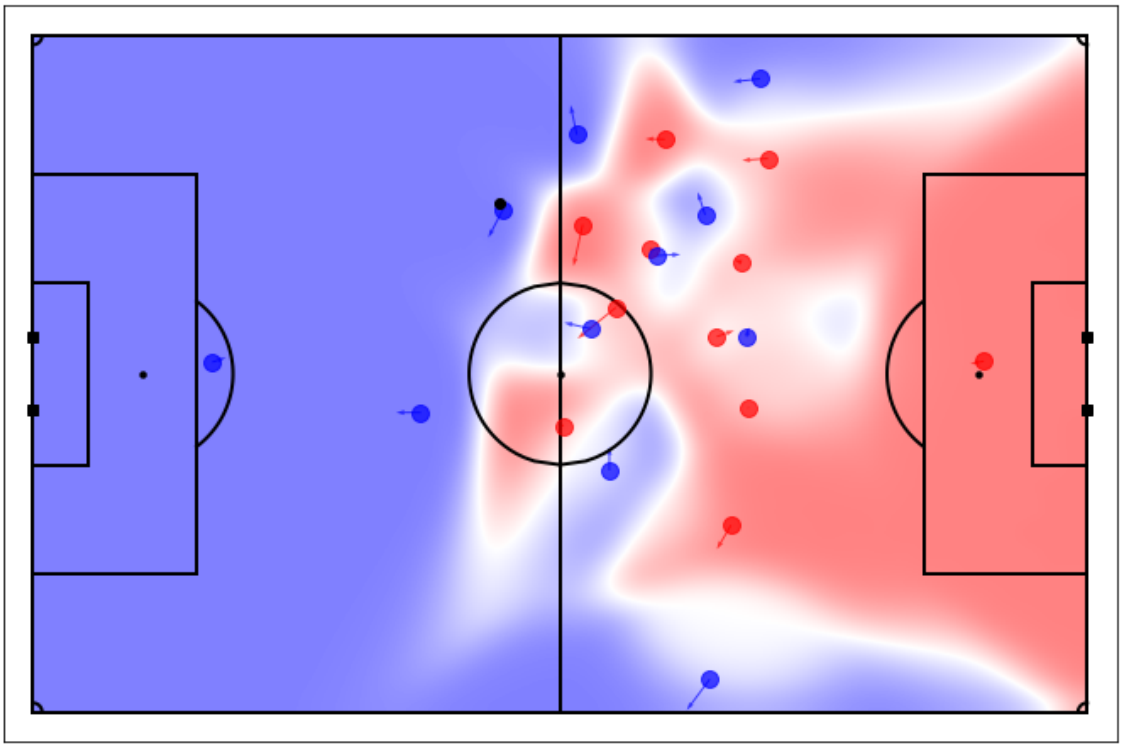
PITCH CONTROL METRICS TO IMPROVE THE PREDICTIONS OF MOMENTS LEADING TO GOALS IN FOOTBALL

SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF MASTER OF SCIENCE

RAMON DOP
10253343

MASTER INFORMATION STUDIES
DATA SCIENCE
FACULTY OF SCIENCE
UNIVERSITY OF AMSTERDAM

07-07-2022



	1st Examiner
Title, Name	Dr. Arnoud Visser
Affiliation	University of Amsterdam
Email	a.visser@uva.nl

Pitch control metrics to improve the predictions of moments leading to goals in football

Implementation of pitch control to look beyond on-ball actions

Ramon Dop
University of Amsterdam
ramon.dop@student.uva.nl

Daily Mentor: Barend Verkerk
Head of Data Science at AZ Alkmaar
b.verkerk@az.nl

Supervisor: Dr. Arnoud Visser
University of Amsterdam
a.visser@uva.nl

ABSTRACT

Methods for evaluating on-ball players' actions in football, such as passing and shooting, are far ahead of those to evaluate off-ball player actions such as blocking the line of passes or controlling space. This thesis suggests the professional Dutch football club AZ Alkmaar a way to improve its assessment of the off-ball qualities of football players. It presents the club a way of improving their baseline model by implementing pitch control metrics. Tracking data is combined with event data (e.g. passes, shots) to provide contextual information. Upon this combined data, a physics-based pitch control model is created. Pitch control is the probability a team will control the ball at location x, assuming the ball were to be passed to location x. Multiple pitch control metrics are generated from this model. The effectiveness of these new metrics is evaluated by implementing them into already existing prediction models and assessing the change in performance of the models. These models predict which events are occurring shortly before a goal is scored. The results show that adding pitch control metrics increase the performance of these prediction models by 6-10%.

KEYWORDS

pitch-control, predictions, tracking, XGBoost, off-ball, physics

1 INTRODUCTION

This thesis suggests a method to improve the assessment of off-ball player actions in football data analysis. It does so by using a pitch control model to create multiple new metrics. These metrics are then used to improve the predictions on football events leading to goals. Football (or soccer) is a dynamic and complicated game of sports with countless variables playing at a time. But in its very essence, it can be broken down into a player being on or off the ball. Since football has 22 players on the pitch, one player will perform on-ball actions at a given time while the other 21 perform off-ball actions. Despite this 1:21 ratio of on-ball vs. off-ball players, metrics quantifying the former are plentiful and high level [11], while for the latter, the options are scarce and low level. Features such as xG (expected goals), xA (expected assists) [6] [1], and PPDA [1], already exist for offensive actions. Measurements of on-ball performance are much more explored as this type of action is strictly related to the change in position or ownership of the ball in play, and thus requires less contextual information.

Off-ball performances are more complex as they relate to the position of other players and the ball in play. Some metrics exist that aim to assess the defensive skills of players, such as interceptions, tackles, and clearances, but they are also related to on-ball actions. Moreover, these are simplistic statistics. More often than not, defenders of better teams (and thus most often better defenders) will have lower scores in such metrics as their team will have the majority of ball possession. One off-ball metric commonly seen in football is kilometers run. Still, this metric has similar issues to the previously mentioned ones, where better players don't necessarily run more.

One advanced metric for measuring the off-ball qualities of football players is *Pitch Control*. Pitch control is the probability a team will control the ball at location x, assuming the ball were to be passed to location x from its current location. For each part of the pitch, this model computes the probabilities for both teams to control the ball at a given time. Liverpool FC's William Spearman introduced this model in 2017 [10]. This thesis will implement the Pitch Control model in AZ Alkmaar's data environment and then evaluate whether this metric does a significantly adequate job at predicting which situations lead to goals being scored. It does so by replicating and implementing this new metric into already existing goal prediction models and assessing the new performance of the models. If this metric seems of valuable use, it will then be used for further data scientific analysis (such as valuing player actions in player recruitment) and strategic analysis by staff members. Thus the thesis proposes to help AZ Alkmaar in assessing the off-ball qualities of football players by implementing a physics-based pitch control model which will quantify the amount of space each team controls throughout a match. This leads to the following research question:

1.1 Research Question

"To what extent can the prediction of actions leading to goals be improved by the implementation of a physics-based pitch control model?"

1.2 Sub Questions

To be able to answer this research question, the following supporting sub-questions are formulated:

- "What data is needed to create a pitch control model?"
- "What are the currently used features for predicting actions leading to goals in AZ Alkmaar's models?"
- "What is the impact of implementing pitch control metrics on the performance of said models?"

[1] https://totalfootballanalysis.com/data-analysis/data-analysis-ppda-its-definition-advantages-and-disadvantages

2 RELATED WORK

To put the pitch control model into the context of the state-of-the-art models and their history, this section will start with relevant work on the off-ball quality assessment in football in the past, followed by the suggested implementation of such a pitch control model.

Some models already assess the qualities of players without necessarily looking at their on-ball performance. Some early models from the start of this century targeted the off-ball qualities of players by focussing on the players' physical capabilities. Rösch et al composed a series of tests to assess the physicality of football players [7]. One can determine the impact of a certain player on the performance of the team by comparing the results a team produces with and without said player. A plus-minus model, which usually measures a player's impact on the game in team sports like ice hockey or basketball, was adopted in 2017 for usage in football [4]. In 2018 a similar approach was followed when a study evaluated the importance of football players for their team based on a weighted plus-minus metric [8]. The EA Sports Players Performance Index (Actim Index) is the official player rating system of the English Premier League. It awards points for good stats as well as for positive team results [5]. AZ Alkmaar has a model running that aims to predict the probability of a goal happening within the next 15 seconds. This model is reporting an AUC of ~0.833 and an RMSE of ~0.127, which will be used as the baseline that the outcomes of this research will be compared to.

In 2018, Liverpool FC's William Spearman introduced a pitch control model to obtain on- and off-ball metrics [10]. Recently, several scientific papers have explained the relevance of a pitch control model [2] [3]. In *Beyond Expected Goals*, Spearman describes the exact workings of a physics-based pitch control model [10].

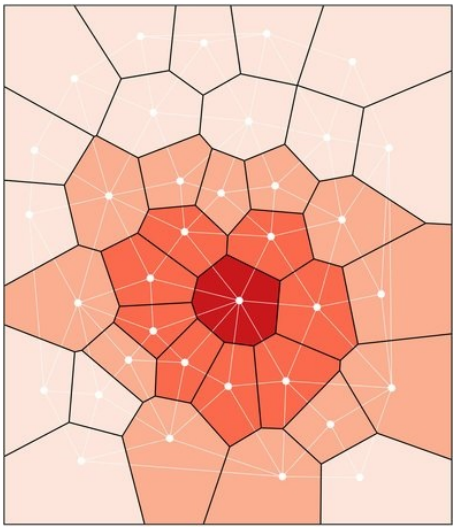


Figure 1: Voronoi diagram

A physics-based pitch control model is created to analyze the amount of pitch control per team. Through the principles of a Voronoi diagram, each player is assigned a part of the pitch it controls for each frame. The simplest model would approximate this by only looking at the player closest to the ball, just as the Voronoi diagram shows in figure 1. But distance is not the sole factor in determining how long a player takes to arrive at a certain point. Current velocity and orientation also significantly influence the amount of time it takes a player to arrive at a location.

$$r_{p-new} = r_p + v_p * t_{react}$$
$$ttc = d(r_b - r_{p-new})/v_p \quad (1)$$

In equation 1 the time to control (ttc) is calculated by first updating the location of the player r_{p-new}. It does so by taking into consideration the player's reaction time. It is assumed the player has a reaction time (t_{react}) of 0.7 seconds; thus, he/she will continue to run at his original velocity (v_p) and direction for that time. This updated location is then used in equation 1. Once the time to control is computed for every player, the probability of control for each player can be calculated. The above shows that computing the probability of control is a task that requires taking assumptions. Spearman has concluded these assumptions to be the following [10]:

- The speed of the ball is constant at 15 m/s (no acceleration and not dependent on grass or air resistance).
- The trajectory of the ball is linear.
- The velocity of the players is constant at 5 m/s.
- Players have a reaction time of 0.7 seconds.
- Players continue moving at their current velocity for 0.7 seconds and then run to the target position at full speed.
- Players take the shortest path to the target position.

However, like physics, little is guaranteed in sports like football. Thus a model should seldom assume that a certain pitch location is controlled for 100% by one team. Similar to how (seemingly) easy chances will not always be converted into goals, pitch is also not always as guaranteed due to factors like players' awareness, direction, errors, and data inaccuracies. For long distances away from the ball, players from both teams will also have the time to adjust their position and will both be able to challenge for the pass. A logistic function (preferred by Spearman over a normal distribution because of its heavier tail[10]) is used to implement some uncertainty in each player's probability of controlling the ball at a certain location.

$$P_j(t, \overrightarrow{r}, T|s) = \frac{1}{1 + e^{-\pi \frac{\sigma * (T - ttc)}{\sqrt{3}}}} \quad (2)$$

Where $P_j(t, \overrightarrow{r}, T|s)$ represents the probability that player j at time t can reach location r within some time T, given that player j is on onside with parameter s. Here σ is one of the parameters of the pitch control model. It is the standard deviation (0.45) of the cumulative distribution function of the time it takes a player to control. It can be thought of as the uncertainty in time it takes a player to intercept the ball, as there is uncertainty in the player controlling the ball based on the time he/she has before the ball arrives. This layer of uncertainty will change the model's output from a binary to a continuous probability. Once these extra layers

are incorporated into the model, the output will look similar to figure 2.

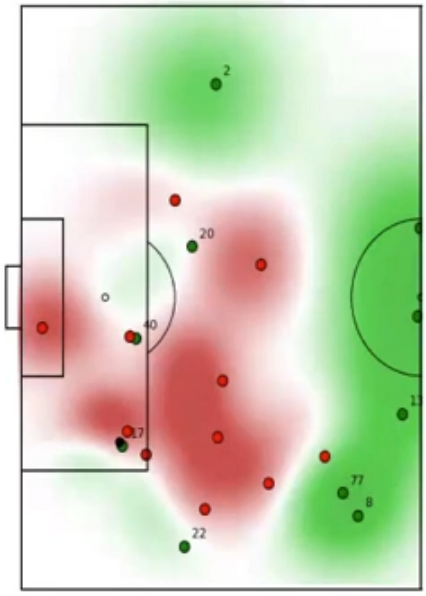


Figure 2: Example of pitch control output

The probability of pitch control per player can then be aggregated over all players of the attacking team (a), and the total value of pitch control for the team controlling the ball has been computed. The inverse of this value is the amount of pitch control for the defensive team.

$$PC_{att} = \sum_{\forall j \in a} P(t, \overrightarrow{r}, T|s) \tag{3}$$

$$PC_{def} = 1 - PC_{att}$$

To create such a pitch control model, the location of the players is needed. This information on the location of the players is obtained in the form of **tracking data** from cameras.

3 METHODOLOGY

In this section the used methods and implementation details will be explained. Firstly the used datasets are examined by providing their origin, volume and flaws. Secondly the section shows how and when both datasets are combined and what the result of the two sets leads to. Lastly, the setup for the created features from the pitch control model is provided. This study is done in a practical manner in the form of an internship. The internship took place at AZ Alkmaar, a professional Dutch football club. The practical part of the thesis consists of three parts: preprocessing the relevant datasets, combining the datasets, and creating the model.

3.1 Data

Since the pitch control model is used to predict which actions lead to goals, knowing the players' location is not enough; more contextual information, such as when goals are scored, is needed. This context comes from the **event data**. This data contains the time and location of events such as passes, tackles, and shots (and their outcome). Tracking data delivers the core information to the model, and event data provides the crucial data to put the tracking data into its context. These two different datasets are then retrieved and combined.

The data used in support of this research is scrutinized on their origin, volume, and potential flaws. This thesis uses data from two sources; Tracking data from ChyronHego and event (meta)data from StatsBomb. This section will provide the understanding, preparing, and analyzing of the two data sources to the point where it is ready to be used to evaluate the off-ball actions of players.

3.1.1 Tracking data. The tracking data is retrieved through multiple cameras placed in each stadium of the highest Dutch football division (Eredivisie). The data is the 5th generation of TRACAB [2], a distributed camera system. These cameras track the x and y coordinates of the 22 players and the ball at 25 frames per second. Since matches usually last for 95 minutes, each player will have around 142,500 ($25 \times 60 \times 95$) rows of information per match. Excluding any exceptions (such as red cards or injuries), each match will contain 22 players and a ball, thus meaning around 3,277,500 rows (142500×23) per match. The columns of the tracking data contain frame numbers, jersey numbers, and x & y coordinates.

Through the location of the player and ball per frame, the velocity and orientation can be computed. However, when analyzing the speed, some anomalies appear as sometimes the speed will exceed any reasonable amount. The unsmoothed data points that exceed the maximum velocity (these are most likely position errors from the cameras) are estimated with a Savitzky-Golay filter[3].

3.1.2 Event data. The event data is bought from StatsBomb [4], a football analytics and data visualization company. The event data is entered manually by StatsBomb and contains all possible actions in a football match (e.g., goals, passes, interceptions). Through the API of StatsBomb, every Eredivisie match of the 2021/2022 season is included. This gives 312 matches (18 teams each play 17 home matches, plus 6 playoff matches) of data. Depending on the match, each match will contain about 4000 events, with each row also containing 52 columns of additional information such as event type, outcome, and player number.

3.2 Combining the relevant datasets

As proposed by section 2, the datasets of tracking and event data need to be linked per match in order to build a pitch control model. As football data is a replication of reality, visualising the data and comparing it with the actual match can help to understand the flaws of the data. Through combining the tracking and event data, the entire match can be replicated virtually as is shown in figure 3. This figure shows the location of the event at the purple cross, as

[2] https://tracab.com/products/tracab-technologies/
[3] https://www.mathworks.com/help/signal/ref/sgolayfilt.html
[4] https://statsbomb.com/data/

well as the type of event, along with the location of the 22 players and the ball in moving picture.

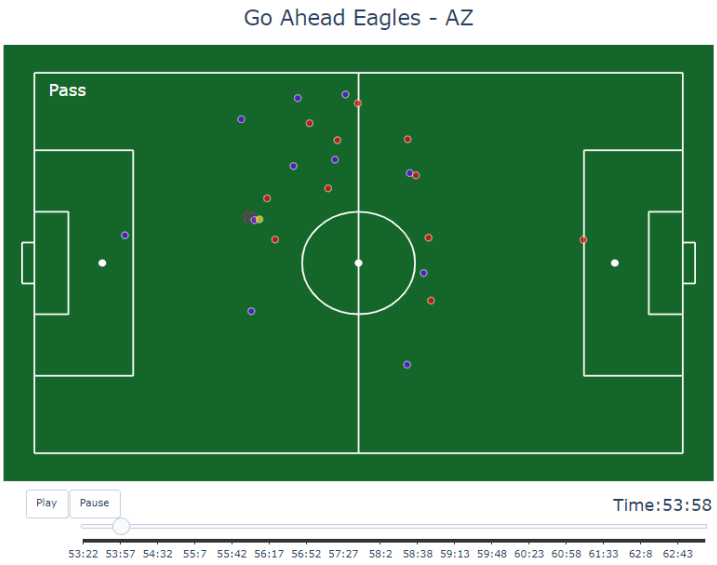


Figure 3: visualization of tracking data combined with event data

Visualising several matches shows that the location and timing of the events in StatsBomb data are, in some cases, considerably off from reality. This can be explained by realising that, while the tracking data is automatically generated by cameras, the event data is entered manually by StatsBomb employees. Even though this data is cross-checked by multiple people before it is retrieved by AZ, it is impossible to expect that this data will be correctly entered at a 0.04 second scale (25 FPS) and to have the correct location on a pitch of 105 by 68 meters. Another reason for this inconsistency in the event data seems to be that some events are happening live (goal kick) while the broadcast is still showing highlights of a previous event (shot at goal). The StatsBomb employee is then unable to see the events happening. Resulting in the missed events being likely to be placed wrongfully with a certain timestamp.

However, since the tracking data comes from cameras, this data will have near perfect accuracy. Tracking data can thus be used to verify the accuracy of the event data. Since each event has a timestamp, a player performing the event, and a location of the event, the tracking data can give an indication on how accurate the event has been handled by StatsBomb. Two things were done to circumvent this unreliability in the event data. Firstly the timing of the events were given some lenience. Rather than matching the event with the exact frame at which it occurred according to StatsBomb, the model looks 12 frames back and ahead to find the smallest distance between the location of the **event** and the location of the **ball**. By doing so the event is placed at a one second scale (25 frames) rather than a 0.04 scale (one frame). StatsBomb is now given a one-second lenience in the preciseness of their timestamps. Some events however do not require the ball to be near, such as 'error' or 'dribbled past'. For this reason the model also looks 12 frames back and forth (resulting in 25 total) to find the smallest distance between the location of the **event** and the **player**.

For the model to be credible, confirmation of successfully matching both dataframes is needed. Once for both distances (ball-event & player-event) the smallest is taken out of the 25 frames, events are then flagged as reliable or unreliable. An event is flagged unreliable if both previously mentioned distances are larger than 2.5 meters. Tracking data is recorded at 25 FPS and the algorithm is looking 12 frames back and forth, this means it is looking ~0.5 seconds back and front. 2.5 meters is taken since, as mentioned in section 2, the players' speed is assumed to be constant at 5 m/s, meaning half a second of lenience can create 2.5 meter of inaccuracy ($5 \times 0.5 = 2.5$).

If more than 10% of the events of a match are unreliable, the match is deemed not reliable enough to feed the pitch control model with its data and is thus skipped. Upon combining both datasets and computing the reliability of each, 113 matches out of the 312 matches of the 2021/22 Eredivisie season were deemed as reliable enough to use for pitch control. This gives ~250,300 relevant events. Once both sets are combined, it is known which frames lead up to goals being scored. A dummy variable *leading-to-goal* is created. Actions (of the scoring team) that happened within 15 seconds prior to a goal will be tagged as leading to goal. The value of 15 here is chosen to stay in line with the currently existing baseline model at AZ.

3.3 Creating the model

Once all relevant data has been combined and is deemed reliable enough it will be used to generate new 'Pitch Control' data. However, not all tracking data will be analysed by the pitch control model; only those rows of the tracking data at which an event occurred. This is for a combination of two reasons. Since there are about 142,500 rows of tracking data per match and the pitch control model analyses one frame at a time, it would be highly unfeasible to analyse all frames of a significant number of matches. Secondly the pitch control values are used to predict actions leading to goals. For this reason only the frames that are directly linked to actions, or events, are used for the pitch control analysis.

On top of this pitch control analysis, different types of metrics are built. Since there are various types of goals that come with their own type of build-up. The build-up of a goal from a counter-attack will not have any high numbers for pitch control, while a slowly built up attack will have very high pitch control values. The following metrics have been created in consultation with AZ Alkmaar's technical staff:

- Relevant Pitch Control
- Increase in Relevant Pitch Control
- Pitch Control behind the defensive line
- Pitch Control in the half-spaces
- Pitch Control between the lines
- Pitch Control around the ball

3.3.1 Relevant Pitch Control. Pitch control is a good indication of how much a team is dominating the pitch at given moment. However, it does not take into account that some parts of the pitch are inherently more valuable than others. If a team is in control of the ball, the pitch control around its own goal is in most cases obsolete. To take this into consideration the probability of pitch

control per team is multiplied by the expected threat (xT) value per grid. Expected threat 'values locations based on not just the immediate shooting threat, but the potential to induce danger later in the possession sequence' [9]. It is a way to quantify the amount of threat a team is exerting on the other team based on the location on the pitch. Figure 4 visualises the importance of each cell when the pitch is divided into 16 by 12 grids. The values in such a layer represent the probability that a team will score within the next five actions. This metric is the core metric of the thesis and is hypothesised to have the highest impact on the predictions of actions leading to goals.

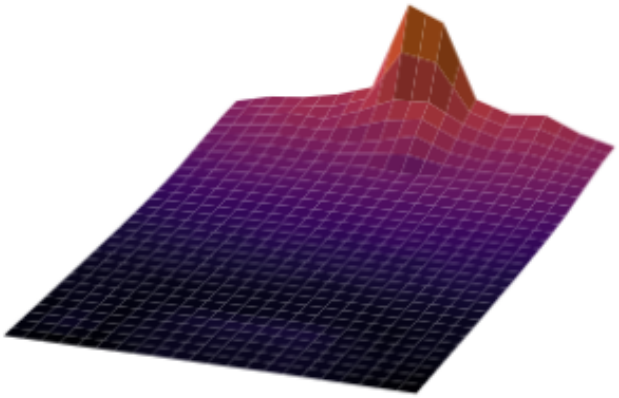


Figure 4: visualization of an expected threat layer, for a team attacking from bottom-left to top-right

3.3.2 Increase in Relevant Pitch Control. The previous metric is deemed to be able to recognize the standard way of scoring by building up build pitch control. However, for goals that came from fast breakaways or counters the metric is not expected to perform well. One way to help the model recognize such dangerous situations is by creating another feature that looks at the increment of relevant pitch control per event. A rapid influx of relevant pitch control could be a good indicator that a dangerous attack has been started.

3.3.3 Pitch Control behind the defensive line. Another intuitive way to be able to foresee possible counter-goals is by looking at the pitch control behind the defensive line. Specifically, this metric only aggregates the pitch control in the grids between the defender closest to the goal line (usually the goalkeeper) and the second closest defender.

3.3.4 Pitch Control in the half-spaces. Upon inspection, the previous metric seemed to be highly influenced by many situations during slow build-up where the players near the sidelines were given space by the defenders as that space is deemed less significant. To circumvent that, this metric ignores the wide areas of the pitch and only aggregates the pitch control between the keeper and the last (deepest) defender. By ignoring the wide areas of the pitch it only aggregates the pitch control in the center area and so-called half-spaces of the pitch.

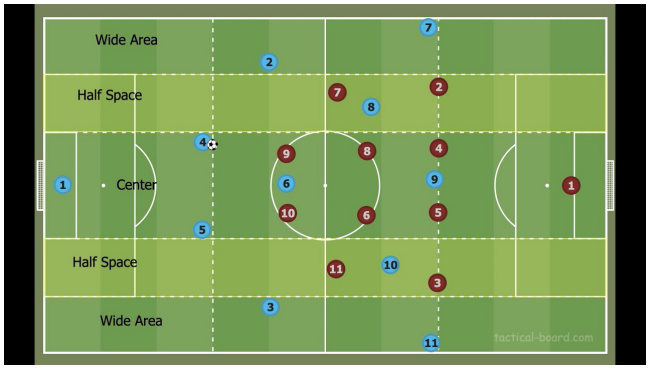


Figure 5: visualization of naming of different parts of a football pitch

3.3.5 Pitch Control between the lines. The next metric that is theorized to help the model predict dangerous counter-attacks is the amount of pitch control 'between the lines'. 'Between the lines' is a common term in football that is associated with space between the line of midfielders and the line of defenders[5]. In this thesis 'between the lines' is simplified as the space between the fourth most forward player (usually the attacking midfielder) and the deepest defender (closest to the goal line). While the fourth most forward player may sound arbitrary, this is chosen as 4-3-3 is a commonly seen formation for teams in the Eredivisie. According to this formation, the fourth most forward player will then be the most forward midfielder.

3.3.6 Pitch Control around the ball. The last metric that is thought to have impact on the amount of threat a team is asserting is the lack of pressure the opposing team has on the ball. This can be quantified by the amount of pitch control the attacking team has around the ball. This metric computes the amount of pitch control of the ball-owning team in a radius of ~11 meters around the ball.

4 EXPERIMENTAL SETUP

Similar to how the result of the merge of tracking data and event data needed to be visualized, should the output of the pitch control model also be visualized. This is shown in figure 6.

This visualization also directly shows the relevance of the implementation of offside in the model. The blue player on the right half of the pitch is given zero pitch as he/she is offside. Before looking at the potential improvement of introducing the pitch control model, the current models should be looked at first.

Currently, one baseline model and one more optimized model are already created by AZ, both solely using data retrievable from event data. The basic model uses *distance to goal* (distance from the event to the goal of the opposing team) and *angle to goal* (angle between the event and the goal of the opposing team) as variables to predict whether an event will lead up to a goal or not. The second model uses the two mentioned variables as well as *under pressure* (the player on the ball is under pressure from a player of the opposing team), *is counter* (the event is part of a counter-attack), and the current *velocity of the ball*. With the previously introduced pitch

[5]https://www.soccercoachweekly.net/other/between-the-lines/

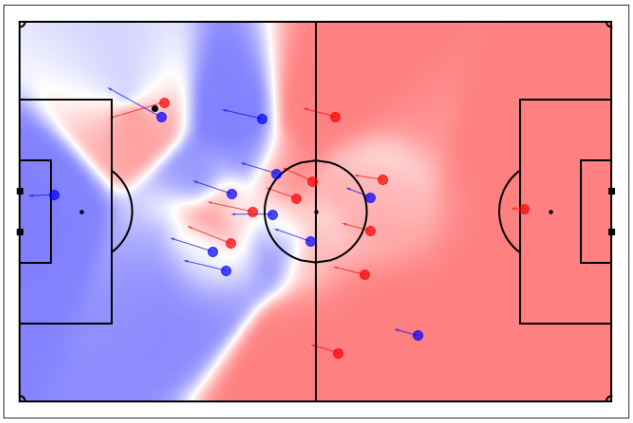


Figure 6: visualization of the output of the Pitch Control model

Table 1: Overview of variables used per model
PC stands for Pitch Control

	Model 1	Model 2	Model 3	Model 4
Distance to Goal	✓	✓		
Angle to Goal	✓	✓		✓
Is Under Pressure		✓		✓
Is Counter		✓		✓
Velocity		✓		✓
Relevant PC			✓	✓
Increase in Relevant PC			✓	✓
PC Behind Defensive Line			✓	✓
PC in half-spaces				
PC Between Lines			✓	✓
PC Around Ball			✓	✓

control metrics, two extra models have been created. One model solely uses the new pitch control metrics, and the last model makes use of all introduced metrics.

While creating the last two models two factors have been taken into consideration. The feature importance of each metric should be higher than zero. If a newly introduced metric doesn't have (significant) predicting power then it should not be used, as models using fewer variables are preferred. Secondly, the variance inflation factor (vif) should be below five for each variable. As a consequence of the latter, the metric aggregating the amount of pitch control behind the defensive line in the half-spaces and center of the pitch is taken out. The metric appeared to be too similar to the metric aggregating the amount of pitch control behind the defensive line and the feature importance of the first was lower. For similar reasons *distance to goal* was not used in the last model as it correlates too heavily with the idea of pitch control (higher values when closer to the opposing goal). This is also shown in table 1.

Each model is an XGBoost Regressor. Here regressors are preferred over classifiers as they fit better with the nature of sport, since there is always a factor of randomness. This way the results of the model also have better interpretability. Rather than having a goal scoring opportunity be labeled as 1 or 0, it is preferred to have it as 0.6 or 0.4 to understand and deal with the output of the model better.

Table 2: Hyperparemeters of the models

Hyperparameter	Value
objective	*reg:logistic*
learning_rate	0.01
colsample_bytree	0.6
gamma	2
alpha	2
max_depth	8
scale_pos_weight	10
n_estimators	500

The hyperparameters of the XGBoost regressors are set as shown in table 2. Scale_pos_weight helps with the highly imbalanced nature of the dataset. The ratio of leading to goal vs. not leading to goal is 108 (2,300 vs. 248,000). The rounded square root of this ratio is then taken (10). N_estimators is the number of trees (iterations) the XGBoost algorithm will use to train. This is determined by the learning curve, as shown in figure 7. This figure shows that the algorithm stops showing significant improvements at around 500 iterations.

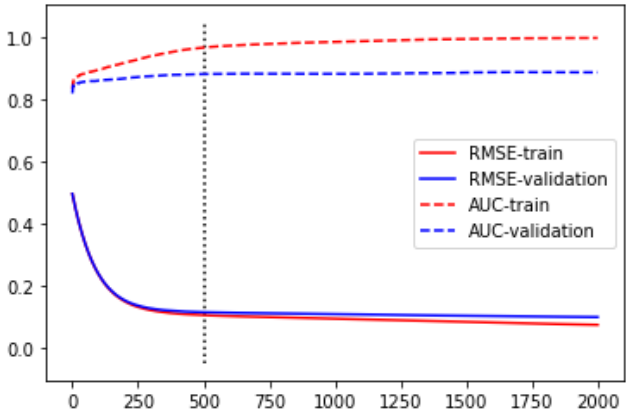


Figure 7: Learning curve

5 RESULTS

The four models, each with their own set of features, as shown in table 1 are used to predict with actions lead to goals or not. As the output of the prediction models is a continuous variable (between 0 and 1) the metrics used to measure the performance enhancement of the models are Area Under the Curve or AUC and Root Mean Squared Error or RMSE. The values of these are shown in table 3.

Table 3 shows that model 4 performs the best in both the AUC (higher value is preferred) and RMSE (lower value is preferred). Model 1 performs the worst, while model 3 is slightly outperformed by model 2 in regards to AUC. This shows that just pitch control metrics aren't enough information and that contextual information

Table 3: Performances of each model

	AUC	RMSE
Model 1; distance & angle	0.786	0.135
Model 2; without PC	0.833	0.127
Model 3; only PC	0.825	0.119
Model 4; all	0.882	0.114

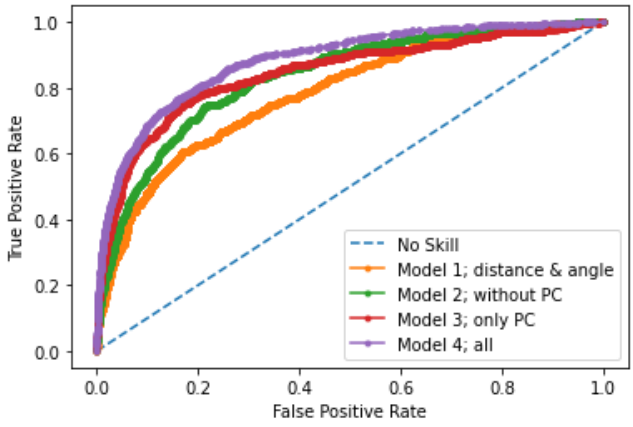


Figure 8: ROC-curve of each model

along with tracking data gives the best results. The ROC-curves in figure 8 tell a similar story. Model 4 performs the best and model 1 the worst. While model 2 and 3 have similar performances. Overall model 2 and model 4 should be compared to each other as these are the two optimised models before and after the implementation of pitch controls respectively. In this regard, the implementation of pitch control metrics result in an increase of 5.91% in AUC as well as a decrease of 9.98% in RMSE. Furthermore the ROC-curves of both models show that the FPR (False Positive Rate) of the latter model is lower or similar than that of the former model at every TPR (True Positive Rate).

Figure 9 shows the feature importance of the best performing model (model 4). This shows that all of the pitch control features used in the model have a significant prediction power on actions leading to goals. The metric that was introduced as the core metric of this thesis, relevant pitch control, has a higher feature importance for the model (~0.10) than the other pitch control metrics (<0.07).

6 DISCUSSION

There have been no previous researches into the improved predictions on actions leading to goals to compare to, outside of AZ Alkmaar's own existing model. The results show that this metric is significantly adequate enough in predicting actions leading to goals. Therefore it can be used in the assessment of football players.

The limitations of this research mainly lie in the assumptions. Some of the assumptions introduced by Spearman's pitch control model can be changed in further research. The speed of each player is set at 5m/s and the reaction time at 0.7s. These values can be replaced by values on a per player basis, in the future.

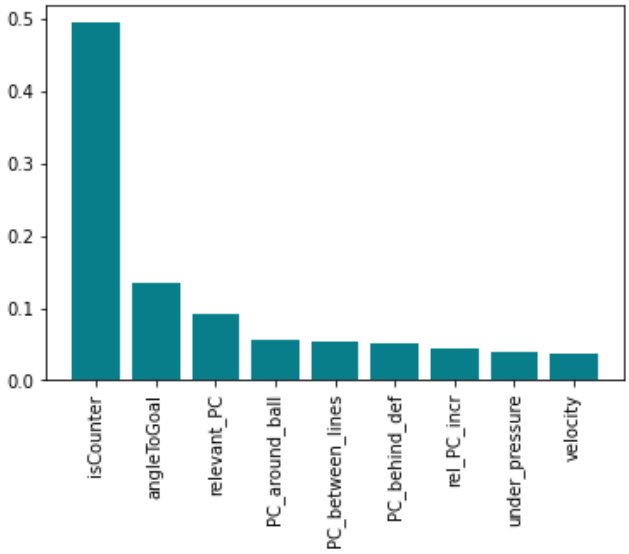


Figure 9: Feature importance of model 4

Further research should also look into improving the cleaning of the event data. As right now, only 36% of the matches are analysed. Another source of data that could be added is the previous and future seasons and/or different football competitions. Although figure 7 has shown that the improvements of the model is marginal after 500 iterations of training. The 90% ruling of reliability (required to use a match for pitch control, as described in section 3.2) could also be tweaked. Specifically when this number is increased when combined with additional data or better cleaning of event data, this could lead to better results.

Other ways to improve the model are by considering more factors when creating the model. Examples of such factors are the curve of the ball, and the difference in drag between grass and air can be implemented. Acceleration could also be taken into account in computing the time to control (*ttc*). This, again, can be individualised based on already available data of the players.

Another parameter that could be tweaked to the model is the maximum amount of seconds an event can happen for it to be tagged as *leading to goal*. In this research it is set at 15 seconds. Reducing this number should lead to higher performances of the prediction model. However, this might not be desirable in this context, as it will become biased towards the most dangerous attacking situations rather than the build up of attacks. It will also reduce the number of events tagged as *leading to goal* and thus imbalance the two classes even further than the current 1:108 ratio.

7 CONCLUSION

The thesis aims to help with the assessment of off-ball qualities of football players. While there are many high-level metrics for analysing on-ball performances, off-ball analysis is harder as it requires more contextual information. The proposed form of off-ball analysis by this thesis is pitch control, the probability for both teams of controlling each part of the pitch. The data needed to do such pitch control analysis is tracking data, combined with

contextual information coming from event data. Multiple pitch control metrics are created from this model. To measure if these metrics are truly predictive for measuring football qualities they are used to predict actions leading to goals. AZ Alkmaar already has an existing model predicting these actions. This model uses the features *Distance to Goal, Angle to Goal, Is Under Pressure, Is Counter,* and *Velocity of the ball.*

The final prediction model combines these existing features with the metrics gained from the pitch control model. Any features that cause too much correlation between the variables are removed. This final model performs better than the currently used model by AZ. The AUC is increased by almost 6% and the RMSE decreases with close to 10% when using the new model. Improving the predictions of actions leading to goal allows for better valuation of player actions in general. As this thesis has shown to help AZ Alkmaar at assessing the off-ball qualities of football players by implementing a pitch control model, this feature can now be used for this purpose. The metrics could be improved further by enhancing the cleaning process of the event data.

REFERENCES

[1] G. Anzer and P. Bauer. 2021. A Goal Scoring Probability Model for Shots Based on Synchronized. *Frontiers in Sports and Active Living* 3 (2021). https://doi.org/10.3389/fspor.2021.624475

[2] U. Brefeld, J. Lasek, and S. Mair. 2018. Probabilistic movement models and zones of control. *Machine Learning* 108, 1 (2018), 127–147. https://doi.org/10.1007/s10994-018-5725-1

[3] J. Fernández and L. Bornn. 2018. Wide Open Spaces: A statistical technique for measuring space creation in professional soccer. *MIT Sloan Sports Analytics Conference Boston, MA* (2018).

[4] T. Kharrat, J. Peña, and I. Mchale. 2017. Plus-minus player ratings for soccer. *European Journal of Operational Research* (2017), 726–736.

[5] I.G. Mchale, Phil Scarf, and David Folker. 2012. On the Development of a Soccer Player Performance Rating System for the English Premier League. *INFORMS Journal on Applied Analytics* 42, 4 (2012), 339–351. https://doi.org/10.2307/23254864

[6] A. Rathke. 2017. An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise* 12, 2 (2017), 514–529. https://doi.org/10.14198/jhse.2017.12.Proc2.05

[7] D. Rösch, R. Hodgson, and L.T.. Peterson. 2000. Assessment and Evaluation of Football Performance. *The American Journal of Sports Medicine* 43, 28 (2000), 29–39. https://doi.org/10.1177/28.suppl_5.s-29

[8] R. S. Schultze and C.-M Wellbrock. 2018. A weighted plus/minus metric for individual soccer player performance. *Journal of Sports Analytics* (2018), 121–131.

[9] K. Singh. 2019. Introducing Expected Threat (xT). Retrieved June 15, 2022 from https://karun.in/blog/expected-threat.html

[10] W. Spearman. 2018. Beyond Expected Goals. *MIT Sloan Sports Analytics Conference Boston, MA* (2018).

[11] M. Van der Werf. 2020. An Overview of Advanced Metrics in Football Analysis. Retrieved June 30, 2022 from https://medium.com/@max_vander_werf/an-overview-of-advanced-metrics-in-football-analysis-4e75fd82bef8