A Statistical Analysis of Inverted Wingers in Soccer

Bachelor Thesis Applied Mathematics
7.5 ECTS

Bo Wardenier (5938535)

June 16, 2021

Universiteit Utrecht

Under supervision of Pr. Dr. Ir. C.W. Oosterlee
Faculty of Science

Acknowledgements

I would like to thank Pr. Dr. Ir. C. (Kees) Oosterlee for the guidance provided during writing my bachelor thesis and for the time he reserved for our weekly meetings.

I would also like to thanks StatsBomb for providing data with which I could conduct my experiment.

Nomenclature

$(X)_{n\geq 0}$ stochastic process.

$g_{x,y}$ goal probability.

h home team.

I state space.

$L(\cdot,\cdot)$ loss function.

$m_{x,y}$ move probability.

P probability function.

$s_{x,y}$ shoot probability.

T transition matrix.

v visiting team.

Acronyms

MLB Major League Baseball.

OBA On Base Average.

SPADL Soccer Player Action Description Language.

VAEP Valuing Actions by Estimating Probabilities.

xG Expected goals.

xT Expected threat.

Table of contents

1 Introduction

The first time I watched the movie *Moneyball*, I did not know what to expect. The movie is based on the eponymous book by Michael Lewis and tells the story of general manager Billy Beane of the Oakland Athletics, a baseball club whose budget is just a fraction of the budget of bigger clubs in the Major League Baseball (MLB), such as the New York Yankees [1]. In the 2001 season Billy is beaten in the World Series, the final games of the season which decide who wins the trophy. After that a lot of key players, such as Jason Giambi and Jason Isringhausen transfer to bigger clubs. To compensate for this loss, Billy attracts a statistician to help him rebuild the team. A quote in the movie from this statistician reads as:

"*There is an epidemic failure within the game what is really happening. And this leads people who run Major League Baseball teams to misjudge their players and mismanage their teams.*"

Until that time, different statistics were used in baseball to determine the level of a player. These statistics among others include the number of stolen bases or the batting average (number of hits divided by the number of at bats), but since these statistics were invented in the late 19th century by H.A. Dobson Washington, they were quite outdated [2]. In the season of 2002 the Oakland Athletics switched from these traditional statistics to the so-called Sabermetrics. These Sabermetrics were the next step in baseball statistics and were invented by Bill James in the 1970's, who worked as a security guard [3]. These Sabermetrics gave a new twist to the old statistics and included on base average (OBA) and slugging percentage. As of today the feud between opponents and supporters of Sabermetrics is still going on.

"*People who run ball clubs, they think in terms of buying players. Your goal shouldn't be to buy players. Your goal should be to buy wins. And in order to buy wins, you need to buy runs.*"

Not just in baseball, but in every sport imaginable data and statistics play a vast part. Similarly in soccer, where old-fashioned statistics, such as the number of goals or assists or even expected goals (xG), are due for renewal. A soccer player can contribute in many different ways to a win for his or her team. A well-timed interception can lead to a scoring opportunity for a forward or a through ball from a defender might rip the defense of the

opposing team apart, but until recently we could not measure the influence of these kinds of actions.

This all comes down to the fact that soccer is a far more complex game to model than baseball. A baseball match follows a certain course of play: the pitcher throws the ball to the batter, he hits the ball (or not) and the ball will fly into the field, after which the field players try to make an out as soon as possible. The actions of the players are quite predictable and the game is therefore easier to understand mathematically than a soccer game is, in which at every second a player can do many different things with the ball.

However, with the rise of machine learning it became a lot easier to extract certain aspects from (big) data sets. The only missing step was modeling a soccer match in a way that all contributions (offensive and defensive) are accounted for. Two different methods have been developed to tackle this problem. The first is designed by Karun Singh, a data science graduate from Cornell University and a soccer enthousiast. He created the method of expected threat, which is explained in chapter three.

The other method was constructed in 2019 by a research team of KU Leuven and employees of SciSports, a company specialized in sports data. They developed a framework for converting the existing data format used in soccer, to a better interpretable language and from there rating players' actions [4]. This method will be explained further in chapter four. Both ways of valuing actions give us the possibility to objectively check assumptions that exist in present-day soccer tactics.

In soccer there exist a few myths about what kind of player should be playing at a certain position, which relies heavily on the tactical view of the trainer. For instance, some think that a left or right back should position himself more towards the goal of the opponent on the pitch, and therefore create more threat. A wider spread and newer myth is that a winger should be inverted, meaning that he prefers the foot opposite of the side that he plays on, so a left winger should prefer to shoot and pass with his right foot and vice versa. We call this kind of player an inverted winger. This phenomenon is what we will be researching in this thesis. In chapter three and four, two methods are proposed for objectively rating soccer players and at the end of chapter four the choice between the two is motivated. In chapter five our experiment is introduced, which advantages and disadvantages belong to an

inverted winger? How does an inverted winger relate to a non-inverted one? In chapter six a conclusion is proposed concerning these questions.

2 Mathematical Background

To make our lives easier, the definitions and explanations that are used in this thesis are summarized in this chapter. When certain definitions or theorems are used, we will refer back to this chapter and the definition or explanation concerned.

Definition 1. Finite Markov Chain
A finite Markov Chain is a stochastic process $(X)_{n\geq 0}$ with finite state space I and transition matrix T. Furthermore, it should hold that $P(X_n = i_n|X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, ..., X_0 = i_0) = P(X_n = i|X_{n-1} = i_{n-1})$, where $i_n, ..., i_0 \in I$.

Definition 2. Transition Matrix
A transition matrix, of n by m, T for a Markov Chain is a stochastic matrix where each entry $t_{i,j}$ describes the one-step transition probabilities from state i to j for every $0 \leq i \leq n, 0 \leq j \leq m$. Furthermore $t_{i,j} \geq 0$ for all $0 \leq i \leq n, 0 \leq j \leq m$ and $\sum_{j=1}^{m} t_{i,j} = 1$ for any $0 \leq i \leq n$.

Furthermore, it holds for a homogeneous Markov Chain that $P(X_n = i_n|X_{n-1} = i_{n-1}) = P(X_1 = i_1|X_0 = i_0) = t_{i_0,i_1}$; the transition probabilities are time independent.

Explanation 1. CatBoost Algorithm
In chapter 4, we will implement the CatBoost algorithm to train our program using supervised machine learning. CatBoost belongs to the family of Gradient Boosting algorithms, and is a technique used for regression and classification tasks. The goal is to let a model $\hat{F}(x)$ estimate output variable, y, based on vector of input variables, x_i. Therefore we introduce a loss function $L(y, \hat{F}(x))$, which we are going to minimize.

$$\hat{F} = \arg\min_{F} E_{x,y}[L(y, F(x))]$$

We will try to construct $\hat{F}$ as a weighted sum of weak learners, $\sum_{i=1}^{M} \gamma_i h_i(x)$, where γ_i define the weights and $h_i(x)$ are the weak learners. We then turn to our trainingsset of pairs $\{x_i, y_i\}$ for $i = 1, ..., n$. We then start by defining $F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$ and for $m \geq 1$ we update our model:

$$F_m(x) = F_{m-1}(x) + \arg\min_{h_m \in \mathcal{H}} \left[\sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h_m(x_i)) \right]$$

For $m = 1, ..., M$. In words, we take the model and add another base learner to minimize the remaining error. This is what is called training the model [5].

3 Expected threat (xT)

3.1 Introduction

In soccer a great number of actions is overlooked, the only rankings that play a role in the evaluation of soccer players are the number of goals or assists for attackers and the number of duels won for defenders, but a lot more happens on the soccer pitch than just these actions. And maybe the more interesting question, how do these actions influence a team's chance of scoring? Many actions in a soccer match do not directly affect the score, and that makes it even more difficult to determine the contribution an action has made. A pass over the axis of the field from left to right does not influence the score directly, but it can create an opening and therefore increase the scoring probability for a team. So we would like to develop a new statistic that captures all of this information and will output a rating based on this information. In this section the method of expected threat is explained.

3.2 Goals

The notion and method of expected threat (xT) is built upon several goals. The goal for this approach is to reward individual player actions on te ball in build up play and close to goal. Expected threat can only value ball progressing actions, which means it is limited to rating passes and dribbles. However, these two kinds of actions made up for over 73% of the total actions in the English, Spanish, German, Italian, French, Dutch, and Belgian top divisions in the 2012/2013 season up and until the 2017/2018 season[4].

Another goal is that the given value should be independent of the outcome of the possession. If Kevin de Bruyne passes the ball brilliantly to Sergio Agüero, but Agüero loses the ball, then the model will still need to take de Bruyne's pass into account. The goal is to determine the value of threat for a position on the pitch. We estimate this by taking into account that a player can shoot from a given position, but that he also has the possibility of moving the ball to a more threatening position on the pitch.

3.3 Set up

The expected threat model is constructed as a Markov chain, as we have defined in definition 1 of chapter 2. The state space for this model is a partition of the soccer field in a grid of n by m. Where each pair of coordinates (x, y) represent another zone on the field. Singh uses a grid of

16 by 12 in his blog and therefore $16 \cdot 12 = 192$ zones, but also mentions that the choice of grid size is dependent on the amount of data you have [7]. The more data you have to the more precise expected threat can be calculated per zone. In spirit of chapter two, we define our state space as $I = \{(x, y) | 1 \leq x \leq n, 1 \leq y \leq m\}$.

In every zone on the pitch a player has a few different options with the ball: he can either shoot the ball from that zone or he can either move the ball. Moving can be passing the ball to a player on the same team or dribbling with the ball. These probabilities are defined as:

Move probability $m_{x,y}$: if a player is in zone (x, y) how often does he choose to pass or dribble as his next action?

Shoot probability $s_{x,y}$: if a player is in zone (x, y) how often does he choose to shoot from that zone?

Note we defined moving or shooting as the only options for a soccer player in this model, therefore $s_{x,y} + m_{x,y} = 1$.

Transition probability $T_{x,y}$: if a player moves (e.g. passing or dribbling) the ball, what is the probability that they will move the ball to each of the other zones?

Goal probability $g_{x,y}$: if a player shoots from zone (x, y), how often does it translate to a goal? This is actually equal to a statistic that is used in soccer more often already, called expected goals (xG). This statistic measures how likely a player is to score from a position.

The problem that we face now, is that we have to assign a value to each zone, representing the direct goal value but also other future rewards it can bring. Using soccer data, the probabilities $s_{x,y}$, $g_{x,y}$ and $m_{x,y}$ can be derived, also the transition matrix T can be determined. In the next part, we are gonna derive the value $V_{x,y}$, which we will assign to each zone and which will represent the payoff at zone (x, y).

If a player chooses to move the ball to a zone (z, w) his payoff will be $V_{z,w}$, but we do not know where the player will move the ball to, and this is where the transition probabilities $T_{x,y}$ come into play. We have $n \cdot m$ different zones where we can move the ball to, staying in the same zone included, therefore

the payoff for moving the ball is:

$$\sum_{z=1}^{n}\sum_{w=1}^{m} V_{z,w} \cdot T_{(x,y)\to(z,w)}$$

Putting this together with the possibility of shooting, and we get:

$$V_{x,y} = (s_{x,y} \cdot g_{x,y}) + (m_{x,y} \cdot \sum_{z=1}^{n}\sum_{w=1}^{m} V_{z,w} \cdot T_{(x,y)\to(z,w)})$$

In words: our value for a zone (x, y) is equal to the percentage of the time a player shoots times the probability that he will score from that position plus the percentage of the time that the the ball is moved times the sum of the payoff over all different zones. Since this captures a level of threat, we can introduce our new variable $\mathbf{xT}_{x,y}$, expected threat. The new formula then becomes:

$$\mathbf{xT}_{x,y} = (s_{x,y} \cdot g_{x,y}) + (m_{x,y} \cdot \sum_{z=1}^{n}\sum_{w=1}^{m} T_{(x,y)\to(z,w)} \cdot \mathbf{xT}_{z,w}) \tag{2.1}$$

We can see that a cyclical dependency follows across the different zones, because the value of one follows from the rest. We can solve this problem by setting $\mathbf{xT}_{x,y}$ equal to 0, and evaluate this formula iteratively. At the first iteration the formula becomes:

$$\mathbf{xT}_{x,y} = (s_{x,y} \cdot g_{x,y}) + (m_{x,y} \cdot \sum_{z=1}^{n}\sum_{w=1}^{m} T_{(x,y)\to(z,w)} \cdot 0)$$

And this results in

$$\mathbf{xT}_{x,y} = (s_{x,y} \cdot g_{x,y})$$

This reminds us again of the xG model, but actually this formula decides how good a position is to shoot from. After the first iteration our variable $\mathbf{xT}_{x,y}$ will not be zero anymore and we can now use (2.1) iteratively. At the first iteration it only calculated the expected threat based on 'shooting', on the second iteration it also considers 'moving then shooting'. On iteration five you consider a maximum of four moves before shooting.

3.4 Translation to expected threat for players

Now we have to remember why we were looking at this approach of expected threat. We wanted to rate soccer players on the basis of their attacking contribution, so now we have to translate this expected threat value to a rating for a player. We can do this by looking at an action made by a player, see where it begins and where it ends. Assume that a soccer player moves the ball from zone (x, y) to (z, w), if we then subtract $\mathbf{xT}_{x,y}$ from $\mathbf{xT}_{z,w}$ we get the value for the ball movement by that player. If we then sum over all the actions made by the player we get their cumulative expected threat level, and we can compare their ratings. This is exactly what Karun Singh has done in his blog with Premier League data of the 2017/2018 season and we see that Kevin de Bruyne tops this list with a mile ahead [7]. This is not a big surprise, but Jose Holebas, a left back from Watford FC, came in third. This is quite unexpected since he is on a shared 51^{st} place when it comes to assists for that season and since he did not score any goals that season [8]. We can conclude that expected threat adds a new point of view relative to the existing statistics.

4 Valuing Actions by Estimating Probabilities (VAEP)

In the next paragraph another method for ranking soccer players is explained. First the existing data format is converted to a new language called SPADL, which is more suitable for machine learning purposes and human interpretability. This .JSON format is the standard for event stream data, an example of this data type can be found in Appendix 0.1. After that the method of estimating probabilities is explained. In the last section both methods are compared and the decision is motivated.

4.1 Soccer Player Language Description Language (SPADL)

The **S**occer **P**layer **A**ction **D**escription **L**anguage or SPADL is designed by researchers from KU Leuven University together with people with domain knowledge [4]. The goal of this language, in contrary to the existing event stream data, is to be simple and better interpretable for humans and computers. The existing data format describes events, whereas SPADL describes actions. There is a python package available that converts the existing event stream data to SPADL. A vector of all actions will be created: $[a_1, a_2, ..., a_m]$, with m the total amount of actions in the soccer game. Each action a_i consists of nine attributes. Each attribute is self explanatory, although a short description per attribute is stated below:

StartTime: the start time of the action
EndTime: the end time of the action
StartLoc: the (x, y) coordinates on the field where the action starts
EndLoc: the (x, y) coordinates on the field where the action ends
Player: the player who performed the action
Team: the team for which the player is playing
ActionType: e.g. shot, pass, dribble
BodyPart: with which body part was the action performed
Result: the result of the action, e.g. succes or fail

In total twenty one different actions have been defined, such as *passes, corners, tackles, penalty shots* and *goal kicks*. Furthermore, we distinguish between four different body parts en six different results of an action.

4.2 Valuing Actions

Valuing **A**ctions by **E**stimating **P**robabilities or VAEP is the name of the method for rating a player by looking at each action and the nine attributes we have defined above. The method is built upon the fact that we can estimate what influence an action has on certain probabilities, namely scoring and conceding in the near future. Suppose we have a game state S_i, which consists of all the actions up to and including time i. Our goal would then be to define the scoring and conceding probabilities for the home and away team for every $i \in [1, m]$, where m is the number of total actions in the game.

$$\Delta P_{scores}(a_i, x) = P_{scores}(S_i, x) - P_{scores}(S_{i-1}, x)$$

We can denote this as the offensive value of an action a_i (and thus changing the game state from S_{i-1} to S_i) by a player from team x, this can be either the home or visiting team. In the same way we can calculate the defensive value of an action.

$$\Delta P_{concedes}(a_i, x) = P_{concedes}(S_i, x) - P_{concedes}(S_{i-1}, x)$$

Valuing an action is then equal to the change in scoring probability minus the difference in conceding probability caused by action a_i, and this is what we will call the VAEP value of an action. Note that if an action increases the scoring probability for a team x, then $\Delta P_{concedes}(a_i, x) < 0$.

$$V(a_i, x) = \Delta P_{scores}(a_i, x) + (-\Delta P_{concedes}(a_i, x)) \tag{4.1}$$

Now that we have a way of valuing actions, we can convert this to player ratings by the following formula:

$$rating(p) = \frac{90}{m} \sum_{a_i \in A_p^T} V(a_i)$$

In this way we can easily determine our time frame (a game, season or a player's whole career), and calculate the rating. Here m is the amount of minutes in that time frame and A_p^T is set of action from player p in time T.

4.3 Estimating Probabilities

The goal of this section is to develop a deeper understanding of how the probabilities and therefore the ratings of the players are determined. We first assume that we have a game state S_i, consisting of a vector of actions $[a_1, ..., a_i]$, our goal is to predict the probability that in the near future a goal will be conceded or scored, for the home team and visiting team. This comes down to estimating the probabilities that

$$P_{scores}(S_i, h) = P(goal(h) \in F_i^k | S_i)$$

$$P_{concedes}(S_i, h) = P(goal(v) \in F_i^k | S_i)$$

$$P_{concedes}(S_i, v) = P(goal(h) \in F_i^k | S_i)$$

$$P_{scores}(S_i, v) = P(goal(v) \in F_i^k | S_i)$$

Where h denotes the home team and v denotes the visiting team and F_i^k is the action vector consisting of the actions $[a_{i+1}, ..., a_{i+k}]$. We can adjust k to our own liking, this the parameter that describes how many steps our model loos back. In Actions Speak Louder Than Goals [4], they chose a k of 10. We only have to estimate two different probabilities, namely $P_{scores}(S_i, h)$ and $P_{concedes}(S_i, h)$.

To estimate these probabilities machine learning is used. These kinds of problems are called probabilistic classification problems, since our goal is to estimate certain probabilities. There are multiple different algorithms you can use, such as CatBoost, Logistic Regression or Random Forest. From [4] it followed that CatBoost was the most precise way, since using this method the probabilities are calibrated in the best way. For the explanation of CatBoost see chapter 2 explanation 1.

In order to make the model understand the game better we have to add some extra features, besides the nine attributes per actions. By a string of actions, we can capture the current speed of play. This takes into account

the distance covered by the action, the angle to the goal, the elapsed time between two actions and whether possession of te ball changed. The second feature we need to add is the game contextual feature. This consists of number of goal scored after action a_i by the team in possession and the defending team, and the goal difference. Teams will play different according to a score, and this feature is designed to capture that information.

4.4 Preferences

Although the methods of expected threat and VAEP have a lot in common, namely valuing on the ball actions for soccer, yet there are quite some differences between the two. First of all, the method of expected threat can only value ball progressing actions. We have in the explanation for expected threat, that it can only value passes and dribbles, while VAEP is not limited to these actions. VAEP can also value tackles and interceptions, which can be also very important.

Second, VAEP has a broader understanding of the game via the extra features added. This understanding is split in game context en action context. For instance VAEP distinguishes between a dribble which leads to a shot or through ball which makes it easier to score for the attacker, we call this action context. VAEP also captures game context, the feature which has been added beside the nine action attributes, which heavily influences the way a team plays.

The last difference is that the method of expected threat is much more robust than that of VAEP. Robustness means that despite fluctuating performances over a season, a few outstanding or bad performances should not influence the rating too heavily. To test this two disjoint sets of the data were created, and the players' ratings were determined. Then the Pearson correlation coefficient was calculated for both methods and expected threat seemed more robust [10].

In conclusion, VAEP captures much more contextual information per action but this is at the expense of being less robust than his counterpart. For the experiment the features are considered a necessary addition, and therefore the method of VAEP is used for the experiment.

5 Experiment

5.1 Introduction

In the early days of soccer the players' roles were very clear; attackers attack and defenders defend. Around the late 1970's many coaches chose the 4-4-2 formation as the teams' tactic [11]. Four defenders, four midfielders and two attackers, that was how the game was played. Soon the players and their coaches stepped out of the traditional roles and fast right and left backs started to join the attack. Soccer tactics became a new line of work with its own hypes. The latest razzmatazz that has soccer tacticians discussing? Inverted wingers.

An inverted winger is an attackers who plays on the side of the field of his weaker foot. This way he can easily cut inside and create danger in different ways. He can directly shoot on goal, give a through pass to the central striker or swing in a cross. Great players and ballon d'or winners are/were inverted wingers, such as Lionel Messi (before he was moved to the center of the field), Johan Cruyff, Robben and Ribéry. These players are exceptions in soccer, but can we generally make a statement about inverted wingers. With the combined method of SPADL and VAEP, we can objectively rate players. Ratings can differ per position on the field, because some players make more actions than others. A midfielder possesses the ball more often than an attacker, but since we are comparing players who play on the same position, this is does not apply to our experiment.

5.2 Methodology

5.2.1 Data

To research our matter of interest we are in need of real-life data, luckily there are few companies specialized in soccer data, such as StatsBomb. There are two different formats of soccer data, namely event stream data and tracking data. Event stream data describes any given event, including the location on the pitch, angle to the goal and the time at which the actions has been made. For research purposes some event stream data has been made available free of charge by StatsBomb, but not in great quantities and only a narrow selection of competitions. Tracking data is a more costly and thoroughly format, since this tracks the location of every player on the pitch at any given moment, but it is therefore expensive and not easy accessible.

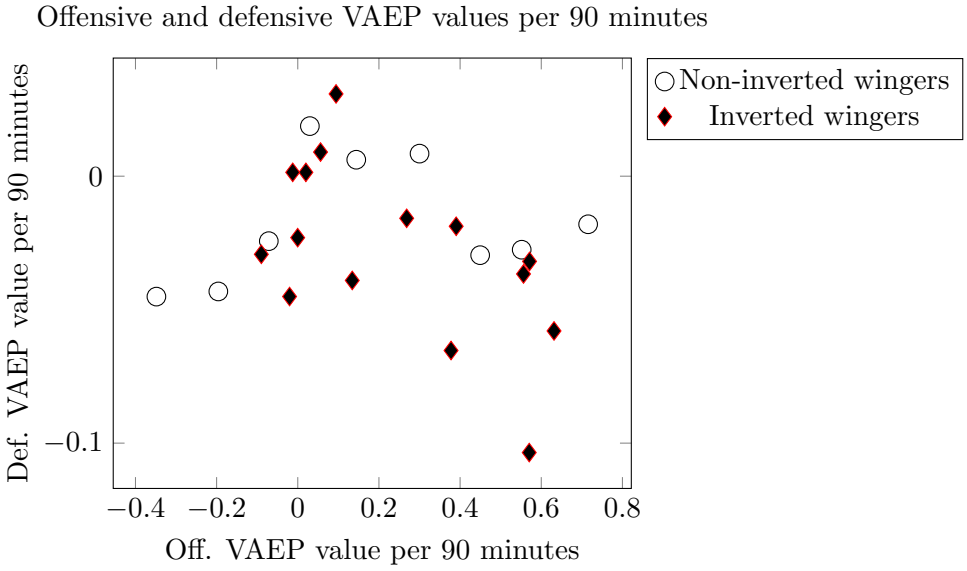
We will be looking at event stream data from the 2018 World Cup, since this is the most general data set that is freely available. This data consists of 64 matches in total, where every match consists of around 1200 actions. Then a list of the wingers who played at the World Cup was made, including their preferred foot and most played position, which can be found in Appendix 0.2. From this we can deduce that there were 55 wingers present at the World Cup, from which 27 are pure inverted.

5.3 Player ratings

For each of these wingers the VAEP value is calculated by using the online Jupyter Notebooks made available by the researchteam from KU Leuven [9]. To provide a clear view, a lower limit of 180 minutes playing time is maintained. From the total of 55 wingers, only 24 played 180 minutes or more, therefore the data set now contains 24 wingers of which 15 are inverted and thus 9 non-inverted wingers.

To determine the ratings, the steps from chapter four are followed. First the event stream data is converted to the SPADL language. Then the features and labels are calculated. For step three, using CatBoost machine learning we determine the scoring and conceding probabilities, and finally these probabilities are used to determine the rating for each player. To get a good insight in het performances per player, we have to adjust for the number a player has played. This results in the VAEP value per 90 minutes, which can be found in Appendix 0.2.

To clearly see the difference in performance between the two kinds of wingers, see the scatterplot below. From equation (4.1) we see that the VAEP value is constructed of two parts, the scoring and conceding probabilities. The offensive VAEP value is equal to the sum of the difference in scoring probabilities over all actions, and the defensive VAEP value is equal to the sum the difference in conceding probabilities over all actions. In the scatter plot the offensive value per 90 minutes is put on the x-axis and the defensive value per 90 minutes on the y-axis, for both hold that higher is better.



Looking at the scatterplot, we can not easily see differences between the two kinds of wingers. The inverted wingers seem a bit more attack oriented, while the non-inverted wingers seem a bit better on the defense. When we look at the average values per 90 minutes, our chimera are confirmed.

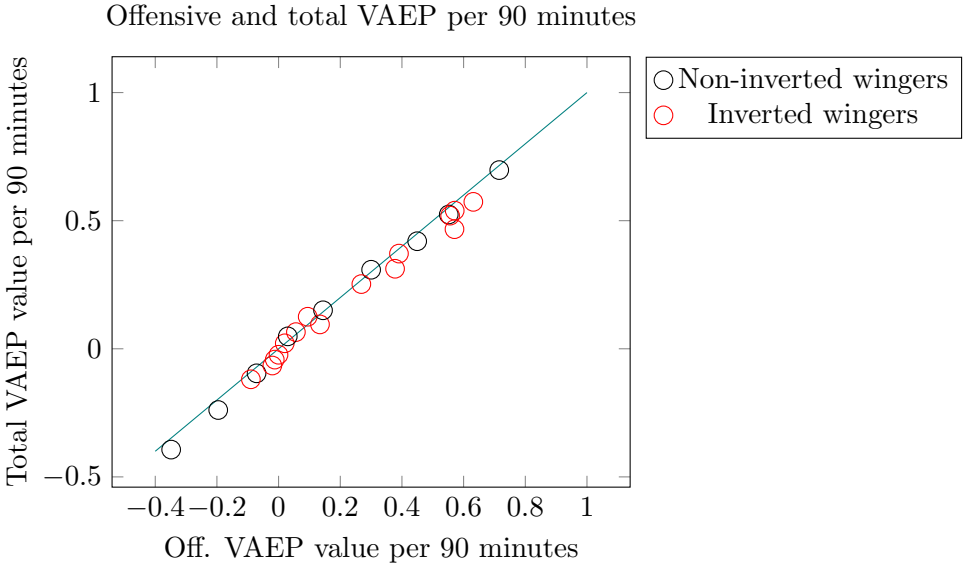
Average VAEP values/90 min.	**Total**	**Offensive**	**Defensive**
Inverted wingers	0,226	0,256	-0,028
Non-inverted wingers	0,158	0,175	-0,017

To motivate what accounts for this difference, the different VAEP values are compared with the goals and assists per 90 minutes during the 2018 World Cup. See the table below.

Goals & assists/90 min.	**Goals**	**Assists**
Inverted wingers	0,270	0,159
Non-inverted wingers	0,287	0,041

Looking at the goals per 90 minutes statistic both kinds do not really get far apart, but the assists present a whole different story. We see that inverted wingers produce more than three times the amount of assists that non inverted wingers do. This could sound a bit strange since the big advantage of inverted wingers is that can easily cut inside and advance to the goal, but it seems that they can attract more opponents and/or find the space for a pass to a teammate.

We also notice that the VAEP value for the wingers is mostly determined by the offensive value, since this is about ten times the size of the defensive part. To support this assumption, notice the scatter plot below, where the offensive VAEP value is put on the x-axis and the total VAEP value is put on the y-axis. The line $y = x$ intersects almost all the points so therefore there is a big dependency.



5.4 Discussion

The findings suggest that inverted wingers are in general more attacking oriented and therefore get a higher overall VAEP rating than the non-inverted wingers. This is mainly due to the amount of assists they have produced. The method of valuing players' actions using VAEP has been used before, but only to create a general ranking of players in the Premier League for the 2018/2019 season [4]. It has not been applied to players on a certain position. The hype for inverted wingers can therefore be explained.

The experiment is eventually conducted on a data set of 24 wingers, of which 15 are inverted and 9 non-inverted. The results could get more weight if the same experiment was conducted on a larger data set, with more wingers and/or matches. Also the fact that the matches were played at the World Cup could have influenced the results, since players play a lot more games at their club in the domestic league than the national team.

Concerning the VAEP method, another feature should be implemented, one that measures if a player performs an action under pressure. To implement this, we do not only need event stream data, but rather tracking data. Tracking data traces every player on the pitch and the ball, and is therefore more extensive and expensive. With this technique we could determine the distance between the player who carries out the action and his opponents nearby. Then pressure can be expressed in a numerical value, which could capture the difficulty of an action even better.

6 Conclusion

In this thesis we took a look at wingers in soccer, especially two kinds: inverted and non-inverted. The goal was to research the difference between the two, using a new developed rating system. The theoretical differences are clear, an inverted winger can easily cut inside and advance to goal, whereas a non-inverted one will cross the ball more often. The experiment showed that inverted wingers are better attacking-wise than their counterpart, but this was mainly due to their higher number of assists, which is surprising. In future research, one could look at the different VAEP values per action to explore in which kind of action the difference in offensive VAEP dwells. If the difference is caused by the difference in number of assists, then it is plausible that inverted wingers have a higher VAEP for passing than the non-inverted wingers. The limitations of my research come down to the fact that data in this field is not as freely available as one would like, therefore the results may be different with a larger data set. This specific research may not be innovating, but it opens doors to objectively looking at soccer performances, rating them and replacing key players by looking at certain characteristics just like *Moneyball*.

7 References

[1] Unknown, (Unknown). *2002 MLB Payrolls*. Retrieved from: http://www.thebaseballcube.com/topics/payrolls/byYear.asp?Y=2002

[2] Andrew J. Schiff, (2008). *"The Father of Baseball": A Biography of Henry Chadwick*. United States: Macfarland and company.

[3] Henry, J. (2006). *Bill James*. Retrieved from: http://content.time.com/time/specials/packages/article/0,28804,1975813_1975844_1976446,00.html

[4] Van Haaren, Decroos, Bransen & Davis (2019). *Actions Speak Louder than Goals: Valuing Player Actions in Soccer*. https://doi.org/10.1145/3292500.3330758

[5] Friedman, J. H. (1999). *Greedy Function Approximation: A Gradient Boosting Machine*. IMS Reitz Lecture 1999. https://doi.org/10.1214/aos/1013203451.

[6] Behrends, E. (2000). *Introduction to Markov Chains.* Advanced Lectures in Mathematics. Vieweg+Teubner Verlag, Wiesbaden. https://doi.org/10.1007/978-3-322-90157-6_1

[7] Singh, K. (2018, 24 december). *Introducing Expected Threat (xT): Modelling team behaviour in possession to gain a deeper understanding of buildup play*. Retrieved from https://karun.in/blog/expected-threat.html

[8] Unknown. (Unknown). *Jose Cholevas: Premier League stats*. https://www.premierleague.com/players/5713/JosC3A9-Holebas/overview

[9] Decroos et al. (2021, 23 may). *Convert soccer event stream data to the SPADL format and value on-the-ball player actions*. Retrieved from

`https://github.com/ML-KULeuven/socceraction`

[10] Van Roy, Robberechts, Decroos & Davis. (2020). *Valuing On-the-Ball Actions in Soccer: A Critical Comparison of xT and VAEP.* Proceedings of the AAAI-20 Workshop on Artifical Intelligence in Team Sports, 1–8. `https://ai-teamsports.weebly.com/`

[11] Wilson, J. (2018). *Inverting the Pyramid: The History of Football Tactics.* Great Brittain: Orion Publishing Co.

Appendices

.1 Example event stream data .JSON format


```
{
  "id" : "643bbb95-9230-4229-97e3-d6a6fa7b5090",
  "index" : 34,
  "period" : 1,
  "timestamp" : "00:00:41.753",
  "minute" : 0,
  "second" : 41,
  "type" : {
    "id" : 30,
    "name" : "Pass"
  },
  "possession" : 3,
  "possession_team" : {
    "id" : 217,
    "name" : "Barcelona"
  },
  "play_pattern" : {
    "id" : 4,
    "name" : "From Throw In"
  },
  "team" : {
    "id" : 217,
    "name" : "Barcelona"
  },
  "player" : {
    "id" : 25854,
    "name" : "Sylvio Mendes Campos Junior"
  },
  "position" : {
    "id" : 6,
    "name" : "Left Back"
  },
  "location" : [ 109.5, 6.3 ],
  "duration" : 1.002633,
  "related_events" : [ "52a98dc5-c393-4ad5-bb85-9c939d787115" ],
```

```
"pass" : {
  "recipient" : {
    "id" : 25879,
    "name" : "Ronaldo de Assis Moreira"
  },
  "length" : 2.9410882,
  "angle" : -1.259798,
  "height" : {
    "id" : 1,
    "name" : "Ground Pass"
  },
  "end_location" : [ 110.4, 3.5 ],
  "body_part" : {
    "id" : 38,
    "name" : "Left Foot"
  }
}
```

.2 Wingers World Cup 2018

Soccer player	Position	Preferred foot	VAEP/90 min
Mohamed Salah	RW	right	0,524267
Kahraba	LW	left	-
Ramadan Sobhi	LW	right	-
Shikabala	RW	right	-
Amr Warda	LW	right	-
Fahad Al-Muwallad	RW	right	-0,393494
Giorgian De Arrascaeta	LW	left	-
Jonathan Urretaviscaya	RW	right	-
Alireza Jahanbakhsh	RW	right	-
Ashkan Dejagah	RW	right	-
Gonçalo Guedes	LW	right	0,021431
Gelson Martins	RW	right	-
Ricardo Quaresma	RW	right	-
Lucas Vázquez	RW	right	-
David Silva	LW	right	-0,042202
Mathew Leckie	RW	right	-0,238956
Robbie Kruse	LW	left	-0,095745
Andrew Nabbout	RW	right	-
Daniel Arzani	LW	right	-
Dimitri Petratos	RW	right	-
Viktor Fischer	LW	right	-
Pione Sisto	LW	right	0,125151
Ousmane Dembélé	RW	right	-
Nabil Fekir	LW/RW	right	-
Florian Thauvin	RW	right	-
Jefferson Farfán	RW	left	-
André Carrillo	RW	right	0,308533
Edison Flores	LW	right	-0,065314
Lionel Messi	RW	left	0,466854
Ivan Perišić	LW	right	0,312392
Ante Rebić	LW	right	0,095169
Marko Pjaca	LW	right	-
Ahmed Musa	LW	left	0,697518

Soccer player	**Position**	**Preferred foot**	**VAEP/90 min**
Victor Moses	RW	right	0,150271
Alex Iwobi	LW	right	-
Douglas Costa	LW	left	-
Neymar	LW	right	0,573503
Taison	LW	right	-
Nemanja Radonjić	LW	right	-
Marco Reus	LW	right	0,371573
Carlos Vela	RW	left	-0,02327
Hirving Lozano	RW	left	0,252594
Son Heung-min	LW	right	0,539201
Kim Shin-wook	RW	right	-
Eden Hazard	LW	right	0,519619
Adnan Januzaj	RW	right	-
Raheem Sterling	LW	right	-0,119094
Ismael Díaz	LW	right	-
Saîf-Eddine Khaoui	RW	right	
Bassem Srarfi	RW	right	-
Naïm Sliti	LW	right	0,06512
José Izquierdo	LW	right	-
Sadio Mané	LW	left	0,419846
Ismaïla Sarr	RW	right	0,048545
Keita Baldé	LW	right/left	-