* definition. As our model is defined from the home teams perspective and mean centred by the mean home possession per zone, intercept β_0 can be interpreted as the home field advantage. In its simplest form of $k = 1$, this model resembles a regression of raw possession on goal difference, taking home field advantage into account.

6.1.1 Controlling for Match Status

As outlined during this thesis, *Match Status* has an effect on expected possession and therefore needs to be considered. Therefore we extend our model the following way:

Define match status $s \in [l, d, w]$ as a dummy variable indicating the three possible *Match Statuses.*

Let X_i be a vector of size $k \times 3$, with k being the number of zones on the football pitch. The possession for match i of home team h per zone k, per status s is then defined as:

$$P_{\text{i,h,k,s}} = \frac{NP_{\text{i,h,k,s}}}{NP_{\text{i,h,k,s}} + NP_{\text{i,a,k,s}}} - \mu(P_{\text{h,k,s}}).$$

$X_{\text{i,k,s}}$ will now contain the mean centred possession for match i, for zone k, for status s defined as

$$X_{\text{i,k,s}} = P_{\text{i,h,k,s}} - \mu(P_{\text{h,k,s}}), \tag{6.2}$$

with $\mu(P_{\text{h,k,s}})$ being the mean possession of home teams h, for zone k, during status s, over all n matches. The linear regression equation now becomes:

$$\begin{aligned} y_i = \beta_0 &+ \beta_{1,l}X_{i,1,l} + \beta_{1,d}X_{i,1,d} + \beta_{1,w}X_{i,1,w} \\ &+ \beta_{2,l}X_{i,2,l} + \beta_{2,d}X_{i,2,d} + \beta_{2,w}X_{i,2,w} \\ &+ \dots \\ &+ \beta_{k,l}X_{i,k,l} + \beta_{k,d}X_{i,k,d} + \beta_{k,w}X_{i,k,w} + \epsilon_i. \end{aligned}$$

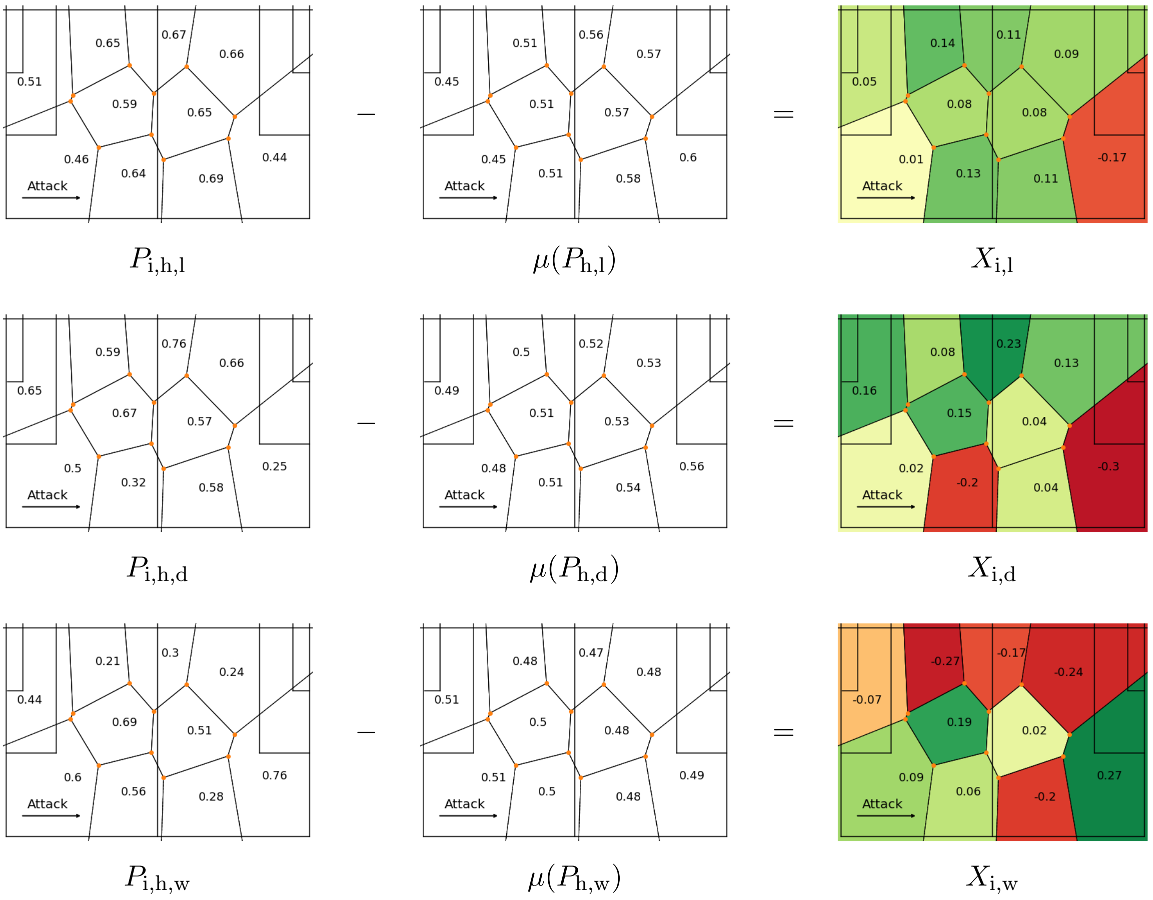


Figure 6.2: The Composition of X_i for the Match Status Controlled Version of the Model with $k = 10$ Voronoi Zones Visualized. The color scale is anchored at the color yellow for a a mean centred possession value of the home team of 0. Green represents a zone with a possession surplus for the home team, red a zone with a surplus for the away team. Darker shades of each color correspond with a larger absolute size of possession differences.

Figure 6.2 visualizes Eq. 6.2 for a sample match for a status controlled model with $k = 10$ and *Voronoi zone* definition.

6.1.2 Dealing with Missing Values

While all matches are drawn for at least some amount of time, the *Match Statuses trailing* and *leading* are not guaranteed to happen and are therefore not always available. Therefore, for the status controlled version of the model, missing values will be imputed by $P_{\text{i,h,d}}$, the mean centred possession per zone k, during status d.

Chapter 7

Results

In this chapter, we will show the results of the nested cross-validation used to fit the model and its hyper-parameters. The results will be compared to multiple baselines, both for the full test sets as well as a subset of even matches. Once the final model is selected, we will visualize its weights.

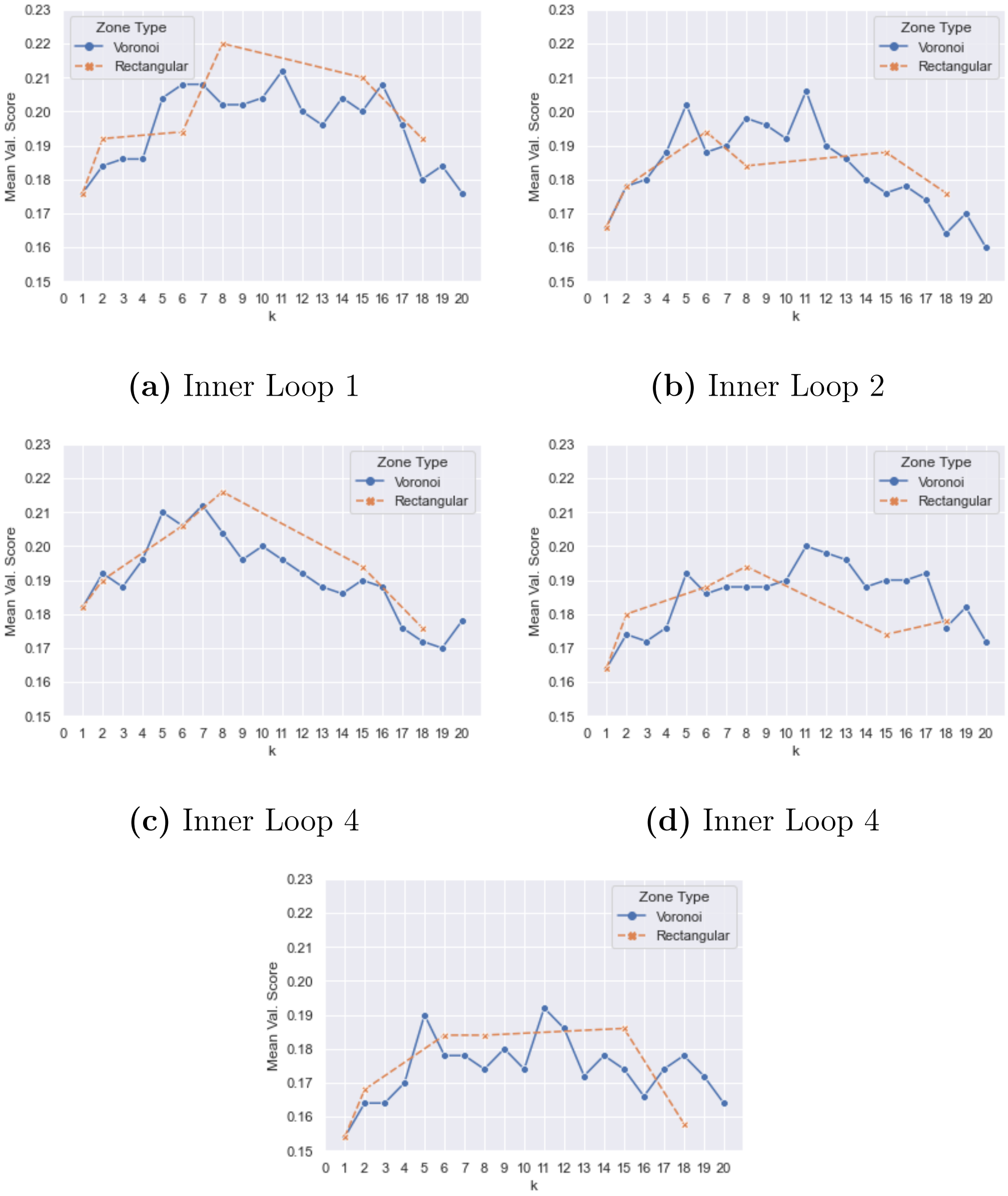
7.1 Results of Nested Cross-Validation

In this section, the results of the Nested Cross-validation (N-CV) are presented. The Inner Loops of the N-CV will serve as a method for model selection, choosing the best performing combination of zone type (*Voronoi* or *rectangular*) and number of zones k. The Outer Loop of the N-CV will allow us to estimate the stability of our model as well as its test performance.For readability purposes, only aggregated results are displayed in this chapter. The full results for all $l \times m = 25$ folds can be found in Appendix C.

7.1.1 Results Inner Loops

For each of the $l = 5$ folds in each fold of each Inner Loop, the model was fit for k in the range of $[1, 2, ..., 20]$ for the *Voronoi zones*. For the *rectangular zones*, all values in the range $[1, 2, ..., 20]$ that allowed for a separation of the pitch into squares of equal size post-mirroring were chosen, resulting in values of k of $[1, 2, 6, 8, 15, 18]$.

For each of the 5 inner loops, the r^2 on the validation set was recorded, and the model with the highest mean score on all 5 folds was chosen to be tested in the respective fold in the Outer Loop.



(a) Inner Loop 1

(b) Inner Loop 2

(c) Inner Loop 4

(d) Inner Loop 4

(e) Inner Loop 5

Figure 7.1: Mean Score per k During the $l = 5$ Folds of each Inner Loop. (a), (b), (c) and (d) and (e) show the mean score aggregated over the $m = 5$ folds, for each of the $l = 5$ Inner Loops.

Figure 7.1 shows the mean score per k over the 5 folds of the Inner Loop, for all 5 folds of the Outer Loop. Starting off at $k = 1$ we see an increase in validation scores until a inflection point is hit, from which the validation score decreases. This general trend holds across all inner loops, but inflection points and steepness of increase and decreases differ. The two zone definitions perform similar inside each inner loop. For each of the two definitions, the k resulting in the highest score is selected for the respective inner loop.

Table 7.1: Test Scores and Chosen Hyper-parameter k for Inner Loops.

	Inner 1		Inner 2		Inner 3		Inner 4		Inner 5	
Zone Type	k	Score	k	Score	k	Score	k	Score	k	Score
Rectangular	**8**	**0.218**	6	0.192	**8**	**0.215**	8	0.193	15	0.187
Voronoi	11	0.212	**11**	**0.206**	5	0.2111	**12**	**0.2**	**11**	**0.191**

Table 7.1 shows the chosen model configurations and the mean validation score for that configuration per Inner Loop. Both methods of zone definition perform on a similar level, with validation scores of around 0.2. For both methods, the Inner Loops agree on k for 3 of the 5 loops.

Table 7.2: Aggregated Results Inner Loops.

Zone Type	Mode k	Mean Score
Rectangular	8	0.201
Voronoi	**11**	**0.204**

Table 7.2 shows the mode of hyperparameter k and the mean score aggregated over all $l = 5$ inner folds. Both types of zone definitions perform on a similar level. As the zone definition based on Voronoi cells achieved the higher mean score, we will proceed to the outer folds with the *Voronoi zones*.

7.1.2 Results Outer Loop.

In this section, the results on the test sets of the $m = 5$ folds in the outer loop are discussed. In addition to the full model resulting from the hyper-parameters chosen in each inner loop, the results of two baseline models will be tested. The first baseline, referred to as the raw model from here on, is a model with $k = 1$ and no control for *Match Status*. This model is equivalent to linear regression model with an intercept and one parameter for the raw possession value of the home team.

The second baseline, referred to as the status model from here on, is a model with $k = 1$ and control for *Match Status*. This model is equivalent to linear regression model with an intercept and three parameters for the possession value of the home team during the three possible *Match Statuses*.

Table 7.3: Results Outer Loop

Model	Outer 1	Outer 2	Outer 3	Outer 4	Outer 5	Mean Score
Raw	0.017	0.056	0.025	0.038	0.084	0.044
Status	0.148	0.185	0.129	0.172	0.244	0.175
Full	**0.204**	**0.239**	**0.155**	**0.196**	**0.247**	**0.208**

Table 7.3 shows that the full model outperforms both baselines in each of the 5 outer folds. The difference in results between the 3 model types shows a similar pattern across outer folds, with the status model outperforming the raw model in test score by approximately 0.13 and the full model by approximately 0.16. When comparing the results of the full model on the test set to the ones on the validation sets in Table 7.1, similar results can be observed, with the score on the test set being slightly higher. Therefore, the model does not show signs of overfitting.

7.1.3 Results on Subset of Matches with Match Status Even.

As a way to control for *Match-up Balance*, the testing of the outer folds of the Nested CV was repeated on the subset containing only the matches with *Match-up Balance 'even'* per fold. Once again, the same is done for the baselines raw model and status model. The results are presented in this subsection.

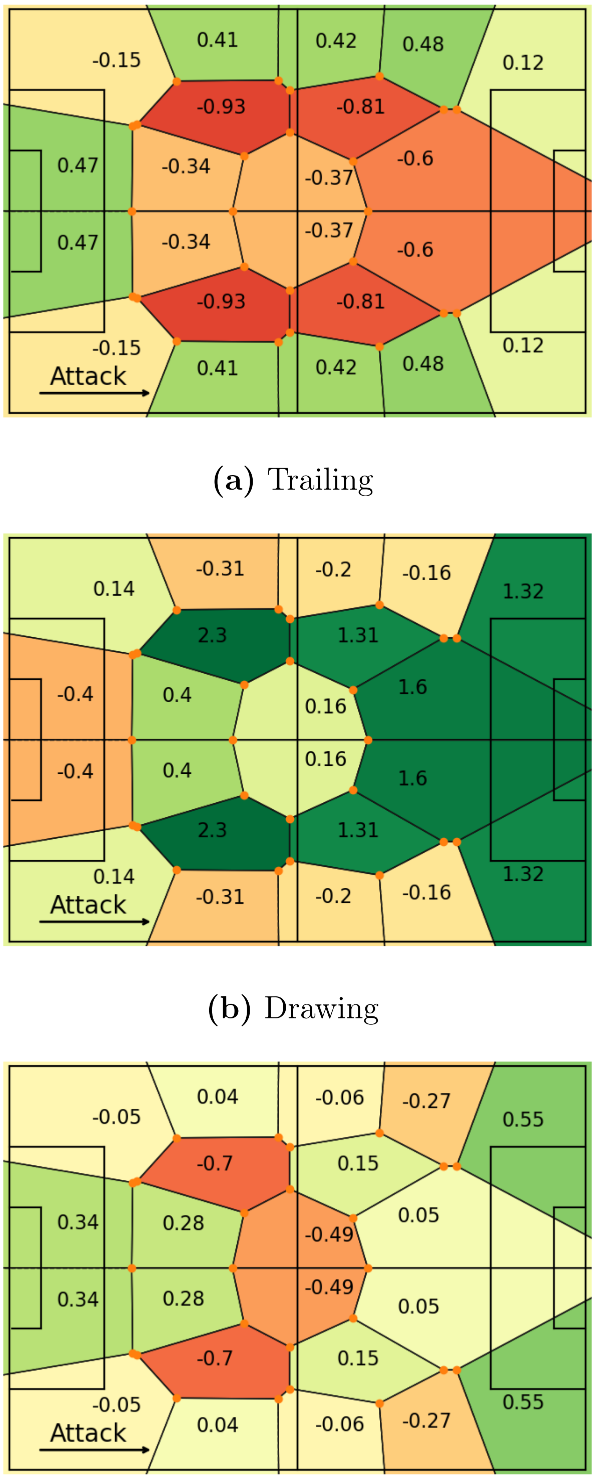
Table 7.4: Results Outer Loop on Even Match-up Balance Subset.

Model	Outer 1	Outer 2	Outer 3	Outer 4	Outer 5	Mean Score
Raw	-0.183	-0.145	-0.158	-0.170	-0.049	-0.141
Status	-0.033	-0.002	-0.160	-0.008	0.131	-0.014
Full	**0.049**	**0.020**	**0.064**	**0.037**	**0.173**	**0.069**

Table 7.4 shows the test results on the *even* subset. Once again, the full model scores highest on all folds. On the even subset, the test score for the raw model turns negative for all folds. The status model only manages to achieve a positive score on outer fold 5. The full model is the only model achieving positive scores on all 5 folds. Similar to the results in Table 7.3, the status model outperforms the raw model by approximately 0.13. The out-performance of the full model increased to approximately 0.2.

7.2 Final Model

After tuning the hyper-parameters through N-CV it is now time to fit the final model on the full data set. While the 5 values of k resulting from the N-CV are to be treated equally, a decision on which one to choose for the final model fit has to be made. We choose $k = 11$, as 3 of the 5 inner loops agreed on this value of k.



(a) Trailing

(b) Drawing

(c) Leading

Figure 7.2: The Weights per Zone of the Final Model Visualized. (a), (b) and (c) show the weights per zone for match status trailing, drawing and leading respectively. The intercept, representing the home field advantage, is $\beta_0 = 0.32$. The color scale is anchored at the color yellow for a weight of 0. Green represents a zone with positive weights, red a zone with negative weights. Darker shades of each color correspond with a larger absolute size of weights.

Figure 7.2 visualizes the coefficients of the final model on the football pitch. The weights have been reflected back along the horizontal axis to aid visualization.

It can be observed that the absolute size of weights is biggest for possession while *drawing*. The zones with the biggest positive effect for match status drawing are the centre left and centre right zones in a teams own half, followed by the zones deep inside the opponents half. Above average possession in the zones in front of a teams own goal and on the left and right edges of the pitch has a small negative effect on goal difference. For match statuses *trailing* and *leading*, the average absolute size of weights is much smaller.

For *trailing* teams, above average possession in the zones in front of their own goal and on the left and right edges of the pitch, have a slight positive effect on goal difference. Possession in the centre of the pitch holds negative weights for trailing teams. *Leading* teams have small positive weights for passes in front of their own goal as well as deep inside the opponents half.

The intercept β_0 of the final model equals 0.32. As the outcome of our model is goal difference, this means that home teams score 0.32 goals more than away teams across the whole data set. The weights per zone can be interpreted as follows: As X_i is a mean centred vector with the possession per zone, the weights will be multiplied with the deviation of a teams possession in a given match to the mean possession of all teams for the same zone and status. E.g. if a team hold 20% higher possession than the average team in the two central zones in front of the opponents goal in sub-figure (b), the model predicts this surplus in possession to lead to a $0.2 \times 1.6 = 0.32$ surplus of goals scored by that team.

Chapter 8

Applications

At this point we have obtained our final model and it is time to showcase its potential in practice. In this chapter we will transform the output of our model into ratio format familiar with raw possession. Once the transformation is done we will show how my metric can be used to analyze teams playing styles on the aggregated season level, on the example of the 2017-2018 La Liga season. Then, we will show how our metric can be applied on the individual match level, showing its full potential over multiple match statuses.

8.1 Transformation

In its current form, our model predicts the goal difference from the perspective of the home team. Theoretically, this prediction can span anywhere on the range of $[-\infty, \infty]$. Raw possession is usually reported as a fraction on the range $[0, 1]$. In this section we will propose a method to transform our models prediction into a range of $[0, 1]$.

This transformation has two advantages. First , it will be easier to report our metric to the casual audience when it comes in a format familiar to them. Second, it will aid the comparison between our metric and raw possession in the rest of this chapter.

A common way to transform a number in the range of $[-\infty, \infty]$ to a range of $[0, 1]$ is by using a sigmoid function. A member of the family of sigmoid functions is any function that is a bounded, differentiable, real and defined for all real input values and has a non-negative derivative at each point and exactly one inflection point (see [14]). We will make use of one of those members in the form of the logistic function

$$f(p) = \frac{1}{1 + e^{-s(p)}},$$

with p being our models prediction and s being a scaling parameter. The scaling parameter s will be chosen by minimizing the distances between the empirical distributions of our transformed metric and raw possession. This is done by minimizing the F-Statistic of the Kolmogorov-Smirnov test for goodness of fit as described in Massey et al. [15].

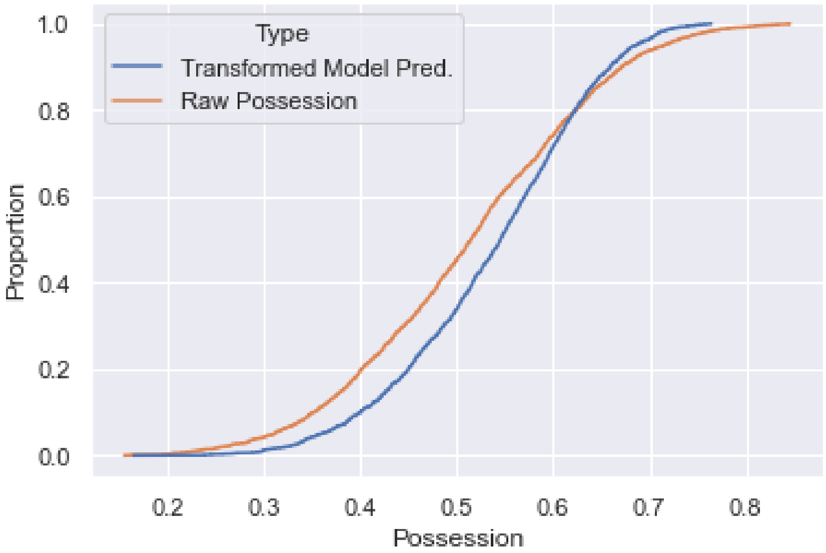


Figure 8.1: Comparison of Cumulative Distribution Functions of Raw Possession and our Transformed Model Prediction. Our models predictions were transformed through a logistic function with scale parameters s = 0.455.

Figure 8.1 compares the empirical cumulative density functions of our transformed metric and raw possession. As a result of the minimization, scaling parameter s was set at 0.455. We observe that the distribution of our transformed metric is less wide than raw possession, with less density at the extreme possession values.

8.2 Application on Aggregated Team Level

Now that our metric and raw possession are on the same scale, it is time to compare the two in practice. We start off by looking at the final standing of the 2017-2018 La Liga season with the mean levels for both raw possession and our transformed metric.

Table 8.1: Final Table of La Liga for the 2017-2018 Season. The dotted lines indicate the cutoffs for Champions League qualification and relegation. The teams in bold will be used for further analysis in this Section.

Position	Club Name	Goal Diff.	Points	Raw Poss.	Trans. Pred.
1	Barcelona	70	93	0.64	0.60
2	**Atlético Madrid**	**36**	**79**	**0.47**	**0.52**
3	Real Madrid	50	76	0.61	0.60
4	Valencia	27	73	0.49	0.52
5	Villarreal	7	61	0.51	0.49
6	Real Betis	-1	60	0.58	0.54
7	Sevilla	-9	58	0.55	0.53
8	Getafe	9	55	0.39	0.45
9	Eibar	-6	51	0.52	0.55
10	Girona	-9	51	0.46	0.49
11	Espanyol	-6	49	0.46	0.47
12	Real Sociedad	7	49	0.57	0.53
13	Celta de Vigo	-1	49	0.57	0.51
14	Deportivo Alavés	-10	47	0.39	0.44
15	Levante	-14	46	0.42	0.44
16	Athletic Club	-8	43	0.49	0.51
17	Leganés	-17	43	0.42	0.46
18	La Coruña	-38	29	0.46	0.46
19	**Las Palmas**	**-50**	**22**	**0.55**	**0.48**
20	Málaga	-37	20	0.44	0.43

Table 8.1 shows the final table of the 2017-2018 La Liga Season. We can observe that while raw possession correlates with a high amount of points and a positive goal difference for some teams, there are notable exceptions. Atlético Madrid as well as Valencia finished in the top 4 with a clearly positive goal difference while having less mean possession than their opponents over the course of the season.

The same phenomenon can be found on the other end of the table. Las Palmas finished the season with only in second to last place with only 20 points and the worst goal difference in the whole league of -50. Nonetheless, they hold significantly more raw possession than their opponents on average. Our metric comes to a different conclusion for these outliers, more in line with the final standings. In the upcoming subsections we will show the reasons for the higher predictive power of our metric on the example of Atlético Madrid and Las Palmas.

8.2.1 Atlético Madrid

Atlético Madrid is well known for its unique play style under their coach Diego Simeone. While most top teams usually play an attacking style with high possession over the whole course of the match, Diego Simeone's Atlético Madrid became famous for a play style relying on efficient offensive and world class team-wide defense. This style is so successful that Atlético is the only team in the last 15 years that managed to beat the giants Real Madrid and FC Barcelona over the course of a whole season, by winning the Spanish title in 2014 as well as 2021 (see [16]).

In this section we will analyze this play style using our metric and showcase why raw possession is not fit well to describe it.

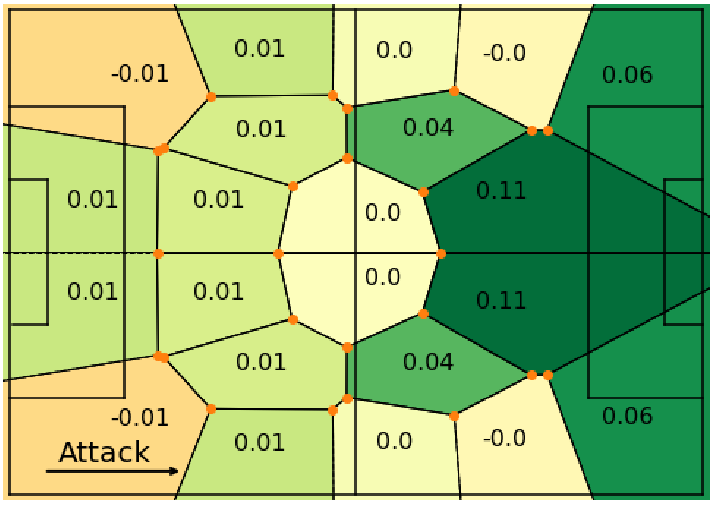
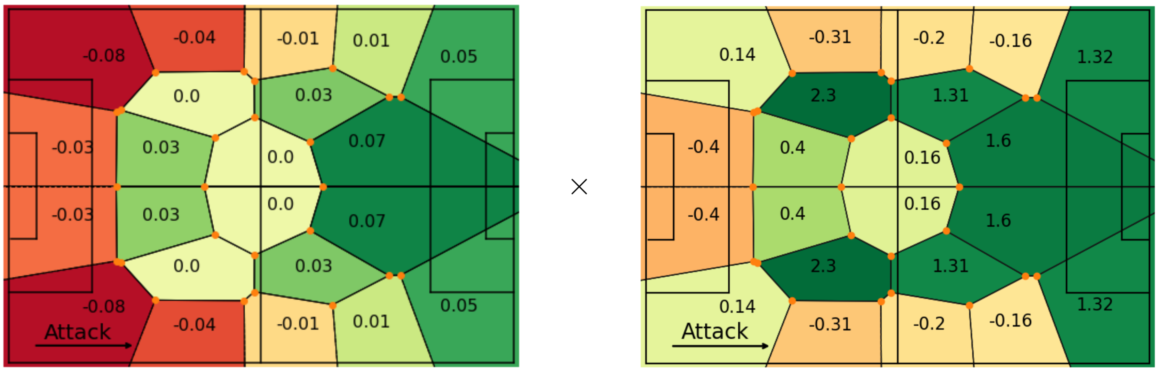


Figure 8.2: Our Final Models Prediction for Atlético Madrid's Performance for Status Drawing in the 2017-2018 La Liga Season. The color scale is anchored at the color yellow for a prediction of 0. Green represents a zone with a positive prediction, red a zone with a negative prediction. Darker shades of each color correspond with a larger absolute size of the prediction.

Figure 8.2 shows the mean untransformed model predictions for status *drawing* per zone for all of Atlético Madrid matches over the course of the 2017-2018 season. We see that the model predicts a positive goal difference based on Atlético's possession while *drawing*. As explained in Chapter 6, the model prediction seen in Figure 8.2 is the result of the multiplication of two components.



Mean Centred Possession for Drawing

Final Model Coefficients for Drawing

Figure 8.3: The Two Components of Atlético Madrid's Possession in the 2017-2018 La Liga Season. The left side shows the possession values mean centred per zone while drawing. The right side shows our final models weights for status drawing. Color scales are anchored at 0 with green colors representing positive and red colors representing negative values. Darker shades of each color correspond with larger absolute values.

Figure 8.3 shows these two components for Atlético Madrid aggregated over all matches of the season. The left side shows the mean centred possession per zone during status *drawing*. It can be interpreted the following way: The 0.07 in the most advanced central zones means that Atlético held 7% higher possession than the average team in our data set in this zone during status *drawing*. On the contrary, Atlético Madrid held 8% lower possession in the outer zones in front of their own goal. The right side shows the coefficients the final version of our model gives to possession in each zone during status *drawing*.

These two components help us explain why raw possession is not suited to explain Atlético's play style. Atlético's possession is highly concentrated in areas around the opponents goal, which our model weights as favorable. In front of their own goal, they hold possession much lower than the average team. As our model gives possesion in these zones a negative or only slightly positive weight, it manages to pick up on this unique distribution of possession. A simple aggregation of possession, without taking spatial context into account does not do Atlético's performance justice.

8.2.2 Las Palmas

Las Palmas played a disastrous 2017-2018 season resulting in the teams relegation. They finished the season with the second to worst goal difference across all La Liga teams in the last 10 years of the competition (see [17]).

The team was three different coaches during the season, with none of them being able to avoid the eventual relegation (see [18]). They have not been able to return to the first Spanish division ever since. Even though the season was a huge disappointment for the club, the club ranked tied 6th in mean raw possession over the 2017-2018 season. In this section we will show that the high possession of Las Palmas was not the sign of a dominant play style, but rather the opposite and explain how our metric manages to pick up on the low quality of the possession Las Palmas held all season long.

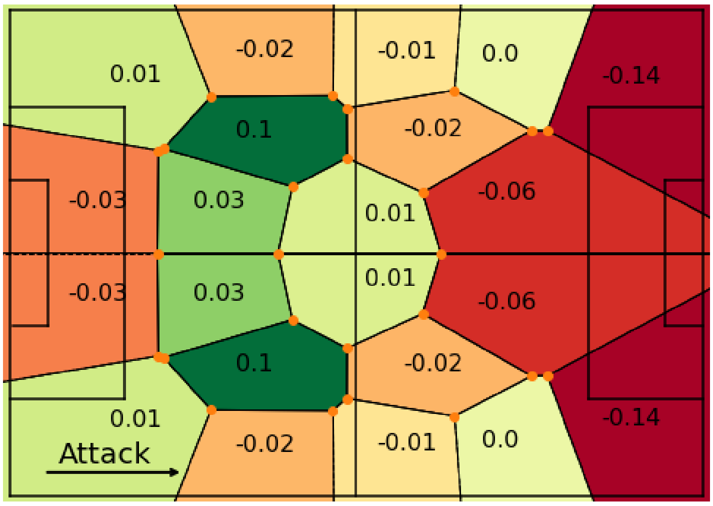
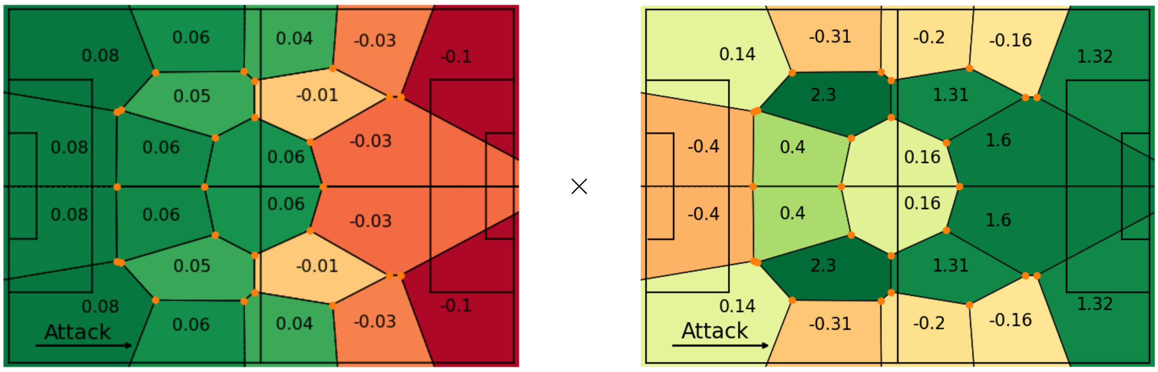


Figure 8.4: Our Final Models Prediction for Las Palmas' Performance for Status Drawing in the 2017-2018 La Liga Season. The color scale is anchored at the color yellow for a prediction of 0. Green represents a zone with a positive prediction, red a zone with a negative prediction. Darker shades of each color correspond with a larger absolute size of the prediction.

Figure 8.4 shows the mean untransformed model predictions for status *drawing* per zone for all of Las Palmas matches over the course of the 2017-2018 season. In contrast to Atlético's prediction, we see that our model predicts a negative goal difference based on

Las Palmas' possession while *drawing*. Once again, we will look at the two components of which this prediction is a product of.



Mean Centred Possession for Drawing

Final Model Coefficients for Drawing

Figure 8.5: The Two Components of Las Palmas' Possession in the 2017-2018 La Liga Season. The left side shows the possession values mean centred per zone while drawing. The right side shows our final models weights for status drawing. Color scales are anchored at 0 with green colors representing positive and red colors representing negative values. Darker shades of each color correspond with larger absolute values.

Figure 8.5 shows these two components for Las Palmas aggregated over all matches of the season. We observe that Las Palmas holds a big surplus of possession compared to the average team in the zones with little to no vertical progression. On the contrary, Las Palmas hold 10% lower possession that the average team in the outer zones with the highest vertical progression. Once again, a simple aggregation of possession without spatial context is not fit well to describe Las Palmas problems. While they do hold a lot of possession, they hold it in zones of the field that do not correlate with winning football matches.

8.3 Application on Individual Match Level

Now, that we have shown how to apply our metric to describe and analyze a teams play style on the aggregated level, we will show how to use it to analyze single matches. First, we will look at the 20 matches for which raw possession and our transformed model prediction had the biggest disagreement on. Then, we will pick an especially interesting example among those matches and will analyze the performance per match status for the teams involved.

Table 8.2: The 20 Matches during the 2017-2018 Season across the Top 5 Leagues with the Biggest Absolute Difference Between Raw Possession and our Models Transformed Prediction. Raw possession and the transformed model prediction are stated from the perspetive of the home team. The table is sorted by absolute difference.

Match-up	Match-day	Result	Raw Poss.	Trans. Pred.	Diff.
Las Palmas - Atlético Madrid	2	1 - 5	0.64	0.24	-0.39
Arsenal - Man. United	15	1 - 3	0.78	0.40	-0.38
Girona - Athletic Club	22	2 - 0	0.37	0.75	0.37
Málaga - Espanyol	18	0 - 1	0.61	0.24	-0.37
Saint-Étienne - Nice	1	1 - 0	0.40	0.76	0.36
Getafe - Las Palmas	17	2 - 0	0.32	0.67	0.35
Bologna - Sampdoria	14	3 - 0	0.31	0.65	0.34
Crotone - Sassuolo	35	4 - 1	0.40	0.74	0.34
Olymp. Marseille - Rennes	5	1 - 3	0.68	0.34	-0.34
Eintr. Frankfurt - Schalke 04	17	2 - 2	0.34	0.68	0.34
Stuttgart - M'gladbach	22	1 - 0	0.32	0.65	0.33
Levante - Athletic Club	15	1 - 2	0.61	0.28	-0.33
Udinese - Juventus	9	2 - 6	0.58	0.26	-0.32
M'gladbach - Eintr. Frankfurt	3	0 - 1	0.73	0.42	-0.32
Girona - Eibar	36	1 - 4	0.60	0.29	-0.31
Lille - Olymp. Marseille	11	0 - 1	0.66	0.35	-0.31
Torino - Napoli	17	1 - 3	0.38	0.69	0.31
Olymp. Marseille - Nantes	28	1 - 1	0.82	0.51	-0.31
Schalke 04 - Stuttgart	3	3 - 1	0.37	0.67	0.29
Real Betis - Real Madrid	**24**	**3 - 5**	**0.58**	**0.29**	**0.28**

Table 8.2 shows the 20 matches in the data set with the biggest difference between raw possession and the transformed model prediction. The possession values and results are given from the home teams perspective. We can observe that our metric shows much higher correlation with the final results. The biggest difference occurred during a match between Las Palmas and Atlético Madrid, the two teams analyzed in Section 8.2.

Many of the matches in this table have the winning team getting ahead very early in the match and holding on to the lead over the whole course of the match. This is expected, as trailing teams have significantly higher raw possession as outlined in Chapter 4. Raw possession has no way to account for this fact. Our model takes into account *Match Status* and is therefore able to differentiate between possession caused simply by *trailing* and possession caused by a dominant play style. A match with a particular interesting game history, including multiple *Match Status* switches and many goals is the match between Real Betis Sevilla and Real Madrid on match-day 24 of the 2017-2018 La Liga season. This match will serve as the example match to show off individual match analysis with out metric in the rest of this chapter.

8.3.1 Real Betis - Real Madrid

On the 18th of February 2018 Real Betis Sevilla hosted Real Madrid for a spectactular match featuring 8 goals. Real Madrid just came off a crucial home win against Paris Saint Germain in the Champions League that same week, with two late goals of Cristiano Ronaldo and Marcelo netting them a 3-1 win and a great position for the second tie. Real Betis Sevilla came into the match with confidence themselves, coming of a 0:1 away win against Deportivo La Coruna on the previous matchday. In this section we will analyze the match by looking at our models prediction for the different *Match Statuses*.

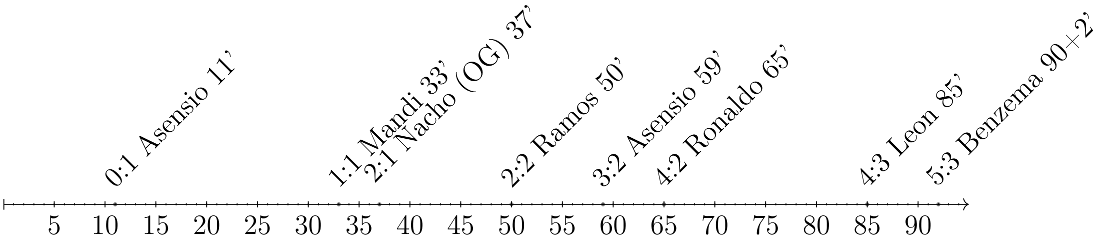


Figure 8.6: Timeline of the Match Real Betis Sevilla - Real Madrid on Matchday 24 of the 2017-2018 La Liga Season.

Figure 8.6 shows the timeline of the match. Real Madrid got an early lead through a header by Marco Asensio in the 11th minute, but Real Betis managed to return the favor, equalizing with a Mandi header in the 33th minute. An unfortunate own goal by Nacho caused by a reflection in the 37th minute meant that Real Betis went into half time with a lead. Real Madrid came out of the half time break with full force, managing to equalize in the 50th minute through another header, this time by Sergio Ramos. They kept dominating the match, taking the lead with goals by Asensio (59') and Ronaldo (65'). But Real Betis was not ready to go down without a fight, bringing their deficit down to one through a goal by Sergio Leon in the 85th minute. Their hopes of a second comeback in this match were stopped by Karim Benzema scoring the final goal of the match in a counter in stoppage time (90+2').

Given that the match featured all three *Match Statuses* it is a perfect example to show the full potential of our metric for match analysis. All predictions are shown from the perspective of the home team Real Betis. During the course of the match Real Betis held 58% raw possession. Our model predicts a goal difference of -1.98 resulting in a transformed prediction of 29%. We will go through our models prediction for each *Match Status* and explain why our model correctly picked up on a dominant performance by Real Madrid and why raw possession failed to do the same.

In this Section, we will only focus only on the models final prediction. If the reader is interested in a decomposition similar to the one in Section 8.2, it can be done with the help of the weights per *Match Status* in Section 7.2.

Trailing

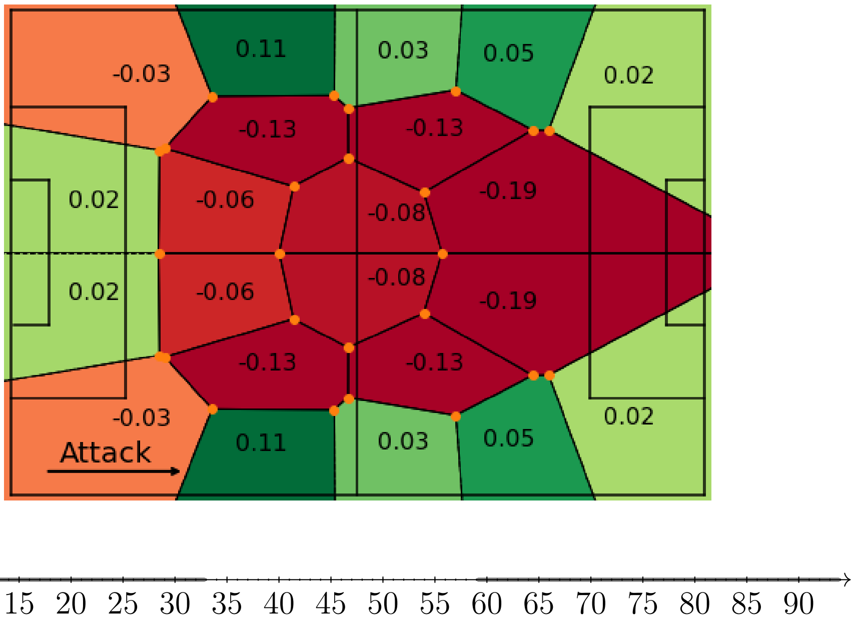
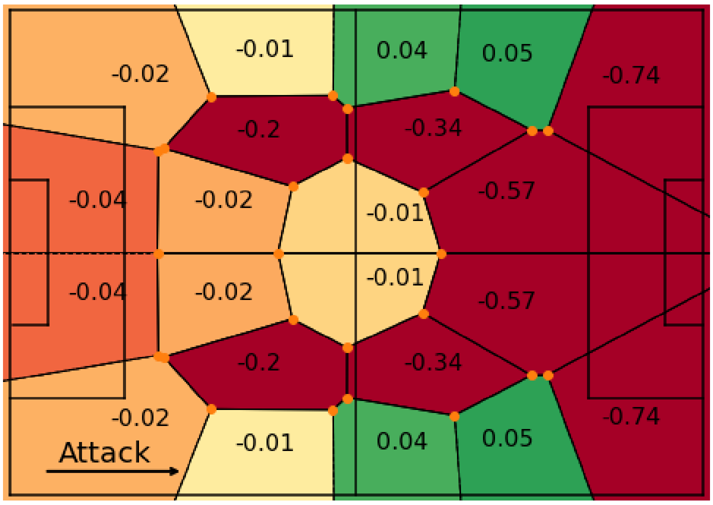


Figure 8.7: Model Prediction for Match Status Trailing for the Match Real Betis Sevilla - Real Madrid on Matchday 24 of the 2017-2018 La Liga Season. The prediction is from the perspective of the home team. On the timeline below the periods for which the match is in the away teams favor are highlighted. The color scale is anchored at the color yellow for a prediction of 0. Green represents a zone with a positive prediction, red a zone with a negative prediction. Darker shades of each color correspond with a larger absolute size of the prediction.

Real Betis was trailing for a total of 53 minutes during the matches regular time and for another 4 minutes of stoppage time. Figure 8.7 shows the untransformed model predictions per zone for this period. The prediction summed up over all zones for status *Trailing* is equal to -0.39. This means that our model rates the possession held by Real Betis while trailing as unfavorable, predicting a negative goal difference for the possession in the majority of the zones. Especially for the possession in the central zones in front of the opponents goal, our model heavily favors Real Madrid.

Drawing



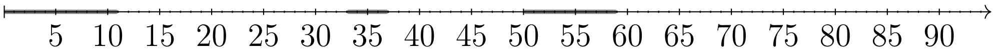


Figure 8.8: Model Prediction for Match Status Leading for the Match Real Betis Sevilla - Real Madrid on Matchday 24 of the 2017-2018 La Liga Season. The prediction is from the perspective of the home team. On the timeline below the periods for which the match is drawn are highlighted. The color scale is anchored at the color yellow for a prediction of 0. Green represents a zone with a positive prediction, red a zone with a negative prediction. Darker shades of each color correspond with a larger absolute size of the prediction.

Real Betis was drawing for a total of 25 minutes during the matches regular time. Figure 8.8 shows the untransformed model predictions per zone for this period. The prediction summed up over all zones for status *Drawing* is equal to −1.86 goals. That means that our model rates Real Madrid as the clearly dominant team for these periods. Especially in the zones close to the opponents goal, both central and towards the edges of the pidge, our model strongly favors the possession of Real Madrid.

Leading

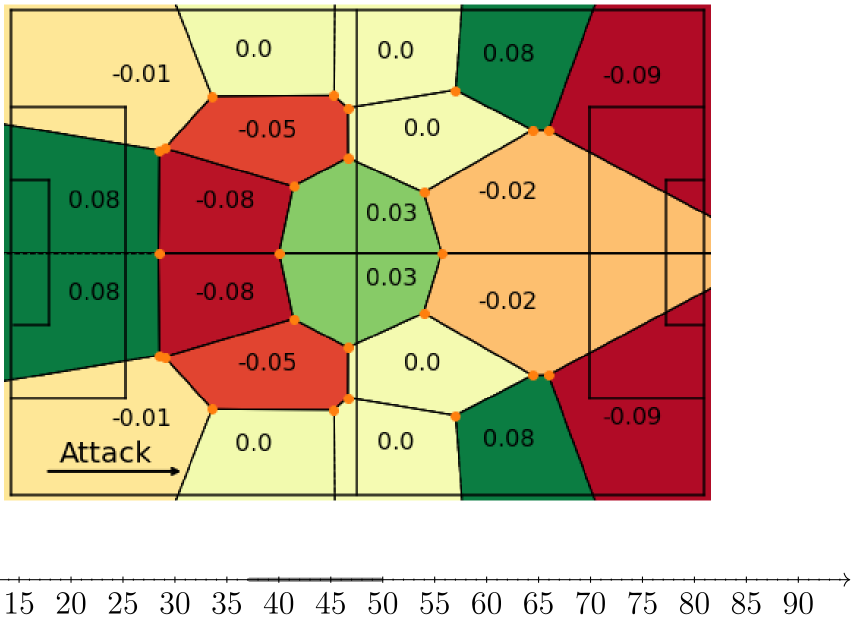


Figure 8.9: Model Prediction for Match Status Leading for the Match Real Betis Sevilla - Real Madrid on Matchday 24 of the 2017-2018 La Liga Season. The prediction is from the perspective of the home team. On the timeline below the period for which the match is in the home teams favor are highlighted. The color scale is anchored at the color yellow for a prediction of 0. Green represents a zone with a positive prediction, red a zone with a negative prediction. Darker shades of each color correspond with a larger absolute size of the prediction.

Real Betis was leading for a total of 13 minutes during the matches regular time. Figure 8.8 shows the untransformed model predictions per zone for this period. The prediction summed up over all zones for status *Leading* is equal to -0.05. That means that our model sees a slight disadvantage for Real Betis during the period in which they were leading the game. Once again, our model prefers Real Madrid's possession in front of the opponents goal. But Real's dominance in these zones is much smaller than for the other two statuses and it gets partly offset by a slight advantage for Las Palmas in the zones in front of their own goal as well as the zones on the edges of the pitch at the start of the oppenents third.

Discussion

Conclusion

This thesis project aimed at improving the existing raw possession metric by adding spatio-temporal context to it with the help of event data. By using subsets of the data we managed to examine the effect of possession controlled for *Match Status* and *Match-up Balance*.

We found that *Match Status* has a significant effect on mean possession, with trailing teams holding 52.1% on average. On the subset of passes played during *Match Status* drawing we found that with 56.1%, eventual winners of matches hold significantly more possession than losers. This indicates that *Match Status* needs to be considered while looking at the effect of possession on the outcome of football matches.

To control for *Match-up Balance*, match-day betting odds were joined indicating the balance of a given match-up. With the helps of these odds the data set was split into half and the effect of possession on match outcome was analyzed on the half containing the most even matches. This analysis showed that in even matches, without also controlling for *Match Status*, the effect of possession turns significantly negative with eventual winners only holding 47,9% of possession. Therefore, also *Match-up Balance* needs to be considered while looking at the effect of possession on the outcome of football matches.

Combining both controls by restricting the analysis to a subset containing only the passes of the even half of matches during status drawing, we find that the combined effect is positive with eventual winners holding 0.521% of possession. Existing studies on the effect of possession like Collet et al. [2] and Fernandez et al. [3] on match outcomes have only controlled for one of the two factors at a time. As controlling for each of these factors in isolation moves the effect of possession on outcome into opposite directions, this thesis provides an explanation for these contradicting results.

The need for spatial context was proven by comparing the kernel density estimates of pass coordinates between winners and losers. These density estimates revealed that winners play a higher share of their passes in the opponents half and around the centre of the pitch. Given this need for spatial context, two ways of separating the football pitch into distinct zones were introduced. One approach simply splits the pitch into rectangular zones of equal size. The other uses k-means clustering to define zones based on the resulting Voronoi cells.

Using a multiple linear regression model we combined our findings to create a new possession metric with improved predictive power. This model regresses a matches final goal difference on the possession per zone per match status. The model was fit using a 5x5 nested cross-validation setup, to allow for exploration of the two different zone types as well as the number of zones. The resulting model features 11 Voronoi zones. The results on the test sets of each of the 5 outer folds were compared to two baselines. One baseline being raw possession, the second one being match status controlled possession without spatial distinction. The raw possession baseline shows by far the worst results, with the status controlled baseline in second place. Our model outperforms both baselines on each of the outer folds.

This also holds for testing on the subset of even matches. Here, raw possession achieves a negative test score. This is caused by the fact that the aggregated effect on raw possession is positive, but the effect of raw possession on the even subset is negative. The second baseline taking into account match status improves upon the performance of the raw possession baseline, but is not able to reliable predict match outcomes either. Only our model including spatial context through zones manages to achieve a positive test score on all outer folds. The improvement in test score by adding these zone gets larger on the even subset.

Combining these results, we can conclude the following. Good teams tend to play with high raw possession, but having high raw possession does not necessarily make you a good team. For even match-ups, what matters most is the spatial distribution of your possession in those match-ups. In general, quality of possession is more important than quantity.

We proposed a way of transforming our models output into the familiar format of raw possession by minimizing the F-Statistic of the Kolmogorov-Smirnov test using a sigmoid function. This transformation allows the statistic to be more accessible for the football audience outside of the academic world. It also allows for a easier comparison between raw possession and our metric.

We showed two real-life applications of our model. In the first one we gave a breakdown of the distribution of possession for Atletico Madrid and Las Palmas in the 2017-2018 La Liga season. This analysis showed that our model is able to distinguish the higher quality of possession of Atletico Madrid from the low quality possession of Las Palmas. The second one was an in-depth analysis of a match between Real Betis Sevilla and Real Madrid. The match featured multiple lead switches and goals and ended up with Real Madrid winning 3-5. We showcased how our metric is able to differentiate between the possession profiles per *Match Status* and how it correctly identifies Real Madrid's dominance over the course of the game.

Both applications clearly show how and why our metric holds higher predictive power than raw possession. Raw possession is not able to distinguish possession per *Match Status*, which leads to biased raw possession values. Additionally, raw possession ignores the spatial distribution of possession. Our model manages to take both of these into account.

Our model setup allows a flexible use of our metric. In its aggregated and transformed form it can be used by sports journalists as a match statistic which is easy to interpret and report on. Simultaneously, coaching staff can use our metric for an in-depth analysis of a teams play style using the untransformed possession profiles per zone and status.

Limitations and Further Work

During zone definition, we decided to mirror the passes along the horizontal axis of the pitch. This mirroring allowed for increased stability of our model, by increasing the sample size per zone while simultaneously decreasing the amount of zones to fit. This mirroring imposes the assumption that football is a symmetric game along the horizontal axis and differentiation between the left and right side of the pitch can be neglected. Our available data set showed that this assumption might be overly simplistic, as there seems to be a slight skew towards the right side of the pitch, potentially caused by the higher share of right footed players. This asymmetry itself might be an interesting topic for an independent study. If this analysis was repeated on a larger sample size, the mirroring assumption could be lifted, potentially revealing interesting insights about the impact of asymmetric play on match outcomes.

Currently, zones are defined through the Voronoi cells based on the centroids of a k-means clustering. The method was chosen as it narrowly beat out the rectangular zone definition in test score during nested cross-validation. But the two methods are very close in performance and it is plausible that the performance of the model could be further improved by finding alternative methods of defining zones. Such methods could feature a higher share of zones in the important zones in front of the opponents goal or an iterative adjustment of zone borders based on the achieved test during model fitting.

Our model splits possession per match status and mean centres all possession value by the mean possession per zone per status. While this approach is a big improvement compared to ignoring match status all together, it is by no means perfect. A team trailing by multiple goals might have a different mean possession per zone than a team trailing by only one goal. Currently, our model does not differentiate between different scenarios inside each match status. Further versions of the model could incorporate this heterogeneity inside match statuses.

On the basis of a larger sample size we propose experimentation with the inclusion of additional features into the model. The model could be enhanced by adding more context to passes themselves in the form of pass accuracy or pass sub-types. An inclusion of features based on different event types might be another promising approach. For example, event type *Duel* could be incorporated into the model by adding the ratio of duels won and lost per zone.

Bibliography

[1] H.P.H. Eggels. Expected goals in soccer: explaining match results using predictive analytics. Master's thesis, Eindhoven University of Technology, 2016.

[2] Christian Collet. The possession game? a comparative analysis of ball retention and team success in european and international football, 20072010. *Journal of sports sciences*, 31, 10 2012.

[3] Javier Fernández, Luke Bornn, and Daniel Cervone. A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Machine Learning*, 110(6):1389–1427, May 2021.

[4] L. Bransen. Valuing passes in football using ball event data. Master's thesis, Erasmus University Rotterdam, 2017.

[5] Stanislaw Weglarczyk. Kernel density estimation and its application. *ITM Web of Conferences*, 23:00037, 01 2018.

[6] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[7] Leona S. Aiken, Stephen G. West, and Steven C. Pitts. *Multiple Linear Regression*, chapter 19, pages 481–507. American Cancer Society, 2003.

[8] Payam Refaeilzadeh, Lei Tang, and Huan Liu. *Cross-Validation*, pages 532–538. Springer US, Boston, MA, 2009.

[9] Richard Simon. *Resampling Strategies for Model Assessment and Selection*, pages 173–186. Springer US, Boston, MA, 2007.

[10] L. Pappalardo, Paolo Cintia, A. Rossi, Emanuele Massucco, P. Ferragina, D. Pedreschi, and F. Giannotti. A public data set of spatio-temporal match events in soccer competitions. *Scientific Data*, 6, 2019.

[11] Creative Commons. Creatice commons license. `https://creativecommons.org/licenses/by/4.0/`, 2021. [Online; accessed 16-November-2021].

[12] Football-Data.co.uk. Bookmaker odds. `https://creativecommons.org/licenses/by/4.0/`, 2021. [Online; accessed 14-July-2021].

[13] Richard Pollard. Home advantage in football: A current review of an unsolved puzzle. *The Open Sports Sciences Journal*, 1, 06 2008.

[14] Jun Han and Claudio Moraga. The influence of the sigmoid function parameters on the speed of backpropagation learning. In José Mira and Francisco Sandoval, editors, *From Natural to Artificial Neural Computation*, pages 195–201, Berlin, Heidelberg, 1995. Springer Berlin Heidelberg.

[15] Frank J. Massey. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[16] Transfermarkt.com. Overview of la liga champions 1929-2021. `https://www.transfermarkt.com/laliga/erfolge/wettbewerb/ES1`, 2021. [Online; accessed 16-November-2021].

[17] Transfermarkt.com. Overview of la liga final tables 1929-2021. `https://www.transfermarkt.com/laliga/tabelle/wettbewerb/ES1`, 2021. [Online; accessed 16-November-2021].

[18] Transfermarkt.com. Overview of las palmas coaches 1951-2021. https://www.transfermarkt.com/laliga/mitarbeiterhistorie/verein/472/personalie_id/1, 2021. [Online; accessed 16-November-2021].

Appendix A

Significance Tests

The number of passes per match are not independent and in the case of the subgroup controlled for *match status.* time spent per status varies from match to match. Therefore, assumptions of standard parametric approaches are violated and these approaches are not suited well to obtain significance levels of the mean possession per subgroup.

Instead, significance levels for the mean possession values per subgroup were obtained by applying Monte Carlo permutation tests to the data. The non-parametric Monte Carlo permutation test will allow us to obtain a asymptotically exact sample of the distribution under H_0, based on which we can obtain the estimated p-values $\hat{p}$.

Under $H_0 : P_{\text{won}} = P_{\text{lost}} = 0.5$, a random assignment of outcome labels should result in a mean possession of 0.5. $N = 10000$ permutations of the data were generated by randomly assigning outcome labels per team to each of the M matches.

For each of those permutations, the sum of passes per outcome or *match status*

$$NP_{\text{outcome}} = \sum_{m=1}^{M} NP_{\text{m,outcome}},$$

was calculated to obtain possession values

$$P_{\text{won}} = \frac{NP_{\text{won}}}{NP_{\text{won}} + NP_{\text{lost}}} \text{ and } P_{\text{lost}} = 1 - P_{\text{won}}.$$

A.1 Possession per Outcome

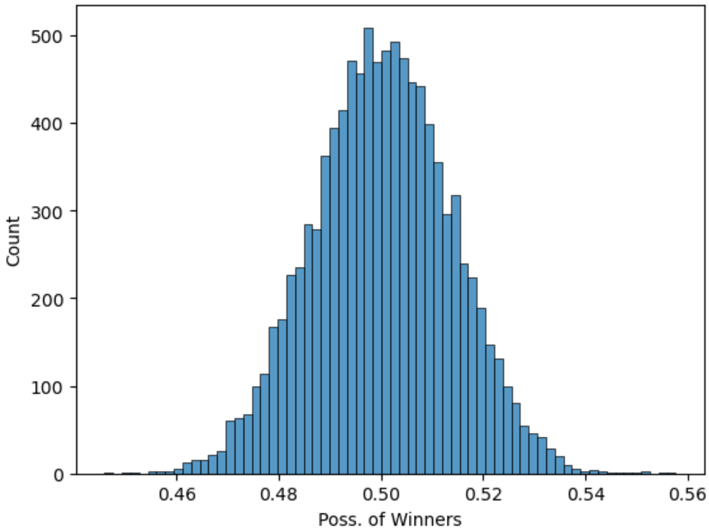


Figure A.1: Outcome: Histogram of the Mean Possession of Winners under H_0. Based on the distribution of possession values shown in this histogram, $\hat{p}$ for $H_1 : P_{\text{won}} > P_{\text{lost}} > 0.5$ was obtained.

A.2 Possession per Outcome Match Status

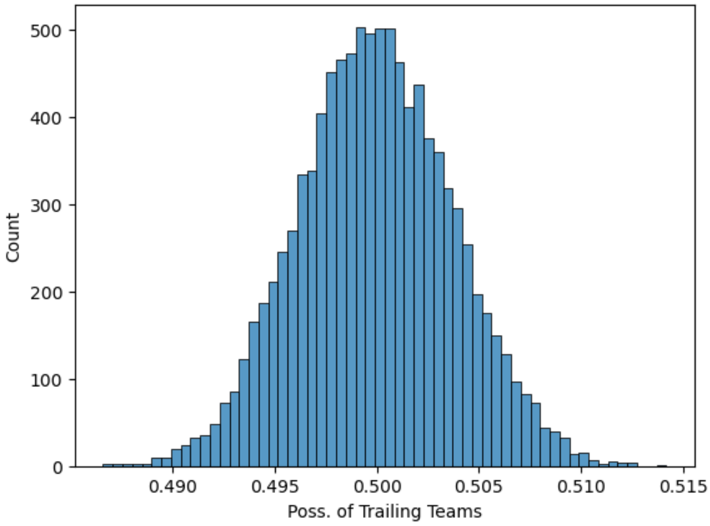


Figure A.2: Possession per Match Status: Histogram of the Mean Possession of Trailing Teams under H_0. Based on the distribution of possession values shown in this histogram, $\hat{p}$ for $H_1 : P_{\text{trailing}} > P_{\text{leading}} > 0.5$ was obtained.

A.3 Possession per Outcome - Match Status

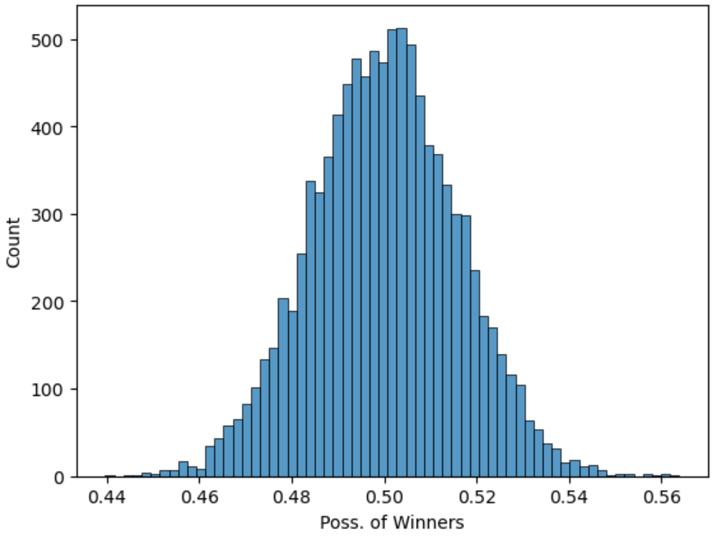


Figure A.3: Possession on Outcome for Match Status drawing: Histogram of the Mean Possession of Winners under H_0. Based on the distribution of possession values shown in this histogram, $\hat{p}$ for $H_1 : P_{\text{won}} > P_{\text{lost}} > 0.5$ was obtained.

A.4 Possession per Outcome - Match-up Balance

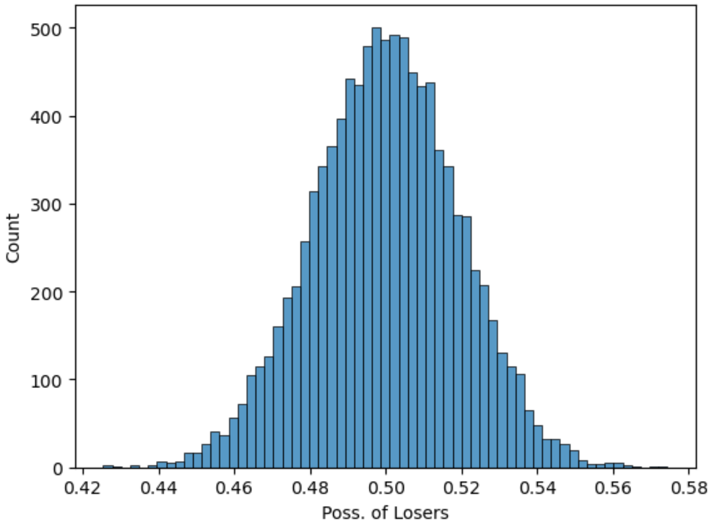


Figure A.4: Possession on Outcome for Match-up Balance even: Histogram of the Mean Possession of Winners under H_0. Based on the distribution of possession values shown in this histogram, $\hat{p}$ for $H_1 : P_{\text{lost}} > P_{\text{won}} > 0.5$ was obtained.

A.5 Possession per Outcome - Match Status and Match-up Balance.

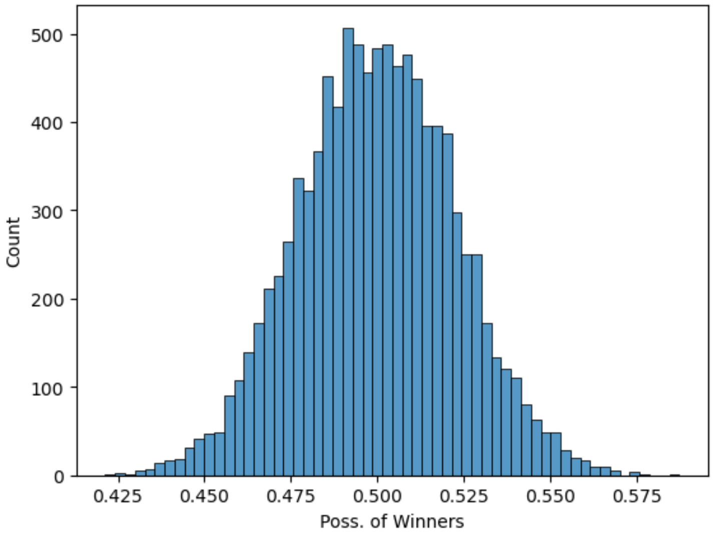


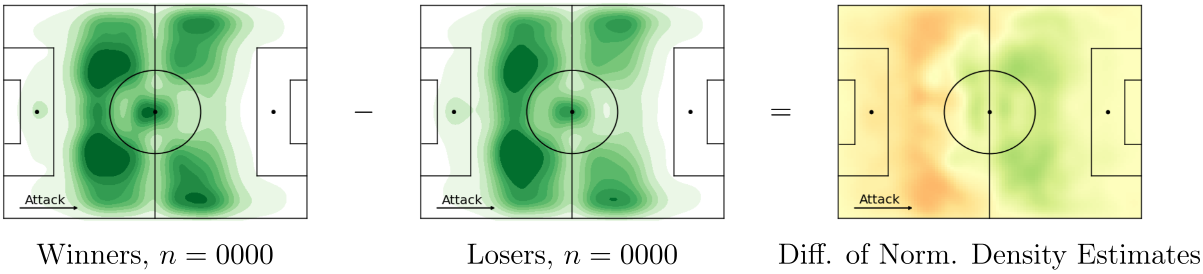
Figure A.5: Possession on Outcome for Match Status drawing and Match-up Balance even: Histogram of the Mean Possession of Winners under H_0. Based on the distribution of possession values shown in this histogram, $\hat{p}$ for $H_1 : P_{\text{won}} > P_{\text{lost}} > 0.5$ was obtained.

Appendix B

KDE of Sub-events of Event Type Pass

Event type *pass* contains multiple *sub-events*. Each of this sub-type comes with their own spatial distribution on the football pitch. This allows us to compare the distribution of winners and losers, analogously to what has been done for all passes combined in Chapter 4. In this appendix we show the distribution per *sub-event* of event type *pass* per outcome, as well as the differnce between both distributions.

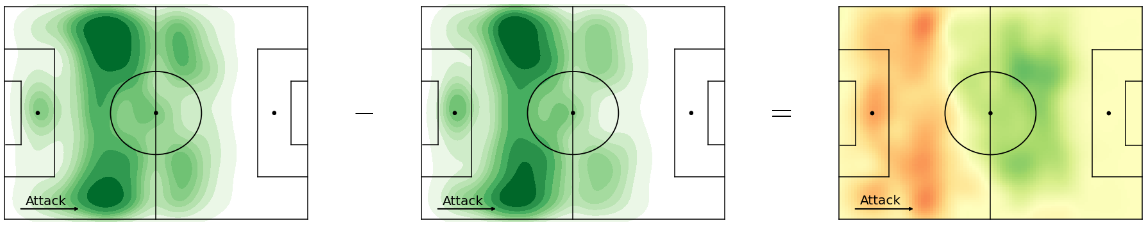
B.1 Simple Pass



Winners, $n = 0000$ Losers, $n = 0000$ Diff. of Norm. Density Estimates

Figure B.1: KDE of Pass Sub-type Simple Pass. The left figure shows the KDE of outcome Winner, the centre figure of outcome Loser and the right figure the difference between the two.

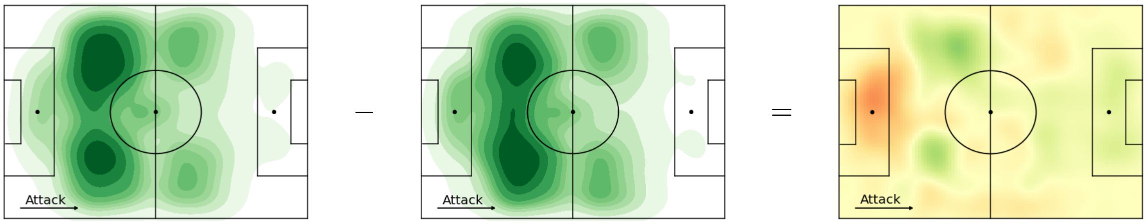
B.2 High pass



Winners, $n = 0000$ Losers, $n = 0000$ Diff. of Norm. Density Estimates

Figure B.2: KDE of Pass Sub-type High Pass. The left figure shows the KDE of outcome Winner, the centre figure of outcome Loser and the right figure the difference between the two.

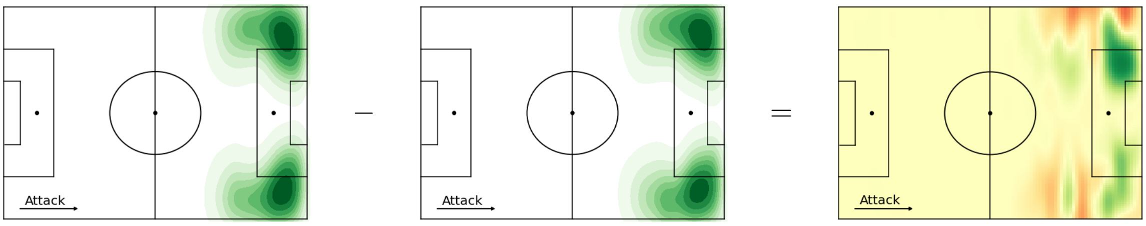
B.3 Head pass



Winners, $n = 0000$ Losers, $n = 0000$ Diff. of Norm. Density Estimates

Figure B.3: KDE of Pass Sub-type Head Pass. The left figure shows the KDE of outcome Winner, the centre figure of outcome Loser and the right figure the difference between the two.

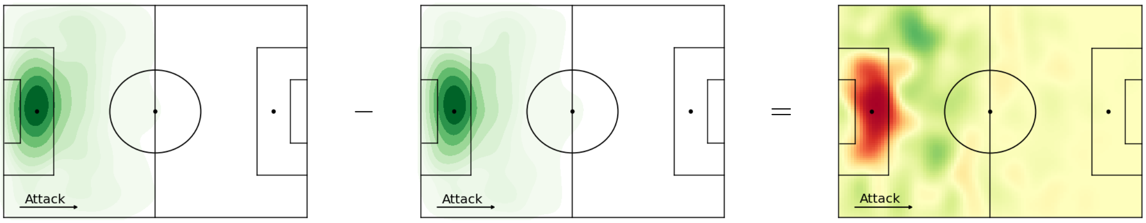
B.4 Cross



Winners, $n = 0000$ Losers, $n = 0000$ Diff. of Norm. Density Estimates

Figure B.4: KDE of Pass Sub-type Cross. The left figure shows the KDE of outcome Winner, the centre figure of outcome Loser and the right figure the difference between the two.

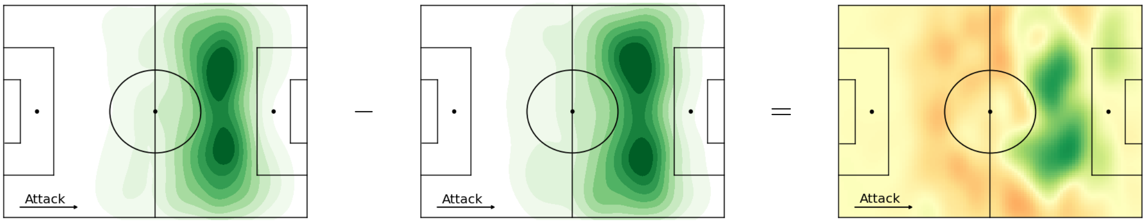
B.5 Launch



Winners, $n = 0000$ Losers, $n = 0000$ Diff. of Norm. Density Estimates

Figure B.5: KDE of Pass Sub-type Launch. The left figure shows the KDE of outcome Winner, the centre figure of outcome Loser and the right figure the difference between the two.

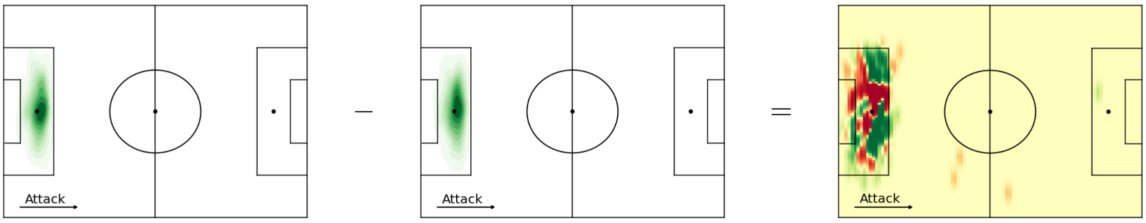
B.6 Smart Pass



Winners, $n = 0000$ Losers, $n = 0000$ Diff. of Norm. Density Estimates

Figure B.6: KDE of Pass Sub-type Smart Pass. The left figure shows the KDE of outcome Winner, the centre figure of outcome Loser and the right figure the difference between the two.

B.7 Hand Pass



Winners, $n = 0000$ Losers, $n = 0000$ Diff. of Norm. Density Estimates

Figure B.7: KDE of Pass Sub-type Hand Pass. The left figure shows the KDE of outcome Winner, the centre figure of outcome Loser and the right figure the difference between the two.

Appendix C

Nested Cross Validation Results

In this Appendix the full results of all $l \times m = 25$ Inner Folds of the nested cross-validation can be found. They are displayed separately per zone definition type *Voronoi* and *rectangular*. For zone definition type *Voronoi* k was tested on the range $[1, 2, ..., 20]$. For zone definition type *rectangular* all values in the range $[1, 2, ..., 20]$ that allowed for a separation of the pitch into squares of equal size post-mirroring were chosen, resulting in values of k of $[1, 2, 6, 8, 15, 18]$.

C.1 Full Results Voronoi Zones

Table C.1: Full Results of all 25 Inner Folds of the Nested Cross-Validation for Zone Type Voronoi

k/																									
Fold	1.1	1.2	1.3	1.4	1.5	2.1	2.2	2.3	2.4	2.5	3.1	3.2	3.3	3.4	3.5	4.1	4.2	4.3	4.4	4.5	5.1	5.2	5.3	5.4	5.5
1	0.15	0.15	0.21	0.19	0.18	0.11	0.13	0.13	0.23	0.23	0.16	0.20	0.26	0.16	0.13	0.16	0.24	0.18	0.13	0.11	0.14	0.13	0.20	0.12	0.18
2	0.15	0.16	0.22	0.21	0.18	0.13	0.11	0.16	0.26	0.23	0.20	0.18	0.30	0.15	0.13	0.18	0.20	0.22	0.17	0.10	0.16	0.16	0.15	0.15	0.20
3	0.15	0.17	0.23	0.21	0.17	0.14	0.11	0.16	0.25	0.24	0.20	0.18	0.31	0.13	0.12	0.18	0.19	0.23	0.16	0.10	0.16	0.15	0.15	0.15	0.21
4	0.16	0.14	0.24	0.23	0.16	0.14	0.11	0.18	0.26	0.25	0.22	0.19	0.30	0.13	0.14	0.16	0.21	0.22	0.16	0.13	0.19	0.14	0.16	0.15	0.21
5	0.16	0.16	0.26	0.26	0.18	0.15	0.15	0.20	0.26	0.25	0.24	0.22	0.30	0.15	0.14	0.20	0.24	0.21	0.16	0.15	0.19	0.13	0.21	0.22	0.20
6	0.18	0.18	0.24	0.26	0.18	0.14	0.10	0.18	0.27	0.25	0.21	0.22	0.30	0.14	0.16	0.19	0.23	0.24	0.14	0.13	0.17	0.13	0.21	0.19	0.19
7	0.16	0.18	0.26	0.26	0.18	0.14	0.11	0.20	0.26	0.24	0.22	0.23	0.30	0.15	0.16	0.21	0.24	0.22	0.16	0.11	0.15	0.11	0.21	0.21	0.21
8	0.15	0.14	0.25	0.25	0.22	0.14	0.14	0.20	0.27	0.24	0.23	0.20	0.30	0.15	0.14	0.21	0.23	0.23	0.16	0.11	0.16	0.08	0.21	0.21	0.21
9	0.16	0.17	0.25	0.26	0.17	0.12	0.13	0.21	0.29	0.23	0.23	0.20	0.29	0.16	0.10	0.21	0.25	0.20	0.18	0.10	0.18	0.10	0.20	0.21	0.21
10	0.18	0.15	0.26	0.25	0.18	0.14	0.11	0.19	0.29	0.23	0.23	0.20	0.29	0.15	0.13	0.22	0.24	0.23	0.17	0.09	0.18	0.09	0.18	0.21	0.21
11	0.20	0.17	0.26	0.26	0.17	0.15	0.14	0.23	0.30	0.21	0.23	0.18	0.29	0.14	0.14	0.23	0.24	0.24	0.21	0.08	0.20	0.12	0.20	0.22	0.22
12	0.19	0.15	0.25	0.26	0.15	0.13	0.11	0.21	0.28	0.22	0.23	0.18	0.29	0.13	0.13	0.24	0.23	0.24	0.20	0.08	0.19	0.13	0.18	0.22	0.21
13	0.19	0.15	0.26	0.23	0.15	0.12	0.11	0.21	0.27	0.22	0.22	0.19	0.28	0.12	0.13	0.24	0.24	0.22	0.19	0.09	0.19	0.12	0.18	0.18	0.19
14	0.23	0.16	0.25	0.24	0.14	0.11	0.12	0.19	0.27	0.21	0.23	0.18	0.27	0.12	0.13	0.24	0.24	0.21	0.16	0.09	0.19	0.13	0.17	0.18	0.22
15	0.21	0.17	0.27	0.22	0.13	0.15	0.10	0.16	0.27	0.20	0.22	0.20	0.25	0.14	0.14	0.26	0.26	0.22	0.17	0.04	0.16	0.11	0.19	0.19	0.22
16	0.20	0.19	0.27	0.23	0.15	0.16	0.10	0.17	0.25	0.21	0.22	0.21	0.26	0.13	0.12	0.25	0.26	0.22	0.18	0.04	0.16	0.08	0.19	0.19	0.21
17	0.23	0.15	0.27	0.22	0.11	0.17	0.08	0.17	0.25	0.20	0.19	0.19	0.25	0.13	0.12	0.24	0.27	0.22	0.17	0.06	0.18	0.10	0.19	0.20	0.20
18	0.19	0.13	0.27	0.21	0.10	0.13	0.08	0.18	0.22	0.21	0.20	0.16	0.25	0.14	0.11	0.23	0.26	0.18	0.18	0.03	0.18	0.14	0.16	0.20	0.21
19	0.20	0.12	0.27	0.23	0.10	0.16	0.08	0.18	0.24	0.19	0.19	0.17	0.26	0.13	0.10	0.24	0.26	0.20	0.19	0.02	0.16	0.11	0.17	0.21	0.21
20	0.19	0.10	0.26	0.21	0.12	0.12	0.07	0.16	0.24	0.21	0.22	0.15	0.26	0.15	0.11	0.21	0.24	0.18	0.19	0.04	0.15	0.12	0.15	0.18	0.22

C.2 Full Results Rectangular Zones

Table C.2: Full Results of all 25 Inner Folds of the Nested Cross-Validation for Zone Type Rectangular

k/																									
Fold	1.1	1.2	1.3	1.4	1.5	2.1	2.2	2.3	2.4	2.5	3.1	3.2	3.3	3.4	3.5	4.1	4.2	4.3	4.4	4.5	5.1	5.2	5.3	5.4	5.5
1	0.15	0.15	0.21	0.19	0.18	0.11	0.13	0.13	0.23	0.23	0.16	0.20	0.26	0.16	0.13	0.16	0.24	0.18	0.13	0.11	0.14	0.13	0.20	0.12	0.18
2	0.17	0.16	0.23	0.21	0.19	0.15	0.10	0.15	0.26	0.23	0.19	0.18	0.30	0.14	0.14	0.18	0.22	0.22	0.18	0.10	0.16	0.15	0.18	0.15	0.20
6	0.15	0.17	0.23	0.25	0.17	0.18	0.10	0.18	0.28	0.23	0.23	0.22	0.30	0.15	0.13	0.20	0.24	0.23	0.17	0.10	0.16	0.15	0.20	0.20	0.21
8	0.18	0.20	0.30	0.27	0.15	0.15	0.11	0.18	0.25	0.23	0.22	0.25	0.28	0.16	0.17	0.24	0.26	0.23	0.17	0.07	0.16	0.14	0.19	0.21	0.22
15	0.23	0.14	0.28	0.25	0.15	0.17	0.11	0.18	0.25	0.23	0.25	0.19	0.26	0.14	0.13	0.23	0.23	0.18	0.18	0.05	0.20	0.13	0.19	0.20	0.21
18	0.18	0.15	0.28	0.25	0.10	0.15	0.10	0.17	0.25	0.21	0.20	0.16	0.28	0.12	0.12	0.24	0.22	0.17	0.19	0.07	0.16	0.09	0.14	0.21	0.19