UPPSALA
UNIVERSITET

UPTEC IT 22022
Examensarbete 30 hp
Augusti 2022

Identifying football players who create and generate space

Joakim Michalak

Civilingenjörsprogrammet i informationsteknologi

UPPSALA
UNIVERSITET

Identifying football players who create and generate space

Joakim Michalak

Abstract

The adoption of data analysis within football has seen significant growth for the past several years. Many different metrics for the players exist which are used as key performance indicators. Metrics utilizing machine learning has been available for a long time, however only recently have we seen a much wider adoption amongst the mainstream audience. These metrics are however mostly on-the-ball metrics. They do not necessarily give credit to players who create space for themselves but do not get to the ball, or players who create space for their teammates. In this thesis, two metrics called Space Control and Space Generation are presented which aim to also capture these off-the-ball movements in attacking situations. The two metrics are computed for all players of the Allsvenskan season 2021 and Eredivisie season 2021/2022. By looking at cumulative values of these metrics throughout the seasons, we successfully identified players who create and generate space in large quantities per match.

Teknisk-naturvetenskapliga fakulteten

Uppsala universitet, Utgivningsort Uppsala/Visby

Handledare: Ola Larsson Ämnesgranskare: David Sumpter

Examinator: Lars-Åke Nordén

Sammanfattning

Under de senaste åren har användningen av dataanalys växt inom fotbollsbranschen. Det är vanligt förekommande för fotbollsklubbar idag att att utnyttja dataanalys för att hitta nya fotbollsspelare att värva till klubben. Fotbollsspelare kan bedömas utifrån olika införda nyckeltal som mäter spelarens prestation, där spelare som just utmärker sig i dessa nyckeltal kan vara intressanta för fotbollsklubbarna. De flesta av nyckeltalen som används flitigt idag mäter däremot för det mesta insatser av spelare med boll. Då fotbollen handlar om mycket mer än vad som bara händer med bollen, finns det ett behov av att även mäta spelare hur väl dom presterar utan boll. Tack vare framfarten i utvecklingen av dataanalys inom fotbollsbranschen har det kommit fler datakällor som tillåter oss att titta på just spelare även utan boll.

Vi föreslår två nyckeltal, Space Control och Space Generation, för att kunna hitta fotbollsspelare som utmärker sig i sitt spel inklusive utan boll. Space Control går ut på att titta på en spelares positionering under en match, och baserat på fysikmodeller ge värden beroende på hur mycket positioneringen bidrar med till att laget kan skjuta i mål. En fotbollsspelare som ofta befinner sig i farliga lägen, där chansen till att göra mål är höga, kan således bli tilldelad höga värden oavsett vad spelaren gör med bollen. Space Generation tittar istället på hur mycket Space Control som en fotbollsspelare hjälper att genera till sina lagkamrater. Detta kan ske när en spelare drar med sig motståndarsspelare i samband med en löpning, samtidigt som det skapar yta åt en lagkamrat. Trots att spelaren i fråga kanske inte höjer sitt egna Space Control värden genom sin löpning, kan spelaren höja sina lagkamraters värden vilket är syftet med Space Generation.

Efter att dessa två nyckeltal implementerats, beräknades dess ackumulerade värden för alla fotbollsspelare under hela Allsvenskan 2021 säsong och Eredivisie 2021/2022 säsong. Dessa värden utvärderades först genom diskussioner hos domänexperter, där resultaten var lovande. Andra utvärderingsmetoden gick ut på att jämföra värdena med relevanta egenskaper från spelet Football Manager 2022 i form av en korrelationsanalys. I spelet har varje spelare olika 36 egenskaper, inklusive Work Rate och Off The Ball, vilket var de två egenskaperna som testades mot våra föreslagna nyckeltal. Korrelationsanalysen visade lovande resultat för offensiva mittfältare och anfallare, men mindre bra för mittfältare. Studien visar på att det är möjligt att införa nyckeltal som mäter fotbollsspelares prestationer utan boll.

Contents

1 Introduction

The adoption of data analysis within football has seen a significant growth for the past several years. Top teams from around the world are today using state-of-the-art data analysis to find better patterns in the game, decrease injuries and finding new players to sign [EDS$^+$16] [TS21]. This can give a significant competitive edge, in a period of time where adoption of data analysis is still not fully grown. Certain statistical measures of football players, such as Expected Goals, have been popularized to a mainstream audience, where the metrics utilizes machine learning. The popularized metrics are often derived using data sources categorised as event data, which contains on-the-ball events collected by humans. These include all events such as shots, tackles, and passes during a football match, with further information about the event such as x, y coordinates of the event and related players to the event in question.

Together with the rise of the analytics, we have also seen a development in the availability of data sources, where we see a wider adoption of tracking data which tracks all of the players on the pitch for the full match [RM16]. This allows us to not only look at on-the-ball actions, but also how players do off-the-ball. In this thesis, the purpose is to quantify the off-the-ball movements for attacking players in football with the help of tracking data and to identify the players who excels in those.

The game of football is much more than just what happens on-the-ball. Good scoring opportunities are often a result of a collective team effort, where both the players on-the-ball as well as off-the-ball are contributing factors. A big part of football is to create space in dangerous positions, ideally where the scoring probabilities are the highest. The space in dangerous positions are not necessarily created by the players on-the-ball. Firstly, the player in the dangerous position may never receive the ball such that event data will not capture the player's contribution. Secondly, a player might have created that dangerous space by making a certain run by dragging the opposition players, where this also does not get captured by only event data. This motivates the usage of tracking data in order to capture movements outside the scope of event data.

By combining physics and stochastic models, we can assign a pitch value for each position on the pitch. The pitch value is the probability that a goal will be scored from that position given the surrounding context. It will depend on different parameters, such as the positions and velocities of each player and the position of the ball. An example of a high pitch value, is an unmarked player in an dangerous zone where the scoring probability is high, where there are none or few players along the linear trajectory between the ball to the unmarked player.

With these pitch values, we can assign the maximum individual contribution each player has per attacking phase. The purpose is to identify which players occupy dangerous spaces most frequently. A common metric to judge attacking players is to use Expected Goals, which utilizes machine learning to assign the probability that a shot ends up as a goal [HKBM21]. However, the new metric proposed using pitch values does not require an on-the-ball action from each player, but rather just to occupy dangerous space. This captures broader situations, where players might not receive the ball but still had a positive contribution to its team.

In this thesis, we will introduce two metrics called Space Control and Space Generation which are based on the pitch values [FB18]. Space Control concerns the maximum pitch value occupied individually for each player for a given frame, whereas Space Generation is the Space Control which a player creates for another teammate by making a certain run. Space Generation looks at players who make a run of over a certain threshold within a timeframe, and assigning the Space Control value of each player who was in a certain radius of the origin position of the generating player within a specific timeframe. These metrics are calculated for all of the matches during the Swedish top tier league Allsvenskan season 2021 and Dutch top tier league Eredivisie 2021/2022. Due to confidentiality reasons, only the results of the Allsvenskan season are presented.

The metrics are evaluated by computing the Pearson correlation coefficient between the metrics and relevant stats from the video game Football Manager 2022. The video game has a big database of football players, where the metrics for the players are collected by crowdsourcing within the community [Hoc16]. Football Manager is branded as a realistic simulator game, where the metrics are supposed to reflect reality.

The intention of the two mentioned metrics are to be used in scouting new players, however we also propose other use cases of the pitch values as future work, where it can be used for coaches to identify dangerous attacking situations.

It is recommended to read the rest of the thesis in parallel with the provided Notebook here: Notebook link. The provided Notebook includes more intuitive examples including videos, and also lowers the barrier to reproducibility. The full code is provided in the Github repository [Mic22].

2 Theory

In this section, the necessary theory required for the proposed metrics is explained.

2.1 Event Data

Event data is a human-collected dataset in which all common on-the-ball events during a football match are recorded. Events such as shots, tackles, and passes are manually recorded by humans with its corresponding timestamp during the game and x, y coordinates. There are many different data providers, and considering the manual process during the collection of the data, human error can not be fully avoided.

2.2 Tracking Data

Unlike event data, tracking data is collected by computers using tracking technology to capture the movement of all of the players and the ball on the pitch during a full match. The dataset contains the x, y coordinates for all players, with the identifier being its jersey number, for the full match with a certain sample rate. This no longer introduces the human error as the event data, but rather some noise is introduced from the imperfect tracking technologies. Common pitfalls for the tracking technologies are situations where players are closely clustered where it struggles to identify the players correctly, and the ball position x, y coordinates when the ball is high up in the air. The dataset also includes the timestamp of the tracked frame, which allows synchronisation between event and tracking data. However, since the event data is manually collected, naively synchronising them will always have a margin of error.

2.3 Pitch Control

The purpose of pitch control is to determine which team has the control of each position on the pitch given the current context. A basic implementation of this could be a Voronoi diagram given the positions of all players, as seen in Figure 1. We can see in the figure that player number 2 in red is moving into the space behind number 5 in green, whereas number 5 is running the opposite way. This shows the flaw of only using the player's position as the context, as player number 2 is more likely to control that space. We want to extend the context to include as much relevant information as possible to quantify the control of the pitch for both teams.

There are many different ways of developing a pitch control model depending on how much of the context you want to include as well as how different parameters are chosen. In this thesis, the pitch control model proposed by William Spearman will be implemented [Spe18]. This pitch control model utilizes a broad context of the game which offers a realistic model. While there exist other state-of-the-art pitch control models,

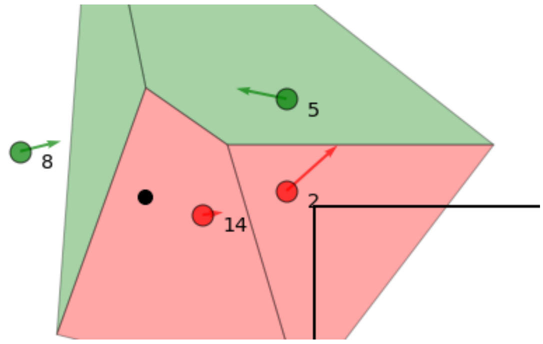


Figure 1 Voronoi diagram using the players' positions. The coloured vertices represent players, where the same colour means the same team, and the black vertex represents the ball. The arrows indicate the players' velocities.

they seem to produce similar results with the main disadvantage of the Spearman's model being the computation time [PA19]. The computation time of the model is not considered to be a bottleneck of this thesis after code optimizations, as the computation time are deemed to be sufficiently fast.

The new model called Potential Pitch Control Field (PPCF) will instead of having a binary control of 0 or 1 as in the Voronoi diagram, rather have continuous probability between 0 to 1. This is to add uncertainty to the model, to better model the game of football since we can never guarantee that a player will successfully control the ball. Another important context to consider is the current ball position. If a player wants to make a pass from one position to another, it is important to consider how many players could intercept the ball along the trajectory of the pass before it reaches its end position. Therefore, if many players are close along the passing trajectory, we should decrease the control probability for the recipient because of the added probability of the pass getting intercepted by the opponents. This will require some assumptions about the reaction time of the players, the potential velocities of the players towards the ball, and the velocity of the ball.

We are trying to calculate the probability of control for a team given the current context. This is done by calculating the individual control of each player j, which is equivalent to the differential equation in Equation 1.

$$\frac{dPPCF_j}{dT}(t,\vec{r},T|s,\lambda_j) = \Big(1 - \sum_k PPCF_k(t,\vec{r},T|s,\lambda_j)\Big) f_j(t,\vec{r},T|s)\lambda_j \tag{1}$$

In Equation 1, the $f_j(t,\vec{r},T|s)\lambda_j$ term is the probability that player j will intercept the ball by reaching location $\vec{r}$ at time t in less than time T, given that the ball will be passed from current position to $\vec{r}$ in a straight-line trajectory. $f_j(t, \vec{r}, T|s)\lambda_j$ is a sigmoid function where s is the standard deviation of the player's arrival time in order to take the uncertainty into consideration. Other important parameters in the model to consider are the reaction times, acceleration, and maximum velocity of the player, the control rate λ_j which is the inverse time it would take player j to control the ball, and the average ball velocity.

Equation 1 is integrated over T from 0 to ∞. The resultant matrix includes a 0 to 1 probability for each grid on the pitch, which is the probability of control for a certain team. An example of the final PPCF can be seen in Figure 2.

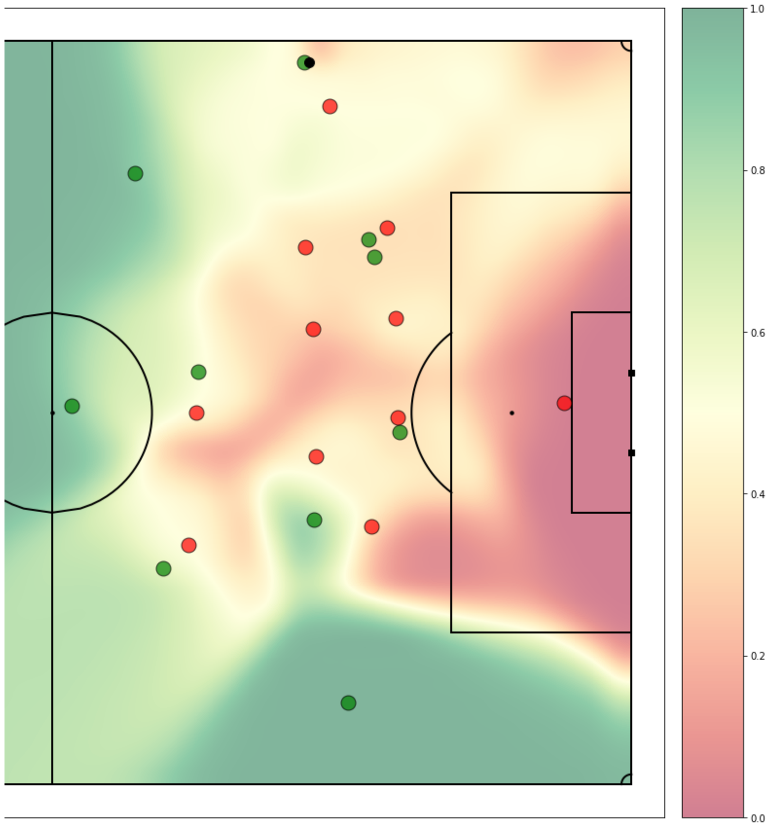


Figure 2 PPCF model in practice, where the green team is attacking from left to right. Green colours indicate that the attacking team has full control, whereas red colours indicate that the defending team has full control. The bright yellow colours are neutral, where neither team is the favourite to control the space.

2.4 Expected Possession Value

The purpose of Expected Possession Value (EPV) is to calculate the probability of scoring for each location on the pitch given the surrounding context of the frame, which can be described by the following equation:

$$EPV = P(Goal|context) \tag{2}$$

Similar to pitch control, there are a lot of ways of creating this model based on how much of the context is used for the model. An example of complex and state-of-art application of EPV is the model proposed by Javier Fernandez et al. [FBC21], which utilizes deep neural networks with the dynamics of the 22 players and the ball all into consideration of the model.

For this thesis, a more simple model will be used which was trained by Laurie Shaw, where the context is the current ball position. An example of the model can be seen in Figure 3. As expected, the highest EPV values are just around the goal. The model was trained by breaking the pitch into certain amounts of zones, and building a transition matrix which has the probabilities of moving between different states. The states are the zones on the pitch, as well as ending up in goal or loss of possession. The probabilities are calculated by looking at how the transition between the states happen in hundreds of real football based on event data. The EPV model is then the probability of moving from each zone on the pitch to the goal state. This model was first proposed by Sarah Rudd, who in similar fashion applied Markov Chain to obtain the transition matrix [Rud11].

2.5 Pitch Control * EPV

Both of the pitch control and EPV models give matrices with values of the pitch, with the dimension being dependent on how many equal-sized cells the pitch is broken up into. The pitch control has for each cell a value of how much between 0 to 1 a team controls that location, whereas the EPV model gives us the probability of scoring of the given cell. The limitation of only using the pitch control model is that it does not consider that not every location at the pitch is equally worth, where intuitively controlling the dangerous zones are more important. We can multiply the two mentioned matrices assuming that they are of the same dimension. This gives us the pitch values for a given context, and will be used for the upcoming metrics introduced in this thesis.

Figure 4 shows an example of the matrix created by multiplying pitch control and EPV. Player number 7 is controlling the space in a location where the EPV is high. This example also demonstrates an example of when this model can be useful. If player number 7 never receives the ball in this example, event data would not capture his contribution

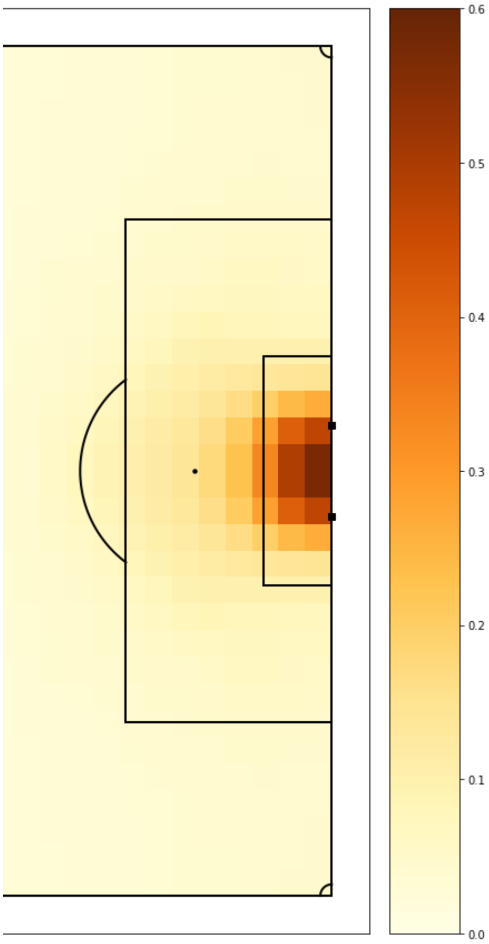


Figure 3 Right part of a football pitch, where the attacking team is attacking from left to right. The colour values are the EPV values for each location.

to the team. However, from the figure we can see that he creates high pitch values by making a smart run.

2.6 Space Control

Space Control is a metric for individual players to capture the types of runs which was shown in Figure 4. As seen in Equation 1, the PPCF is firstly calculated per player. Hence, we have individual pitch control matrices for every individual player and can multiply with the EPV matrix. Rather than using the team's PPCF as seen in Figure 4, only each individual's PPCF is used to calculate the Space Control. The highest value in the resultant matrix is then assigned as the Space Control value for the player, for the

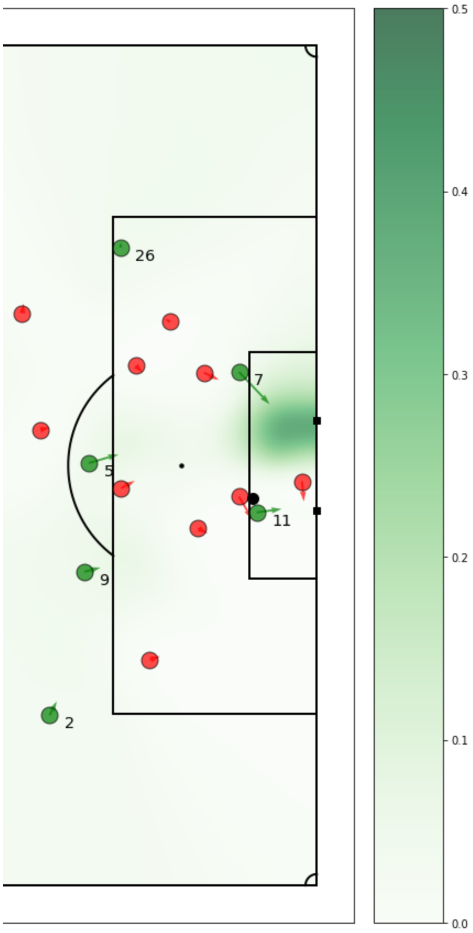
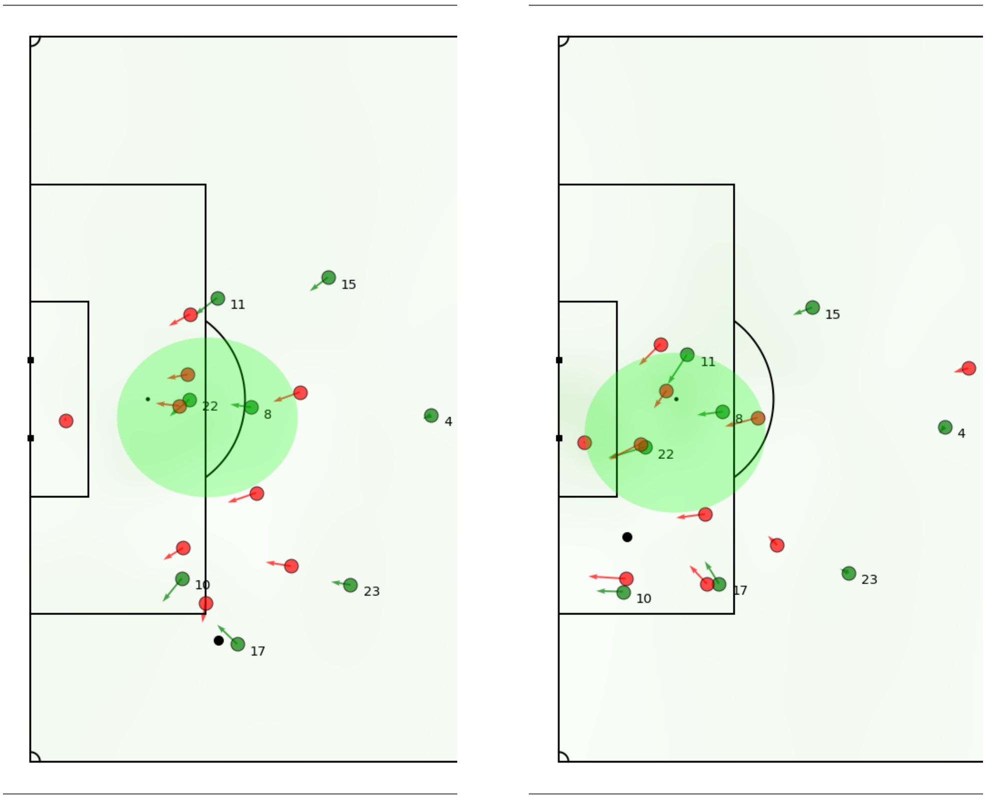


Figure 4 Right part of a football pitch, where the green team is attacking from left to right.

given frame. This metric allows us to capture smart movements from players, which were previously not captured by event data.

2.7 Space Generation

Rather than only looking at how much your own movement contributed in individual Space Control, we can also look at how much Space Control was created for a player's teammates by making a certain run. Players will sometimes make a run which drags opponent defenders along, which could potentially create higher Space Control for the teammates, even though individually it did not result in higher Space Control for the runner. An example of this is shown in Figure 5, where player number 22 makes a run towards the goals which drags the defenders with him. This leads to space created for his teammate number 8, who ran towards the origin position of the player 22. In this example, player number 22 has generated space and should thus be assigned a Space Generation value, dependent on the Space Control by player number 8.

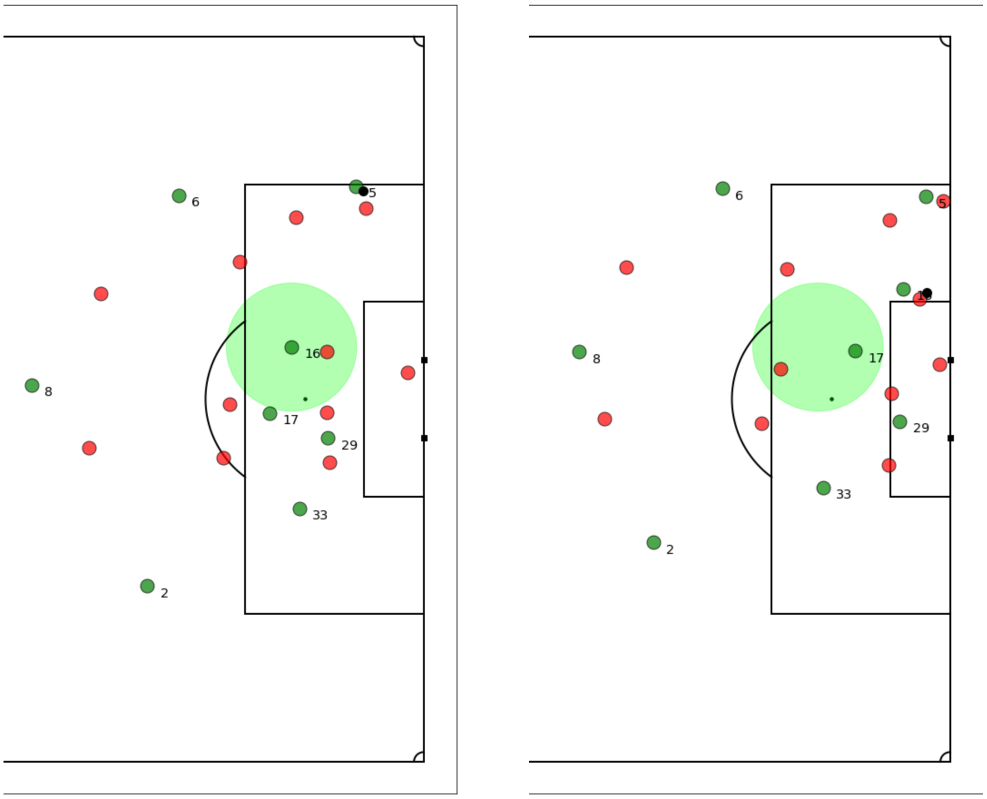


(a) Player number 22 makes a run towards the goal.

(b) 1.5 seconds after, space has been created for number 8 thanks to number 22 who dragged the defenders closer to goal.

Figure 5 Example of a player who generates space control for a teammate.

The approach for calculating the Space Generation is similar to Javier Fernandez et al. work [FB18]. We look at situations where the attacking team is in possession of the ball, and a player makes a run of at least x metres within a time window $[t, t + w]$. For each such player, we look for all teammates within a radius of r metres of the origin position at time t and their corresponding Space Control value. After the player makes a run of more than s metres, we look in the same radius for the teammates' Space Control values of all frames until time $t + w$. If any of the combined Space Control values within the radius are higher than at time t, we will assign the combined Space Control value to the player making the run. A concrete example of this is shown in Figure 6.



(a) Player number 16 will make a run of more than s metres within w seconds, where a circle of radius r metres will be created. The radius is empty.

(b) New frame in the interval $[t, t + w]$. Player number 17 has entered the radius and the Space Control value is higher than at the beginning of the run.

Figure 6 Example of Space Generation with parameters w = 3, r = 6, s = 4

2.8 Football Manager metrics

Football Manager 2022 includes a big database of football players around the world, including all of the players who played during Allsvenskan 2021 and Eredivisie 2021/2022. Each outfield player has in total 36 different attributes within the categories technical, mental and physical, each ranging from 1 to 20. The relevant attributes for this thesis are Off The Ball and Work Rate, which are the attributes which will be tested against the proposed metrics to measure the correlation.

The Off The Ball metric is described as ”This attribute reflects the player’s ability to move when not in possession of the ball, making themselves available to receive a pass in a dangerous position”. The Work Rate metric is described as ”This attribute reflects the player’s willingness to work to his full capacity, going above and beyond the call of duty".

2.9 Expected Goals

An Expected Goal model is a predictive model, which estimates the probability of scoring for a given shot. The models' parameters are derived from the given context during the shot. Common parameters are the distance and angle between the shooting player and the goal.

2.10 Pearson correlation coefficient

To evaluate the proposed metrics, we want to measure the linear correlation between the metrics and Football Manager's metrics. This is done by computing the Pearson correlation coefficient r, with its definition in Equation 3.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3}$$

We have for $r = \pm 1$ a complete linear relationship between variables x and y, whereas for $r = 0$ we have complete linear independence between variables x and y.

3 Implementation

The implementation of the metrics, from the datasets including preprocessing to the final metrics, is explained in this section.

3.1 Dataset

Event and tracking data were both necessary to calculate the Space Control and Space Generation values for full seasons. For this thesis, Wyscout was the event data provider whereas Signality was the tracking data provider for Allsvenskan and TRACAB was the tracking data provider for Eredivisie.

3.1.1 Wyscout

The Wyscout data contains of one file per football match, where each file has all of the human-collected events on-the-ball throughout the match as distinct rows. Each row has a corresponding timestamp as well as other information regarding the event. Each row is also assigned a possession identifier. Wyscout defines a possession as a sequence of events with the ball of the same team. Consecutive events from the same team would thus be assigned with the same possession identifier.

Each possession has different tags, such as whether it is an attacking possession or if the possession ends up in a shot. This is useful since this thesis looked at movement in attacking scenarios, in which we could calculate the desired metrics only during the intervals of each attacking possession. Rather than calculating the metrics for full matches which would be computationally expensive, the metrics were only calculated during the interval of the attacking possessions given by the Wyscout event data. Thus, the purpose of the Wyscout event data was to find for which tracking data frames the metrics should be computed.

3.1.2 Signality and TRACAB

The calculation of the metrics required the usage of tracking data to capture all of the players on the pitch, rather than only the on-the-ball events, which was provided by Signality and TRACAB. Similar to the Wyscout data, Signality collects each match into separate $json$ files but also separated by each half of the match. Each row in the files contains a timestamp, two lists for all of the home players respectively away players who are tracked on the pitch with their x, y coordinates where each player is separated by its jersey number, as well as the ball x, y coordinates. The TRACAB files contain the same information but in a different structure. Both tracking datasets have a sampling rate of $25Hz$.

3.2 Preprocessing

3.2.1 Wyscout

Each Wyscout $json$ file was loaded into a Pandas DataFrame. Pandas is a Python library which allows for intuitive data analysis and manipulation [M$^+$10]. The purpose of the Wyscout data was to find attacking situations. This was done by filtering the data by only including events which were tagged as attacking possessions. The metrics were computed for each player during the attacking possessions, such that we needed the start and end time of each attacking possession provided by the Wyscout data, which required synchronisation between the two datasets.

The synchronisation issue was handled by looking at the offset between the first row of both datasets, where a new column was added to the Wyscout DataFrame by adding the offset to the Wyscout timestamps. The first event in the Wyscout data is always the kickoff pass made in the football match. However, in the Signality files, the first frame is not always the pass frame which leads to possible synchronisation errors. In this thesis, perfect synchronisation was not required such that the error margins are acceptable.

With the new timestamps which took the offset into account, all unique attacking possessions were stored in a list with their corresponding start and end timestamps. The timestamps were found by taking the timestamp of the first and last event for each unique attacking possession.

3.2.2 Signality and TRACAB

For each match, the two Signality $json$ files for the halves were merged into the same Pandas DataFrame. After the merge, each match had approximately 135000 rows considering each football match is 90 minutes long and the sample rate being $25Hz$. Considering the future operations on the dataset which required the positions of the players, the list of all players was flattened such that each resulting DataFrame had a separate column for each player's x and y positions.

The TRACAB dataset was preprocessed similar to the Signality dataset, such that both resulting DataFrames had identical column names. In that way, all future operations worked for both processed datasets.

The tracking technology of the Signality dataset was not perfect, and thus missing data was a common occurrence. This was handled by applying linear interpolation of each player's x, y coordinates between real values. Since we had the coordinate columns for each player, it also included players who are not on the pitch. Hence, the limit area of the linear interpolation was between real values.

Each x position was normalized to the $[-53, 53]$ interval and y to $[-34, 34]$, such that $(0, 0)$ is the centre of the pitch and to represent the coordinates in metres. The data was also down-sampled to $5Hz$ to make future operations less computationally expensive, while still maintaining enough information.

Velocities for each player were calculated by taking the x, y coordinate difference between consecutive rows and dividing by the time difference between the rows. Velocities which exceeded 12 m/s were set to none as the velocities are unrealistic and likely to be noise. The velocities were then smoothed by applying moving average with a window of 7.

3.3 Models

For the two implemented metrics, Space Control and Space Generation, the prerequisite models are the Pitch Control and Expected Possession Value. The Pitch Control model was originally created by Laurie Shaw and is based on the Spearman's model, with slight modifications [Sha21]. The modifications were done by applying Numba as well as adapting the code for the Signality data. Numba is a Just-In-Time compiler which translates certain Python code into machine code, which drastically increases the performance [LPS15]. The parameters of the Pitch Control model were chosen to be equivalent to Shaw's code. Shaw also provides a *csv* file containing the Expected Possession Value for each position on the pitch, where each grid is the value of transitioning to a goal state given its current position [Sha21]. This *csv* file was used for the Expected Possession Value model.

3.3.1 Computing the metrics for full season

Given that we had the Pitch Control and Expected Possession Value models, we could calculate the Space Control and Space Generation metrics for each player given a certain frame. The next step was to compute the metrics for each player during a full season, such that the players could be evaluated on this metric. However, there are multiple ways of how and when to assign the metrics for each player.

As previously described in 3.2.1, we wanted to look at attacking possessions where the timestamps were provided by the Wyscout data. The new timestamps which took the offset into account were used, such that we could loop over all attacking possessions and iterate through all tracking data frames in the attacking possession. The two metrics could then be computed for each of the frames, under the condition that at least one player in the attacking team was close to the ball. For this thesis, a player was considered to be close to the ball if the distance is smaller than 2 metres.

Rather than adding up all Space Control and Space Generation values per attacking possession, we instead tried to find the best combination of the values where each consecutive assigned value must be at least three seconds away from the previous assigned value. As an example, in a nine second attacking situation, each player could at most be assigned three different values for each metric.

The best combination was greedily searched by looking at the highest value of both Space Control and Space Generation within the attacking situations. Thereafter, it looked for the highest value in both right and left directions within the time interval $[t + 3, t + 4.5]$ for the right direction and $[t - 3, t - 4.5]$ for the left direction, where t is the timeframe in seconds of the current frame. This was computed until it reached the start and end frame of the attacking possession. The final result gave a table for each player in the attacking team, where each row was a Space Control or Space Generation value for a given player. Each row with the same player and for the same metric would have a time difference of at least three seconds.

3.4 Evaluation

The evaluation was performed by computing the Pearson correlation coefficient as described in Equation 3, which required variables x and y with their the means. The assumption for this thesis was that Space Control + Space Generation normalized per 90 minutes played should have a linear relationship with Football Manager's metrics Off The Ball and Work Rate. Hence, for each player i we have:

$$\begin{aligned} x_i &= (\text{Space Control / 90 min})_i + (\text{Space Generation / 90 min})_i \\ y_i &= (\text{Off The Ball})_i * \frac{(\text{Work Rate})_i}{20}. \end{aligned}$$

To benchmark the coefficients, we compared it to instead using a Expected Goals (xG) model. The Wyscout event data has pre-computed xG values for each shot, such that we computed the cumulative xG for each player normalized per 90 minutes played. Thus, we have:

$$\begin{aligned} x_i &= (\text{xG / 90 min})_i \\ y_i &= (\text{Off The Ball})_i * \frac{(\text{Work Rate})_i}{20}. \end{aligned}$$

4 Results

In this section, the top players for Space Control and Space Generation are presented, as well as the correlations between the metrics and the Football Manager attributes.

4.1 Top Players in Allsvenskan 2021

A top 10 list is presented for Space Control and Space Generation for Allsvenskan 2021 with respect to their playing position. The top 10 lists for Eredivisie are not presented due to confidentiality reasons. The lists are filtered for players with more than 600 minutes played during the season.

Player	Team	Age	Minutes Played	Space Control per 90 min
Christoffer Nyman	IFK Norrköping	29	857	1.53
Nikola Djurdjic	Degerfors IF	36	611	1.35
Kalle Holmberg	Djurgårdens IF	29	1473	1.30
Antonio-Mirko Čolak	Malmö FF	28	2128	1.28
Bojan Radulovic	AIK	22	632	1.27
Astrit Selmani	Hammarby IF	25	1962	1.26
Jasse Tuominen	BK Häcken	26	629	1.24
Christian Kouakou	IK Sirius	27	2464	1.22
Per Frick	IF Elfsborg	30	2027	1.10
Marcus Antonsson	Halmstads BK	30	2283	1.04

Table 1 Top 10 Space Control list for strikers in Allsvenskan season 2021.

Player	Team	Age	Minutes Played	Space Control per 90 min
Joel Asoro	Djurgårdens IF	23	999	1.11
Isak Jansson	Kalmar FF	20	1200	1.08
Niklas Bärkroth	Djurgårdens IF	30	1437	1.01
Jacob Ondrejka	IF Elfsborg	19	760	1.00
Nils Fröling	Kalmar FF	22	1739	1.00
Adi Nalić	Malmö FF	24	1572	0.96
Emmanuel Banda	Djurgårdens IF	24	685	0.96
Oliver Berg	Kalmar FF	28	2600	0.95
Veljko Birmančević	Malmö FF	24	1961	0.95
Akinkunmi Amoo	Hammarby IF	19	1852	0.94

Table 2 Top 10 Space Control list for attacking midfielders in Allsvenskan season 2021.

Player	Team	Age	Minutes Played	Space Control per 90 min
Erik Lindell	Degerfors IF	26	1367	0.79
Vladimir Rodić	Hammarby IF	28	1056	0.79
Williot Swedberg	Hammarby IF	18	756	0.77
Adam Ståhl	IK Sirius	27	1586	0.75
Simon Olsson	IF Elfsborg	24	1318	0.74
Anders Christiansen	Malmö FF	31	1486	0.72
Moustafa Zeidan	IK Sirius	23	1565	0.65
Hampus Finndell	Djurgårdens IF	21	2192	0.64
Dennis Collander	Örebro SK	19	1449	0.63
Axel Lindahl	Degerfors IF	27	721	0.63

Table 3 Top 10 Space Control list for midfielders in Allsvenskan season 2021.

Player	Team	Age	Minutes Played	Space Generation per 90 min
Christoffer Nyman	IFK Norrköping	29	857	1.22
Bojan Radulovic	AIK	22	632	0.95
Per Frick	IF Elfsborg	30	2027	0.88
Kalle Holmberg	Djurgårdens IF	29	1473	0.82
Jasse Tuominen	BK Häcken	26	629	0.75
Antonio Čolak	Malmö FF	28	2128	0.73
Christian Kouakou	IK Sirius	27	2464	0.67
Astrit Selmani	Hammarby	25	1962	0.66
Samuel Adegbenro	IFK Norrköping	26	2244	0.66
Marcus Antonsson	Halmstads BK	30	2283	0.63

Table 4 Top 10 Space Generation list for strikers in Allsvenskan season 2021.

Player	Team	Age	Minutes Played	Space Generation per 90 min
Isak Jansson	Kalmar FF	20	1200	0.71
Joel Asoro	Djurgårdens IF	23	999	0.62
Adi Nalić	Malmö FF	24	1572	0.59
Niklas Bärkroth	Djurgårdens IF	30	1437	0.57
Yukiya Sugita	IK Sirius	29	1067	0.57
Alexander Bernhardsson	IF Elfsborg	23	796	0.56
Emmanuel Banda	Djurgårdens IF	24	685	0.51
Oliver Berg	Kalmar FF	28	2600	0.49
Noah Shamoun	Kalmar FF	19	672	0.49
Kevin Yakob	IFK Göteborg	21	897	0.48

Table 5 Top 10 Space Generation list for attacking midfielders in Allsvenskan season 2021.

Player	Team	Age	Minutes Played	Space Generation per 90 min
Simon Olsson	IF Elfsborg	24	1318	0.46
Anders Christiansen	Malmö FF	31	1486	0.44
Williot Swedberg	Hammarby	18	756	0.43
Hampus Finndell	Djurgårdens IF	21	2192	0.37
Kristoffer Khazeni	IFK Norrköping	26	898	0.32
Ísak Jóhannesson	IFK Norrköping	19	1305	0.31
Moustafa Zeidan	IK Sirius	23	1565	0.3
Johan Karlsson	IK Sirius	20	610	0.29
Robert Gojani	IF Elfsborg	29	1217	0.29
Dennis Collander	Örebro SK	19	1449	0.28

Table 6 Top 10 Space Generation list for midfielders in Allsvenskan season 2021.

4.2 Correlation

x	Player Position	r coeficient
Space Control + Space Generation	Striker	0.30
xG	Striker	0.32
Space Control + Space Generation	Attacking Midfielder	0.57
xG	Attacking Midfielder	0.19
Space Control + Space Generation	Midfielder	0.16
xG	Midfielder	0.10

Table 7 Pearson correlation coeficient for different metrics and positions.

In table 7, we can see the r coefficients for each position with an alternating x variable between the proposed metrics and benchmark metric xG. The sample consists of all players in those positions, who played more than 600 minutes in Allsvenskan 2021 and Eredivisie 2021/2022. The y variable for all cases is Off The Ball $* \frac{\text{Work Rate}}{20}$.

5 Discussion and conclusions

The top 10 lists seem to produce promising results. In brief conversations with people of domain knowledge, the results seem somewhat inline with their intuition. This is partly due to the large appearance of players in top teams within their leagues, which as expected rank highly in these lists. Players in lower ranked teams will naturally tend towards lower values within Space Control and Space Generation, as they do not create as many opportunities.

The correlations between the proposed metrics and the Football Manager's metrics seem interesting, as there seem to be moderate correlation for strikers ($r = 0.30$), even stronger for attacking midfielders ($r = 0.57$), but only weak for midfielders ($r = 0.16$). In order to evaluate these coefficients, the correlation between xG and the Football Manager's metrics was also computed, such that we can use them as a benchmark. The proposed metrics Space Control and Space Generation were only slightly worse for strikers, significantly better for attacking midfielders, and somewhat better for midfielders.

One reason for the weak correlation for the midfielders is that defensive midfielders are included in that list, who will naturally player further away from the opposition's goal during the games and not necessarily move into the opposition's penalty box. The two metrics, Space Control and Space Generation, are mainly rewarded for movements in dangerous zones, which are most occurring in the opposition's penalty box. The Football Manager metric Off The Ball, given its definition, seems to apply differently based on the role of the player.

A defensive midfielder who is deemed to be good off-the-ball is not necessarily a player who makes movements into dangerous zones, but rather a player who opens up space in parts of the pitch further away from the opposition's goal, which are not captured by Space Control and Space Generation. Hence, the two proposed metrics are not useful to measure the off-the-ball movements of players who do not advance on the pitch as other players.

A disadvantage of this evaluation is that the Football Manager data does not necessarily reflect reality. Although there is an attempt to model the players' attributes perfectly, it can not be perfect. The approach of crowdsourcing the attributes of the players will not yield perfect results.

The purpose of this thesis was to identify players who create and generate space. By introducing the two metrics, Space Control and Space Generation, and computing them for full seasons of Allsvenskan 2021 and Eredivisie 2021/2022 we created top 10 lists for strikers, attacking midfielders and midfielders. These metrics can be used in a complementary way together with more common on-the-ball metrics such as xG in order to evaluate attacking football players.

6 Future work

In this thesis, the metrics are proposed with the intention of using it in scouting purposes. However, the use cases can also expand outside of scouting. For football coaches who quickly want to find dangerous attacks for a given match, an option with today's popular metric is to look at attacks which end up with high xG. This can be used in order to find patterns of how dangerous attacks are created. However, those would require the attacks to end up in a shot. By instead using Space Control, we can for every frame look at the combined values for all players, and in such a manner find dangerous attacks. This no longer requires the attack ending up with a shot. An example of this is if a player is through on goal but gets intercepted just before the shot, where it would result in zero xG due to no shot taken but high Space Control.

A common theme within the top 10 lists was that most of the players belonged to the top teams of the league. This is expected due to how the metrics work, but must be taken into consideration when inspecting the lists. There are certain players who rank highly in both metrics, but play for a top team within the league. This does not necessary imply that the player is the best off-the-ball player in the league, but can rather be an effect of playing for one of the top teams. It is likely that the player would not reach as high values playing in lower ranked teams, even though the abilities of the player would be the same. For future work, it could be interesting to also measure the quality rather than just the quantity. Instead of normalizing per 90 minutes, a possible suggestion could be to normalize per attacking situation.

To further improve on the metrics, it would be useful to do more qualitative analysis in cooperation with people of domain knowledge.

References

[EDS$^+$16] F. E. Ehrmann, C. S. Duncan, D. Sindhusake, W. N. Franzsen, and D. A. Greene, "GPS and Injury Prevention in Professional Soccer," in *Journal of Strength and Conditioning Research*, no. 30, 2016, pp. 360–367.

[FB18] J. Fernández and L. Bornn, "Wide open spaces: A statistical technique for measuring space creation in professional soccer," 03 2018.

[FBC21] J. Fernández, L. Bornn, and D. Cervone, "A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions," *Machine Learning*, vol. 110, no. 6, pp. 1389–1427, may 2021.

[HKBM21] M. Herold, M. Kempe, P. Bauer, and T. Meyer, "Attacking Key Performance Indicators in Soccer: Current Practice and Perceptions from the Elite to Youth Academy Level," in *J Sports Sci Med*, no. 20, 2021, pp. 158–169.

[Hoc16] A. Hocquet, "Football Manager: Mutual Shaping between Game, Sport, and Community," *Journal of media studies and popular culture = Revue d'études des médias et de culture populaire*, vol. 6, no. Special Issue, Apr. 2016.

[LPS15] S. K. Lam, A. Pitrou, and S. Seibert, "Numba: A llvm-based python jit compiler," in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, 2015, pp. 1–6.

[M$^+$10] W. McKinney *et al.*, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, vol. 445. Austin, TX, 2010, pp. 51–56.

[Mic22] J. Michalak, “JoakimMich,” https://github.com/JoakimMich/masterthesis, 2022.

[PA19] F. J. Peralta Alguacil, “Modelling the collective movement of football players,” 2019.

[RM16] R. Rein and D. Memmert, “Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science,” in *SpringerPlus 5*, no. 1410, 2016.

[Rud11] S. Rudd, “A framework for tactical analysis and individual offensive production assessment in soccer using markov chains,” in *New England Symposium on Statistics in Sports*, 2011.

[Sha21] L. Shaw, “LaurieOnTracking,” https://github.com/Friends-of-Tracking-Data-FoTD/LaurieOnTracking, 2021.

[Spe18] W. Spearman, “Beyond Expected Goals,” in *MIT SLOAN Sports Analytics Conference*, 2018.

[TS21] P. Thakkar and M. Shah, “An Assessment of Football Through the Lens of Data Science,” in *Data. Sci.*, no. 8, 2021, pp. 823–836.