Universiteit Leiden
The Netherlands

Opleiding Informatica

The Risks and Rewards of Pressure in Football.

Guus Toussaint
s1805819

Supervisors:
Dr. A.J. Knobbe & Dr. L.A. Meerhoff

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

01/08/2019

Abstract

With the development of positional recording systems in football, new tactical analyses are possible. These analyses could be employed to quantify the risks and rewards of specific actions such as applying pressure: The risk is to leave spaces open for the attacking team, but the reward could be to recover the ball. Defensive pressure is when the team that is not in possession of the ball is actively trying to take back control of the ball. There are multiple ways of applying pressure often related to the position on the pitch where pressure is applied. In this research we refer to these different types of defensive pressure as *zones*. The aim in this thesis was to research the different types of pressure, both the expert opinion of the KNVB and the data-driven types of defensive pressure, and compare them based on their related risks and rewards. In consultation with the KNVB we devised a set of rules for the expert opinion. For the data-driven defensive pressure types we used k-means clustering. This resulted in four different pressure zones for the expert opinion. Since no data-driven number of clusters could be found we also chose four different zones for the data-driven types of pressure. We compared these types of pressure with two features *terrain gain* (the amount of terrain the attacking team moved towards the defending goal in meters) and *time to possession* (the amount of time it took to regain control of the ball in seconds). After analyzing the results we concluded that *time to possession* is not influenced by the type of pressure that is applied, both the expert opinion and data-driven types of pressure showed no significant difference in average *time to possession* between the different zones. For the feature *terrain gain*, however, it showed that the data-driven type of pressure divided the risks along a linear path, i.e. the more the defending play moves away from the defenders goal the higher the risks are. For the expert opinion type of pressure the risks resemble a U-shape. The pressure zone closest to the defending goal and the zone furthest away from the defending goal represent the zones which carry the greatest risks. The zones situated between the two carry the lowest risks.

Contents

1 Introduction **1**
- 1.1 Related work . . . 1
- 1.2 Research objectives . . . 2

2 Methodology **3**
- 2.1 Data . . . 3
- 2.2 Expert opinion . . . 3
- 2.3 Data Processing . . . 4
  - 2.3.1 Features . . . 4
  - 2.3.2 Filtering . . . 5
- 2.4 Data Mining . . . 6
  - 2.4.1 K-Means . . . 6
- 2.5 Risks and Rewards . . . 7

3 Results **9**
- 3.1 Expert opinion . . . 9
- 3.2 Events . . . 10
- 3.3 Clustering . . . 10
  - 3.3.1 K-means . . . 10
- 3.4 Risks and Rewards . . . 11

4 Conclusions and Discussion **13**

References **15**

Appendices **16**

1 Introduction

With the development of positional recording systems for football players a lot of new approaches to tactical analysis became possible, e.g. using the data to determine whether all players are in their correct positions according to the formation. For a coach it is very important to know what the risks and rewards are for certain actions, e.g. the knowledge that a tackle within the penalty box can lead to a penalty. With the newly available data it is now possible to do a quantitative analysis for such risks and rewards. A critical part in a football match is applying pressure. Applying pressure (sometimes called *pressing*) is when the team that is not in possession of the ball is actively trying to take back control of the ball. There are different ways a team can apply pressure, e.g. a team can stay close to their own goal when defending as a way of giving the opponent very little space to create a scoring opportunity. On the other hand a team can defend very high up the field with the advantage being that if they regain control of the ball they are close to the opponents goal and thus making a potential scoring opportunity much more likely. In order to make a correct decision about what type of pressure should be applied it is crucial to know what the different types of defensive pressure are and what risks and rewards are related to these types of pressure.

1.1 Related work

There is an increasing demand for tactical analysis in sports. Football teams are slowly incorporating tactical analysis in training, preparation and recruitment of new players. This development means that other football teams can't fall behind. Tactical analysis in football also called 'sports analytics' is an interplay between sports experts and data. This type of analysis makes it possible to quantify the 'gut feeling' of a coach or other expert. This in turn makes it possible to check whether there statistical proof to backup the opinion of the experts.

The current view on applying defensive pressure is that there are four different types of defensive pressure. These types are referred to as 'zones' and are related to where the defenders of the team that is applying pressure are located on the football field when pressure is applied. The zones are distributed over the length of the field and refer to 'very low', 'low', 'high' and 'very high' types of pressure. The zone closest to the defenders goal is seen as 'very low' pressure and the more the defenders move away from their goal the higher the pressure gets. The zone closest to the attackers goal is seen as 'very high' pressure.

The data gathered from the positional recording systems is spatio-temporal data. Because the position of every player and the ball is recorded at a frequency of 10 Hz this leads to a dataset which contains more than 2,4 million data points for a 90 minute football match. This abundance of data points can make it difficult to formulate the correct questions for the data. In order to do tactical analysis on a football match it is required to have a high level knowledge of football tactics, this is due to the fact that you cannot do a tactical analysis without knowing what tactics are at play. Tactical analysis also has some subjectivity, different experts have different views and therefore require a different analysis. The need of high level knowledge and the element of subjectivity makes tactical analysis in football a complex process.

As described by Memmert et al. [1] a lot of research is being done in this field to meet the increasing demand. As shown by Bialkowski et al. [2] tactical analysis can lead to a better model than just analyzing the matches with the naked eye, i.e. the model created by Bialkowski is significantly better then experts at identifying different teams. Most research however has been focused on the classification of individual events such as passes [3] or on the performance of individual players [4].

Little research is done to discover the data-driven types of defensive pressure and how they relate to the expert opinion, as described earlier this can be very interesting because it allows us to research whether the expert opinion on defensive pressure is also reflected in the spatio-temporal data. Andrienko et al. [5] has done research in identifying the different types of pressure however his research is more focused on the type of pressure an individual player is applying and not on what kind of pressure a team is applying. It is important to be able to detect and analyse what type of pressure a team is applying, because that is what determines the choices made by a coach. After determining the expert opinion and data-driven types of defensive pressure applied by a team, a comparison between the two based on the risks and rewards can be very interesting. The average time elapsed to recover the ball for a certain type of pressure can be seen as either a risk. The average distance the ball moves towards the defenders goal for a certain pressure type can be seen as a rewards. Being able to see how these risks and rewards behave for certain types of pressure is very interesting, because it improves the information a coach has available when choosing a pressure type to be applied.

1.2 Research objectives

The first aim of this thesis is to detect and categorize the expert opinion of the KNVB, henceforth this will be referred to as 'the expert opinion', and the data-driven types of defensive pressure based on the spatio-temporal data. In order to achieve this aim, a rule-set for the expert opinion on defensive pressure has to be created. For the detection and categorization of the data-driven types of defensive pressure a data mining algorithm has to be implemented. The second aim of this thesis is to compare the expert opinion and data-driven types of defensive pressure based on the risks and rewards. In order to achieve this goal, first the risks and rewards must be determined and calculated for each type of pressure. Secondly some sort of statistical comparison between the expert opinion and data-driven defensive pressure types must be conducted. This research will answer the following question: How do the expert and data-driven types of defensive pressure relate to each other based on the risks and rewards?

2 Methodology

This chapter describes the methods and data used for achieving the aims and objectives of this research.

2.1 Data

For this research the data of 106 half football matches of the dutch national team were used. This data is provided by the KNVB, in order to comply to the privacy regulations the team and player names are all anonymized. The data consists of the X and Y coordinates of all 22 players and the ball. This data is recorded at 10 Hz.

Table 1: Example of the raw spatio-temporal data obtained from a football match showing the X- and Y-coordinates for each point in time as well as some additional information.

Time stamp	**X**	**Y**	**Player Name**	**Shirt**	**Team**
1800100	-19.08	15.513	ball		
1800200	-18.016	17.048	ball		
1800300	-16.98	18.537	ball		
1800400	-15.935	20.038	ball		
1800500	-14.886	21.546	ball		
1800600	-13.851	23.035	ball		
1800700	-12.799	24.546	ball		
1800100	-47.744	-634	Jag9a4 Z9ap	1	NL001
1800200	-47.692	-502	Jag9a4 Z9ap	1	NL001
1800300	-47.636	-373	Jag9a4 Z9ap	1	NL001
1800400	-47.576	-249	Jag9a4 Z9ap	1	NL001
1800500	-47.515	-131	Jag9a4 Z9ap	1	NL001
1800600	-47.452	-19	Jag9a4 Z9ap	1	NL001
1800700	-47.388	86	Jag9a4 Z9ap	1	NL001

Table 1 shows an example of the raw data. Timestamp refers to the time of the recording. X and Y refer to the positions on the field of the subject. X is the length of the football field and the Y is the width of the football field. The center spot of the football field is the point *(0,0)*. Player Name is a unique identifier for each player and the ball. Shirt is the shirt number of the players. Team refers to the team that the player belongs to. The Shirt and Team columns are left empty for the ball, because the ball has neither a shirt number nor does it belong to any team.

2.2 Expert opinion

Because a comparison between the data-driven classification of defensive pressure and the expert opinion will be made, a concrete formulation of the expert opinion is required. This will be done in consultation with experts from the Dutch national football team. We will have a meeting with a performance analyst at the KNVB. We will formulate the current views of applying defensive pressure in such a way that it can be implemented as a rule-set, i.e. a set of rules that will define

what type of defensive pressure is applied. This rule-set will serve as the expert opinion on the different types of defensive pressure. So it is important that the rule-set is realistic (it has to be implemented) and representative (the experts have to agree on the rule-set).

2.3 Data Processing

In order to be able to detect and categorize the different types of pressure some data processing has to take place. For the processing of the data we used an existing framework created in Python called TacticsPy as described by Meerhoff et al. [6], this framework makes it possible to implement new features and events based on the raw spatio-temporal data from a football match.

The output of the pipeline is event-based which means that all the features are calculated per event. An event is a set of rules based on features, e.g. a pass which can be detected based on the distance between the ball and a player, the speed and direction of the ball.

In order to detect when a player is put under pressure by the defensive team, and thus detecting that the defensive team is applying pressure, a new event needs to be created. This event will be called a *pressure event*. A *pressure event* is a sequence of time in which a team is applying pressure. The detection of a *pressure event* will be done by applying rules on the features described in section 2.3.1.

2.3.1 Features

The features needed to detect and analyse a *pressure event* are as follows:

1. Distance to closest defender:
   This is the euclidean distance between the player with the ball and the closest player from the opposing team in meters.

2. Minimal distance to closest defender:
   This is the minimal value of the feature distance to closest defender during an *pressure event* in meters.

3. Point of pressure:
   This is the X-position (length of the football field) of the player which is put under pressure in percentage of the football field where the defenders goal is 0% and the attackers goal is 100%.

4. Centroid of the last two defenders:
   This is the X-position (length of the football field) of the centroid of the two field players (does not include the goal keeper) which are located closest to their own goal in percentage of the football field where the defenders goal is 0% and the attackers goal is 100%.

5. Terrain gain:
   This is the distance an attacking player moved towards the goal of the defending team during a pressure event in meters. If the attacking player moved away from the goal of the defending team this value is negative.

6. Time To Possession:
   This calculates the time from the start of the event until the defensive team has regained possession of the ball in seconds.

Feature *minimal distance to closest defender* was used to determine whether an event was really a pressure event. Features *point of pressure* and *centroid of the last two defenders* were used to determine what type of pressure event it is. The features *terrain gain* and *time to possession* were used to determine the risks and rewards of the pressure event.

2.3.2 Filtering

First, every instant where a player is in control of the ball is marked as a *potential pressure event*. Then, these events were filtered with a specific rule-set to eventually created a set of *pressure events* that are interesting for the tactical analysis.

The filtering process to go from the *potential pressure events* to the real *pressure events* went as follows. As described by Andrienko et al. [5] a player is put under pressure when the distance to the closest defender is less than 9 meters. So first the events where no player from the defensive team is within 9 meters of the player with the ball were removed. Then the events where there are less than 10 or more than 11 players per team were removed, this can occur due to flaws in the data. The third filter removed all the events that took longer than 20 seconds, this was done to eliminate goal kicks, corners and other non-standard situations during a football match.

Finally, some tactical filters had to be applied. We wanted to analyse the pressure events where the defensive team is applying organized pressure, i.e. the defensive team is applying pressure in a structured manner. Therefore, the sequences that occur just after a turnover are not usable. This is due to the defensive team being unorganized since they just switched from an attacking to a defensive style of play. This problem was solved by removing the sequence of *potential pressure events* that are directly after a turnover and have a positive *terrain gain* (the attacking team is moving towards the defending teams goal) instead of a negative *terrain gain*, e.g. in Figure 1 *potential pressure events* 1 trough 5 are illustrated if we apply the filter described above pressure event 3 will be removed. When a player from the attacking team passes the ball to a teammate the type of pressure the defensive team is applying does not change, however with the current list of *potential pressure events* passing the ball to a teammate results in a new *potential pressure event*. Therefore, the final filter that was applied was merging all the individual *potential pressure events* where the team in possession of the ball does not change. Because the type of pressure is defined by the first occurrence of organized pressure the features *point of pressure* and *centroid of last two defenders* were not combined when merging the *potential pressure events* but are taken from the first event in the sequence, e.g. in Figure 1 the features *point of pressure* and *centroid of the last two defenders* for the combined *pressure event* 1-2 are taken from *potential pressure event* 1.

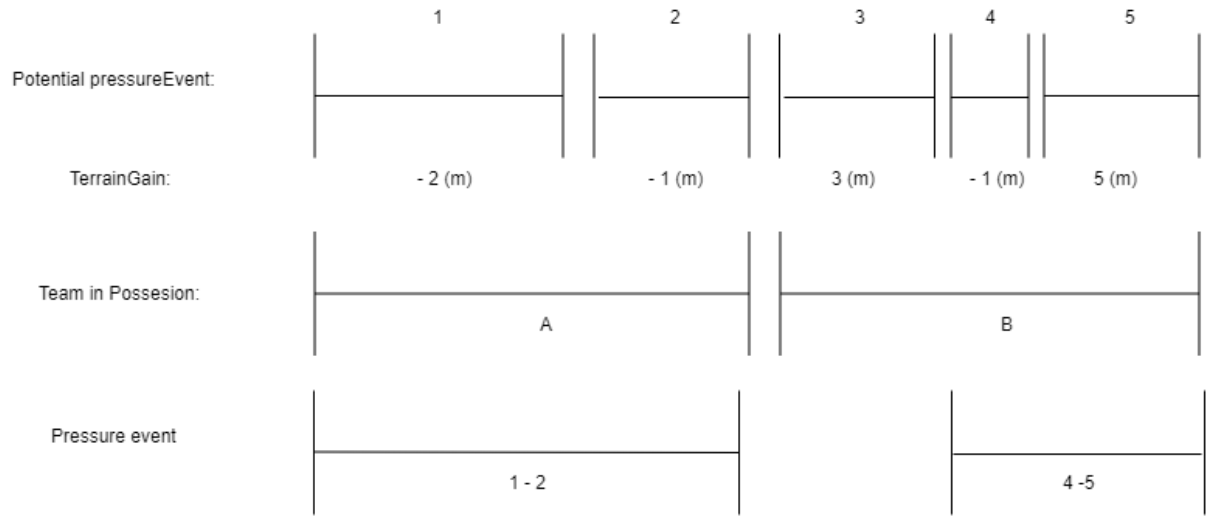


Figure 1: Illustration of the filtering process from *potential pressure events* to *pressure events*. First the feature *TerrainGain* is used to filter out some *potential pressure events* and finally the feature *team in possession* is used to combine the remaining *potential pressure events* into *pressure events*.

2.4 Data Mining

In order to achieve the aim of identifying the data-driven types of pressure as well as the expert opinion on defensive pressure the following approach was chosen. To determine the data-driven types of defensive pressure the features *point of pressure* and *centroid of the last two defenders* were used. This is a problem of finding similar cases in an unlabeled dataset, therefore, a clustering approach was used to determine the data-driven types of defensive pressure. Ergo, the outcome of the clustering algorithms, which is a set of clusters, will represent the different defensive pressure zones.

2.4.1 K-Means

The clustering algorithm that was used is k-means. K-means starts by randomly assigning k centroids, the algorithm then assigns each point to it's closest centroid based on the euclidean distance. After this is done the mean of the clusters is calculated, this serves as the 'new' centroid for the clusters. The steps of assigning the data points to centroids and calculating the 'new' centroid is repeated until there is no change in the centroid values. K-means requires you to manually define the number of clusters. If no clear number of clusters could be detected k is equal to the number of pressure zones for the expert opinion. This was done in order to evaluate the difference in risks and rewards between the expert opinion and data-driven types of pressure.

2.5 Risks and Rewards

In order to achieve the aim of identifying the risks and rewards of defensive pressure and comparing the expert opinion and data-driven types of pressure based on these risks and rewards the following approach was chosen. To determine the risks and rewards for the different types of pressure the features *terrain gain* and *time to possession* as described in section 2.3.1 were used.

In order to determine whether the expert opinion and the data-driven types of defensive pressure present significantly different risks and rewards a two-way analysis of mean variance (ANOVA) was conducted. A two-way ANOVA was chosen because it allows us to examine the effect of two factors on a dependent variable. In our case the two factors are expert opinion (zone 1 to 4) and data-driven (zone 1 to 4). The dependent variables are *time to possession* and *terrain gain*, because we have two dependent variables and the two-way ANOVA can only examine one at a time the test had to be executed for *time to possession* and *terrain gain* separately. The two-way ANOVA test is a hypothesis based test, in our case the following null-hypotheses will be tested:

1. The means of all expert opinion types of pressure are equal.
2. The means of all data-driven types of pressure are equal.
3. There is no interaction between the expert opinion and data-driven types of pressure.

If the first hypothesis is rejected, it states that the dependent variable (*time to possession* or *terrain gain*) changes significantly for each value of expert opinion pressure. The same is true for the second hypothesis only now the effects of the data-driven types of pressure will be tested. The third hypothesis will test whether the effects of expert opinion types of pressure on the dependent variable are significantly different than the effects of data-driven types of pressure.

The two-way ANOVA has six assumptions that need to be met in order to conduct the test. The first assumption states that the dependent variable (*time to possession* and *terrain gain*) should be a continuous variable. The second assumption requires the independent variables to be recorder in two or more independent groups, i.e. zone 1 through 4. The third assumption states that there are no duplicate observations, in our case *pressure events*. The fourth assumption says that their should not be any significant outliers in the dataset. The fifth assumption requires the data to be normally distributed. The sixth and final assumption requires the variances for each combination of the groups to be roughly equal.

The dependent variables are *time to possession* and *terrain gain*, these variables are measured at the continuous level, therefore the first assumption is met. The independent variables are measured in four categorical groups each, and thus the second assumption is also met. Since, every observation corresponds to a unique *pressure event* the third assumption is also met. For the fourth assumption a check will be done to detect the significant outliers, if present they will be removed from the data and thus this assumption is also met. To check the normality of the dependent variable, and thus meet the fifth assumption, a visual inspection based on the probability density function will be done. In order to meet the sixth and final assumption a Levene's test for homogeneity of variances will be conducted.

If the results of the two-way ANOVA test are significant for the interaction between the two independent variables, a Bonferroni pairwise comparison test was conducted between the expert opinion and data-driven pressure zones.

3 Results

Now that the methodology used for this research has be explained it is time to present the results obtained from following the methods described in section 2.

3.1 Expert opinion

In order to do a comparison of the expert opinion on defensive pressure types and data-driven pressure types it is necessary to know the expert opinion. The experts at the KNVB define the different types of defensive pressure as follows: There are four different zones of defensive pressure called balstart 1 through 4. Every zone refers to the zone the last two defenders are in. The zones are distributes on the football field as shown in Figure 2.

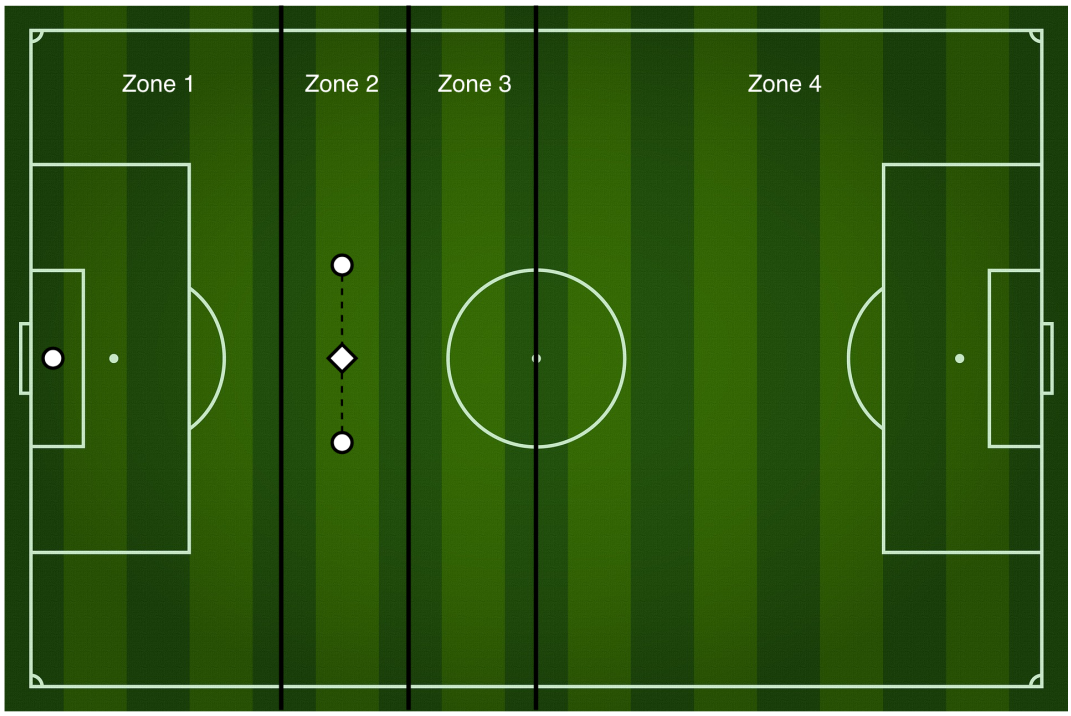


Figure 2: The Balstart zones on the football field. The dots refer to players and the diamond refers to the centroid of the last two defenders. The defending teams goal is located on the left. When a player from the defending team moves within 9 meters from the player with the ball the *centroid of the last two defenders* is calculated. The position of this centroid is the type of defensive pressure. Zone 1 represents the first 25% of the football field from the perspective of the goal of the defending team, zone 2 is between 25% and 37,5%, zone 3 is between 37.5% and 50% and zone 4 is from 50% and upwards.

3.2 Events

The output of TacticsPy contains 46,992 *potential pressure events* (*PPEs*). As described in section 2.3.2 some filters have to be applied in order to turn the *PPEs* into *pressure events*. The first filter that was applied removes the *PPEs* where no player is within 9 meters of the player with the ball. This filter removed roughly 17% (8,033 events) of the original *PPEs*, leaving 38,959. The second filter that was applied removed all the *PPEs* where a team contains less than 10 or more than 11 players. This filter removed 133 *PPEs*, leaving 38,826. The third filter removed all the *PPEs* that have a duration longer than 20 seconds. This filter removed 497 events, leaving 38,329.

Then the tactical filters were applied. The first tactical filter removed all the *PPEs* where a defending team isn't in an organized formation. This removed roughly 22% (10,193 events) of the original *PPEs*, leaving 28,136. The last filter which combines the *PPEs* into *pressure events* combined the remaining 28,136 *PPEs* into 8,461 *pressure events*.

Table 2: An example of the event table created after the filtering of the pressure events. The features *CentroidLastTwoDef* and *PointOfPressure* are shown in percentage of the length of the football field, where 0% is the defending teams goal and 100% is the attacking teams goal.

eventStart	Balstart	CentroidLastTwoDef (%)	PointOfPressure (%)	TimeToPos (s)	TerrainGain (m)	nth_pressureEvent
26.7	1.0	12.22	15.78	37.59	-0.71	2
81.0	1.0	17.54	19.73	1.09	-1.44	4
82.1	3.0	48.58	80.88	6.30	0.02	5
103.1	1.0	8.99	18.69	62.09	8.50	6
188.1	2.0	31.36	56.96	22.50	0.10	8
244.3	1.0	11.40	24.51	48.09	-1.55	10
303.4	3.0	45.63	61.67	40.50	-7.87	11
343.9	3.0	42.89	77.98	19.40	6.78	15
404.4	2.0	31.25	64.00	2.90	5.17	19
410.2	3.0	37.82	63.04	2.19	0.87	20
416.9	2.0	29.33	53.15	23.10	3.01	21

Table 2 shows an example of how the output looks. The rows are the *pressure events* and the columns represent the features as described in section 2.3.1, calculated for each event.

3.3 Clustering

Now that the the rule-set for the expert opinion has been created and all *potential pressure events* have been filtered and combined into *pressure events* it is time to determine the data-driven types of pressure. From the *pressure events* created, the features *point of pressure* and *centroid of the last two defenders* are used as a basis for the clustering.

3.3.1 K-means

In an ideal situation we would be able to identify the correct data-driven number of defensive pressure types, however, as described in the appendices this was not the case. In order to do a comparison between the data-driven and expert opinion types of defensive pressure the k-means algorithm was used to created the different zones of data-driven defensive pressure. As described in section 2 we choose $k = 4$ this represents the expert opinion on defensive pressure and thus allowed us to research the difference in the distribution of the zones compared to the expert opinion. The data-driven and expert opinion zones of defensive pressure are shown in Figure 3.

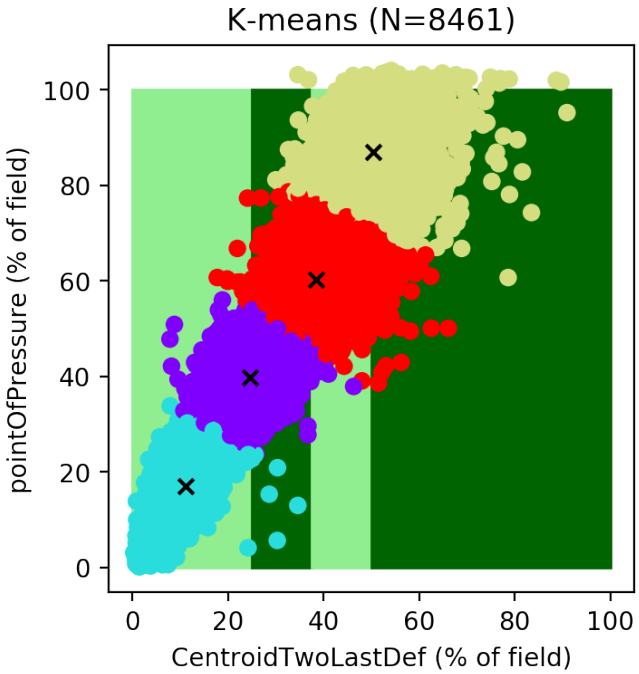


Figure 3: The four clusters with the 'x' in the middle represent the four data-driven pressure zones. The four bars in the background represent the four expert opinion pressure zones. % of the field refers to the position on the field where 0% is the goal of the defending team and 100% is the goal of the attacking team.

3.4 Risks and Rewards

Now that both the expert opinion and the data-driven types of defensive pressure are created a comparison by analyzing the difference in the risks and rewards between the two types of pressure is possible.

As described in section 2 a two-way ANOVA test was conducted to test the effects of the independent variables (expert opinion and data-driven types of pressure) on the risk and rewards. Because we are testing two different dependent variables two separate two-way ANOVA test had to be conducted. The assumptions of the two-way ANOVA were accepted, this is explained in the appendices. First we will present the results of the two-way ANOVA test with the dependent variable being *time to possession* and secondly we will present the results with the depend variable being *terrain gain*.

The results of the two-way ANOVA test with the dependent variable *time to possession* showed no significant difference between the data-driven zones of pressure, $F(3, 8449) = 2.305, p = 0.075$. Also, no significant difference was found for the expert opinion zones of pressure, $F(3, 8449) = 0.241, p = 0.868$. The test showed that there was no significant interaction between expert opinion and data-driven, $F(5, 8449) = 0.516, p = 0.765$. Thus resulting in the acceptance of all three null-hypothesis, as described in section 2. Proving that the different types of pressure whether it is expert opinion or data-driven provide no significant difference in the rewards related to *time to possession*.

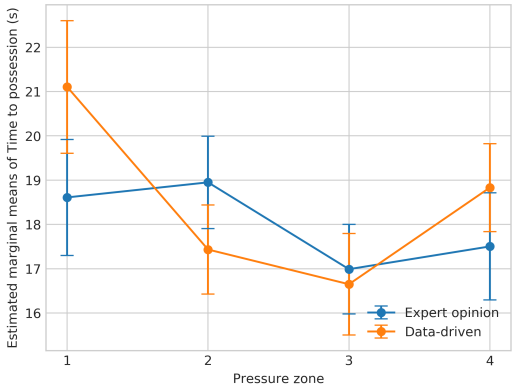


Figure 4: The estimated marginal means of the dependent variable time to possession, shown for each value of the independent variables expert opinion and data-driven.

The results of the test with the dependent variable being *terrain gain* showed a significant difference between the data-driven zones of pressure, $F(3, 8449) = 189.684, p = < 0.01$. A significant difference was also detected between the expert opinion zones of pressure, $F = (3, 8449) = 0.87.390, p = < 0.01$. And finally the interaction between the expert opinion and data-driven zones also showed a significant difference, $F(5, 8449) = 48.129, p = < 0.01$. Thus resulting in the rejection of all three null-hypothesis. In order to identify which expert opinion and data-driven zones differ significantly from each other a Bonferroni pairwise comparison test was conducted. The pairwise comparison resulted in a significant difference between all groups, the only comparison that does not have $p = < 0.01$ is data-driven zone 2 and expert opinion zone 2 this comparison has $p = 0.01$.

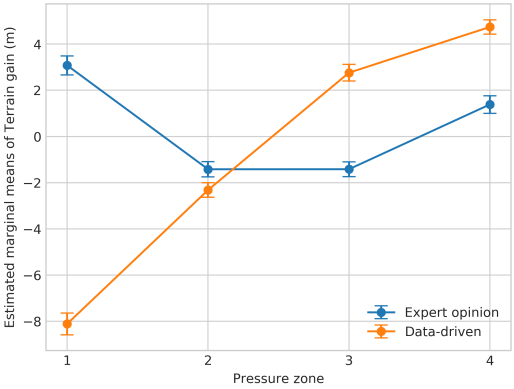


Figure 5: The estimated marginal means of the dependent variable terrain gain, shown for each value of the independent variables expert opinion and data-driven. Illustrating the main effects of the two-way ANOVA test.

Figure 5 described the estimated marginal means for *terrain gain* per zone of expert opinion pressure and data-driven pressure. Another interesting result that can be noted is that for the expert opinion zones of pressure the difference in the dependent variable *terrain gain* between zones 2 and 3 is insignificant, $p = 1.00$.

4 Conclusions and Discussion

The question posed in this thesis was: How do the expert and data-driven types of pressure relate to each other based on the risks and rewards? We can conclude that for the data-driven pressure zones the risks related to *terrain gain* follow a linear path. For the expert opinion on defensive pressure the risks related to *terrain gain* follow a path that resembles a U-shape, but still divide the risks in a significant way. The pressure zone closest to the defending goal and the zone furthest away from the defending goal represent the zones which carry the greatest risks. The zones situated between the two having the lowest risks, i.e. the average *terrain gain* is negative which means that the ball is moving away from the defending teams goal. For the data-driven types of defensive pressure the pressure zone closest to the defenders goal yield the lowest risks. The more the data-driven pressure zone moves away from the defending goal the higher the risks are, for pressure type 2 the average distance the ball moves away from the defending goal is significantly smaller than for pressure type 1. For pressure types 3 and 4 it holds that the ball on average moves towards the defending goal and thus posing a greater risk. If a coach want to 'play safe' a low pressure type is advised. This conclusion is in accordance with the tactics employed today. Many coaches choose to defend close to their own goal when no risks can be taken, e.g. when a team is ahead of its opponent by only one goal and the match is almost over. This tactic is sometimes referred to as 'park the bus'. The rewards related to the average time until the defending team has regained possession of the ball are not significantly different for either type of pressure. In practice this means that a coach cannot significantly alter the average time to regain control by changing the teams defensive tactic. However these findings are limited to the risks and rewards related to *terrain gain* and *time to possession*. No concrete conclusions can be drawn about how the expert opinion and data-driven types of pressure relate to each other based on the total risks and rewards. Furthermore, the robustness of the statistical comparison can be put into question since no statistical evidence is provided for the fifth and sixth assumptions of the two-way ANOVA-test. These assumptions are regarded as met solely due to the large sample size of the population.

The first aim of this thesis was to identify both the expert opinion and the data-driven types of defensive pressure. In order to achieve this aim a clustering approach on the features *centroid of the last two defenders* and *point of pressure* was chosen. This let us to the conclusion that based on these two features no clear data-driven number of clusters could be determined. There might be a way to find the data-driven number of cluster by using different features as a basis for clustering, however, this was not further explored in this research.

There are some limitations to the research conducted in this thesis. The data used for this research only consists of data from the KNVB, this may cause the data to be over fitted on the playing style of the dutch national team since half of the data (each match also includes an opponent) is from the dutch national team. Furthermore, the risks and rewards are limited to *time to possession* and *terrain gain*. As one can imagine there are more risks and rewards that play a role when applying defensive pressure. Finally the fifth and sixth assumptions of the two-way ANOVA test should be regarded as met based on some sort of statistical test and not solely on the fact that there is a large sample size.

Based on the known limitations of this research future work should include the following: The clustering in this research only used the features *point of pressure* and *centroid of the last two defenders*, however, in further research it would be interesting to investigate more and other features as a basis for the clustering of the different types of defensive pressure, e.g. the spread of the defensive team. Another component that could be researched more is adding more risks and rewards to the analysis. Where a reward could be creating a scoring opportunity directly after regaining control of the ball. Conducting more research with different types of risks and rewards will give a better understanding of the overall risks and rewards for pressure. It would also be interesting to apply the risks and rewards on certain teams, e.g. only using the data for specific teams it would be possible to analyse what works best against certain teams. This can be of great interest for a coach since it creates the possibility do a quantitative analysis on the performance of the opponent.

For a coach it is important to have a clear distinction in the risks and rewards for certain actions, like applying defensive pressure, this can lead to a better preparation and a better knowledge of what risks are taken and what rewards are granted when applying a certain type of pressure. This research shows that using the data-driven types of defensive pressure the risks and rewards regarding *terrain gain* are distributed in a linear way, and thus resulting in a intuitive distribution. The further you move away from your own goal, the greater the risk is that your taking.

Acknowledgements

This research was conducted in collaboration with Tim Janssen from the koninklijke nederlandse voetbalbond (KNVB).

References

[1] D. Memmert, K. A. Lemmink, and J. Sampaio, "Current approaches to tactical performance analyses in soccer using position data," *Sports Medicine*, vol. 47, 06 2016.

[2] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews, "Identifying team style in soccer using formations learned from spatiotemporal tracking data," *IEEE International Conference on Data Mining Workshops, ICDMW*, vol. 2015, pp. 9–14, 01 2015.

[3] M. Horton, J. Gudmundsson, S. Chawla, and J. Estephan, "Classification of passes in football matches using spatiotemporal data," *ACM Transactions on Spatial Algorithms and Systems*, vol. 3, 07 2014.

[4] E. Nsolo, P. Lambrix, and N. Carlsson, *Player Valuation in European Football*, pp. 42–54. 04 2019.

[5] G. Andrienko, N. Andrienko, G. Budziak, T. Landesberger, and H. Weber, "Exploring pressure in football," pp. 1–3, 05 2018.

[6] L. A. Meerhoff, A.-W. de Leeuw, F. R. Goes, and A. Knobbe, "Mining soccer data: Subgroup discovery of tactics from spatio-temporal data,"

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[8] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

Appendices

K-means

To determine the correct number of clusters the elbow method was used. When using the elbow method we calculated the sum of squared errors (SSE) (sometimes called the Sum of Squared distances) as described in section 2.3.2 of Scikit learn [7] for each number of clusters. The SSE refers to the squared distance for each node to it's closest cluster. When we plot the SSE for each k we can see a elbow shaped curve, with this curve we can determine what the potentially correct number of clusters is. This is done by identifying the value of k where there is a clear 'elbow', i.e. a point in the graph where the slope of the line after this point is less steep than before this point. If the SSE line results in a curve that does not show a clear 'elbow' it indicates that the data used is not optimal for clustering. The k-means algorithm was run for k 2 to 8. The threshold of 8 was chosen based on the assumption that if there are more than 8 types of defensive pressure it would be hard to spot the different types of pressure with the naked eye. Thus not being useful for a coach during a real football match. The elbow curve for these values of k is shown in Figure 6.

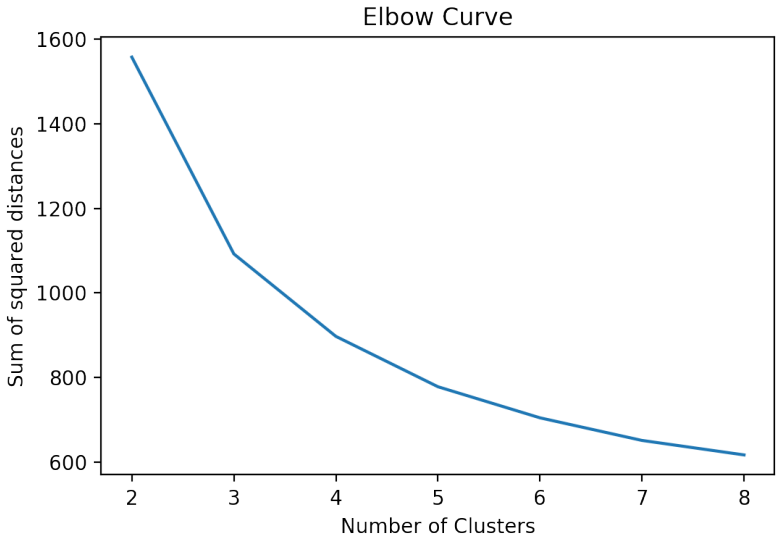


Figure 6: The elbow curve showing the sum of squared errors for each k. The curve shows no clear 'elbow' and therefor a optimal value of k cannot be chosen.

Affinity Propagation

The second clustering algorithm we used to detect the data-driven types of defensive pressure was affinity propagation. Affinity Propagation as described by Frey et al. [8] is a clustering algorithm that exchanges messages between all data points until high quality clusters are created. The messages contain the willingness of that point to be the other points exemplar. An exemplar is the data point that represents the other data points, this can be seen as the cluster centroid. The points that have the highest willingness to be an exemplar are chosen as the clusters. This contributes to the research aim of finding the data-driven types of defensive pressure because it does not require us to specify the number of clusters. Therefore, it might be able to give us insights in what number of clusters (i.e. defensive pressure zones) is ideal for the data that we would not have found otherwise. The results from the affinity propagation algorithm are shown in Figure 7, as can be seen the algorithm results in a distribution with 1910 separate clusters, thus showing 1910 different types of defensive pressure. As described earlier it must still be usable in the real world and therefore the clusters generated by the affinity propagation algorithm are discarded.

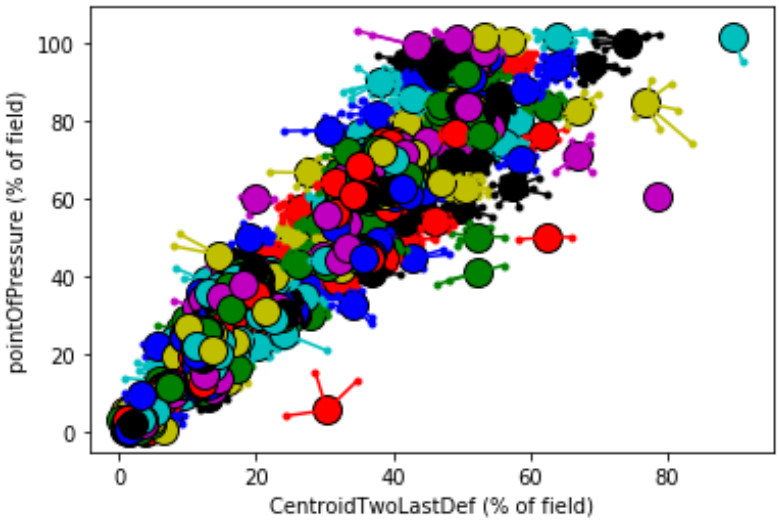
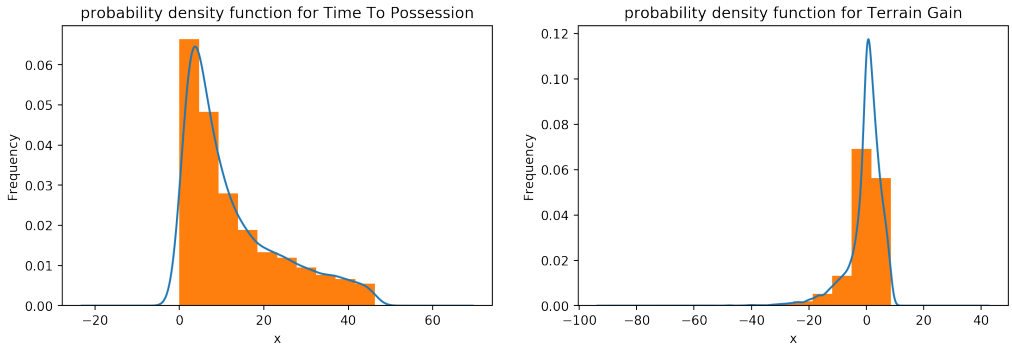


Figure 7: The results of the affinity propagation algorithm, which shows the 1910 different defensive pressure zones.

Two-way ANOVA assumptions

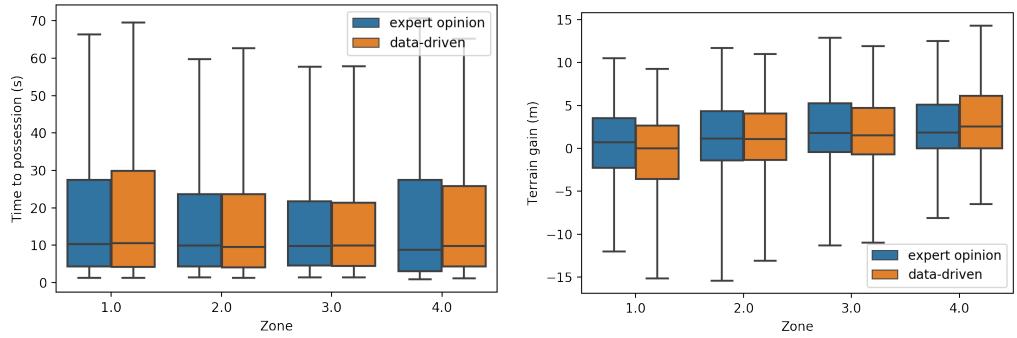
In order to conduct the two-way ANOVA the dependent variables *time to possession* and *terrain gain* must be roughly normally distributed. Figure 8(a) shows that the data is slightly positively skewed, however, since our sample size is greater than 8000 this assumption is regarded as met for the dependent variable *time to possession*. Figure 8(b) shows that the data follows a roughly normally distributed pattern, hence the assumption is also met for the dependent variable *terrain gain*.



(a) Time to possession (b) Terrain gain

Figure 8: Probability density functions for the 95 percentile of the features *time to possession* and *terrain gain*.

The final assumption that needs to be checked is the homogeneity of variances. The Levene's test results was significant, meaning that the null-hypothesis that the error variance of the dependent variables across all groups is equal was rejected for both dependent variables. However, due to the large sample size we can assume that the homogeneity of variances is still accepted. Figure 9 shows that the deviations of the dependent variables for each group are roughly equal.



(a) Time to possession (b) Terrain gain

Figure 9: Box plots of dependent variables *time to possession* and *terrain gain* distributed over the different zones for each type of defensive pressure.