https://doi.org/10.26219/heal.aueb.5673

SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

Statistical analysis for the player evaluation in team sports

By

Charalampos I. Zafeiropoulos

A THESIS

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
September 2017

https://doi.org/10.26219/heal.aueb.5673

https://doi.org/10.26219/heal.aueb.5673

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

Στατιστική ανάλυση για τις επιδόσεις των αθλητών

Χαράλαμπος Ι. Ζαφειρόπουλος

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Σεπτέμβριος 2017

https://doi.org/10.26219/heal.aueb.5673

https://doi.org/10.26219/heal.aueb.5673

DEDICATION

Dedicated to
my parents, Ioannis and Asimina

https://doi.org/10.26219/heal.aueb.5673

https://doi.org/10.26219/heal.aueb.5673

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Prof. Ioannis Ntzoufras, with whom I had an excellent collaboration. He showed patience and understanding during the development of this master thesis. The guidance and information he provided were both substantial.

https://doi.org/10.26219/heal.aueb.5673

https://doi.org/10.26219/heal.aueb.5673

VITA

I was born in Athens, in 1989. I graduated from high school in 2007. In October of the same year I started my undergraduate studies at the Department of Mathematics of the National and Kapodistrian University of Athens and I obtained my degree in September 2013. In June 2014 I fulfill my military obligation. In October 2015, after I had worked for 1 year, I enrolled in the Full Time Master program of Statistics offered by the Department of Statistics of the Athens University of Economics and Business.

https://doi.org/10.26219/heal.aueb.5673

https://doi.org/10.26219/heal.aueb.5673

ABSTRACT

Charalampos Zafeiropoulos

Statistical Analysis for the Player Evaluation in Team Sports

September 2017

Statistics knows great development in sports from mid 90s and after. Numerous models and indices have been created in order to estimate one match's outcome or the competitive balance of a round-robin tournament. Most of them differ either in the response variable or at the type of the model and the distribution of the response variable. In recent years, statistics also blooms in the field of Training. Statistical models have been constructed to estimate the most desirable level of training load. Professional athletes are also used as sample subjects from many researchers because of their physical condition. Coaches and managers also take into consideration statistics in order to make the best possible additions in their roster or prepare the best possible team for the upcoming season. Similarly, common people also care about statistics in sports either for betting or to verify their favorite team's chances of winning the trophy of major league. In season 2015-2016 Leicester made history when the team was crowned Premier league champion for the first time. This success happened two years after their promotion from English Championship. My aim is to estimate the most important contribution of Leicester's player via statistical models.

Player performance is the main field of sports that my thesis focuses. Most known models and metrics have plus-minus form and they use the number of goals as response variable with big variety of explanatory variables. My approach follows the same direction. However, as explanatory variables I chose only the number of minutes that each player spend in the pitch. My proposal is Poisson generalized linear model. Poisson distribution seems to explain the nature of my response variable goals. As link function I used the logarithm because goals' values meet the requirements of logarithm.

First Player performance estimation took place through descriptives and model for each player. Estimations for the attacking performance were different for each methods. On the other hand, estimations for the defensive ability were alike in both methods. Plots also shown that, Ulloa was more profitable for his team than Okazaki who is main opponent for the position of Vardy's partner. Mahrez also had decent attacking and defensive ability. As for Teams, Leicester and Tottenham were 2 out 3 best attacking and defensive teams. Manchester

https://doi.org/10.26219/heal.aueb.5673

City was the best attacking team and Manchester United was the best defensive team. Leicester also shows bigger variance than the other teams due to the fact that its' player was also included in the model. Championship table confirmed my results about teams.

Last I divide each game into shift according to events such as substitutions and red cards. Each player that played on each shift, had as count the number of minutes of the shift divided by 90. I presented a model for all players. I faced the problem of multi-collinearity which led to some extreme results but the model fitted with the saturated. Results shown, that the starting team was more effective in defense with the exception of defenders. However, substitutions seem to be more productive. Last, the results about Teams' brought the same conclusion with the previous models.

https://doi.org/10.26219/heal.aueb.5673

Περίληψη

Χαράλαμπος Ζαφειρόπουλος

Στατιστική Ανάλυση για τις Επιδόσεις των Αθλητών

Σεπτέμβριος 2017

Η Στατιστική των αθλημάτων γνώρισε μεγάλη ανάπτυξη από τα μέσα της δεκαετίας του 90 και μετά. Πολλές έρευνες έχουν πραγματοποιηθεί με στόχο να εκτιμηθούν οι παράγοντες που επηρεάζουν την έκβαση ενός αγώνα. Πολλοί δείκτες έχουν δημιουργηθεί για να εκτιμήσουν το επίπεδο ανταγωνιστικότητας ενός πρωταθλήματος. Μοντέλα έχουν δημιουργηθεί αναφορικά με το επίπεδο πίεση στην προπόνηση των ομάδων. Επίσης πολλές έρευνες έχουν πραγματοποιηθεί σε δείγματα αθλητών λόγω της πολύ καλής φυσικής τους κατάστασης. Προπονητές και τεχνικοί διευθυντές λαμβάνουν υπόψη αυτές τις μετρήσεις για την απόκτηση παιχτών που θα ανεβάσουν το επίπεδο της ομάδας αλλά και για την καλύτερη δυνατή προετοιμασία για την αγωνιστική χρονιά. Ομοίως πολλοί άνθρωποι λαμβάνουν υπόψη αυτές τις έρευνες κυρίως για λόγους στοιχήματος αλλά και για να εκτιμήσουν την πορεία της αγαπημένης τους ομάδας στο πρωτάθλημα που αγωνίζεται. Την αγωνιστική χρονιά 2015-2016 η Λεστερ πραγματοποίησε μια από τις μεγαλύτερες εκπλήξεις της σύγχρονης ιστορίας, καθώς ανακηρύχθηκε πρωταθλήτρια Αγγλίας για πρώτη φορά στην ιστορία της. Αυτή η επιτυχία μοιάζει ακόμα μεγαλύτερη αν αναλογιστεί κανείς πως 2 χρόνια πριν είχε ανέβει στην πρώτη κατηγορία. Ο στόχος της εργασίας είναι να εκτιμήσουμε την επίδραση των πιο σημαντικών παιχτών μέσω ενός μοντέλου.

Η απόδοση των παικτών είναι ένα πεδίο των αθλημάτων όπου η στατιστική γνωρίζει μεγάλη άνθιση. Οι πιο γνωστοί δείκτες έχουν την τάση να εξισώσουν τα θετικά και τα αρνητικά για να εκτιμούν την απόδοση ενός παίχτη. Τα περισσότερα μοντέλα τείνουν να χρησιμοποιούν τα γκολ σαν εξαρτημένη μεταβλητή για να εκτιμήσουν την απόδοση ενός παίκτη χρησιμοποιώντας μια μεγάλη ποικιλία ανεξάρτητων μεταβλητών. Αντίστοιχη ήταν και η προσέγγιση που ακολουθήσαμε. Παρόλα αυτά σαν ανεξάρτητη μεταβλητή χρησιμοποίησα μόνο τους χρόνους που κάθε παίχτης έπαιξε στο εκάστοτε αγώνα, Θεωρήσαμε πως τα Poisson γενικευμένα γραμμικά μοντέλα ήταν τα κατάλληλα για να κάνουν την εκτίμηση. Η Poisson κατανομή επιλέχθηκε γιατί έδειχνε να συνάδει με την φύση της εξαρτημένης μεταβλητής. Σαν συνάρτηση σύνδεσης επιλέχθηκε ο λογάριθμος. Οι τιμές της εξαρτημένης μεταβλητής τηρούν τις προϋποθέσεις του λογαρίθμου. Αρχικά η εκτίμηση της απόδοσης των

https://doi.org/10.26219/heal.aueb.5673

παιχτών πραγματοποιήθηκε μέσω της περιγραφικής ανάλυσης και μέσω μοντέλων για κάθε παίχτη. Οι εκτιμήσεις για τις επιθετικές επιδόσεις ήταν διαφορετικές σε κάθε μέθοδο. Ενώ στην εκτίμηση της αμυντικής επίδρασης η κατάταξη των παικτών ήταν σχεδόν παρόμοια. Ο Ulloa ήταν πιο παραγωγικός από τον βασικό του ανταγωνιστή, για την θέση του συμπαίκτη του Vardy, Okazaki. Ο Mahrez επίσης ήταν το ίδιο καλός και σε επίθεση και σε άμυνα. Όσον αφορά τις ομάδες η Leicester και η Tottenham ήταν δύο από τις 3 καλύτερες ομάδες σε επίθεση και άμυνα. Η Manchester City είναι η πιο παραγωγική ομάδα ενώ η Manchester United καλή αμυντικά ομάδα. Ο βαθμολογικός πίνακας επιβεβαίωσε τα αποτελέσματά μου.

Στην συνέχεια τα δεδομένα χωρίστηκαν σε περιόδους ανάλογα με γεγονότα όπως τις αλλαγές και τις κόκκινες κάρτες. Κάθε παίκτης που έπαιξε στην περίοδο είχε σαν συντελεστή τον αριθμό των λεπτών που έπαιξε διαιρεμένο με 90. Μετά παρουσίασα ένα μοντέλο για όλους τους παίκτες. Το μοντέλο παρουσίασε το πρόβλημα της πολυσυγγραμικότητας το οποίο οδήγησε σε αρκετά ακραία αποτελέσματα. Τα αποτελέσματα έδειξαν πως ,με εξαίρεση τους αμυντικούς, η αρχική ομάδα ήταν πολύ αποτελεσματική στην άμυνα. Παρόλα αυτά, οι αλλαγές ήταν πιο παραγωγικές στην επίθεση. Τέλος, τα συμπεράσματα για τις ομάδες ήταν ίδια με τα προηγούμενα μοντέλα.

https://doi.org/10.26219/heal.aueb.5673

Table of Contents

https://doi.org/10.26219/heal.aueb.5673

https://doi.org/10.26219/heal.aueb.5673

https://doi.org/10.26219/heal.aueb.5673

List of Tables

https://doi.org/10.26219/heal.aueb.5673

https://doi.org/10.26219/heal.aueb.5673

List of Figures

https://doi.org/10.26219/heal.aueb.5673

https://doi.org/10.26219/heal.aueb.5673

https://doi.org/10.26219/heal.aueb.5673

1. Introduction

1.1. Purpose of Thesis

Sport is a physical activity that aims to improve the physical ability of the participants while providing joy. Sport also provides entertainment to spectators. Since old times, sports were influenced by politics. For instance, Hitler used the summer Olympics at 1936 that were hosted in Berlin to promote the Nazi ideology. Nowadays, sports know great growth in both reputation and revenue. Many companies use sports to advertise their products. Also, a lot of people bet on sports outcome. To sum up, sports industry affects a big part of our population.

Statistics is a science that collects, presents, and analyzes data via statistical methods such as descriptive statistics and inferential statistics. In descriptive analysis we use plots and numeric values while in inferential analysis we make hypothesis tests for the population based on data from a sample. It was a matter of time before sports and statistics began to be related. Numerous studies have been made in order to estimate some sports factors such as prediction, player performance, sports economics, training etc. Each study may differ in the dependent variable, explanatory variables or the type of model that it is used. In my thesis, I try to estimate player performance via statistical models by using only player's participation times. I propose Poisson generalized linear model with the number of goals as dependent variable. My data refers to Leicester's championship course in Barclay's Premier league for period 2015-2016.

1.2. Structure of Thesis

My thesis is divided in three chapters. First chapter refers to statistical methods and models that were used in other papers with regard to sports and was divided in four sections. First section was about the match's outcome. Second section was about sports economics and league's competitive balance metrics. Third section was about player performance estimations. The last section was about training methods since professional athletes were used as sample due to their physical condition. Second chapter refers to a detailed presentation of player's performance indices. In this chapter I also presented a paper that used player performance indices in order to estimate each player's worth in relation to his weekly wage. Third chapter refers to the statistical model I finally concluded. This chapter was divided in three sections. First section refers to the descriptive analysis I performed at the dataset. I

https://doi.org/10.26219/heal.aueb.5673

identified that Leicester's coach used close rotation in goalkeeper's and central defender's position, while the rotation in midfielders and forward was pretty wider. I also presented correlation coefficients between Leicester's players, where the values of right defenders were greater than the rest of the team. At the second section I presented the reasons that led me to choose Poisson as my response variable's distribution via plots and tests. Last section includes player's impact through descriptives and model. First I calculate the number of goals that Leicester scored and conceded when the player was in the pitch as attacking ability and defensive ability, respectively. Then I proposed 22 models for each player. Comparisons between the values shown, that descriptives overestimate the attacking performance of player that participated in few matches. On the other hand, model underestimate players with many matches. In defense, estimations are pretty equal in both methods. Last I presented the attacking and defensive ability for each team via errorbars. Leicester, Tottenham and Manchester City were most productive teams while Aston Villa, Norwich and Bournemouth were the worst attacking teams. Furthermore, Leicester, Tottenham, and Manchester United were the best defensive teams. On the other hand, Aston Villa, Norwich and Crystal Palace were the worst defensive teams. Championship table confirmed my findings.

In the Fifth I divide each game into shift according to events such as substitutions and red cards. Each player that played on each shift, had as count the number of minutes of the shift divided by 90. This way my dataset had further information about the impact of each player in Leicester's goals. I presented a model for all players where I used sum to zero method in order to obtain coefficient for all Teams. Then, I checked model requirements using plots and tests. Players' results shown that starting team was good in defense but not that good in attack. Last, the results about Teams' impact on the number of goals were similar with the results of the previous models.

https://doi.org/10.26219/heal.aueb.5673

2. Football Analytics and Prediction

2.1. Introduction

In this chapter I am going to present statistical methods related with sports. I divide this chapter in 4 sections. First section refers to prediction, which is important especially to bookmakers. Second section refers to sports economics and competitive balance, which are important specifically to each league. Third section refers to player performance indices, which are important especially to managers and fans. Last section refers to training which is important specifically to coaches. With the exception of player performance, the rest sections refer to football which is the most known sport all over the world.

2.2. Prediction

Prediction is an important field in statistics of sports, which knows big development through years. Betting is the main reason of that development. Determining sports betting is a difficult task. However, betting in sports exists for thousand years. From ancient Rome where emperors used to bet on circus or chariot races to nowadays, where many people bet on the outcome of a match. Moreover betting exists in almost all sports that are known.

Statistical football prediction is a method which is used to predict the outcome of a match along with the use of statistical tools. The goal is to outperform the predictions of bookmakers that set the odds on the outcome of the match. The most widely used statistical approach to prediction is ranking. Football ranking systems assign a rank to each team based on their past game results, so that the highest rank is assigned to the strongest team. The outcome of the match can be predicted by comparing the opponent's ranks. Several different football ranking systems exist, for example, FIFA is widely known world ranking or the World football Elo ratings.

2.2.1. Knockout Competition

All prediction methods could be categorized according to tournament type, time dependence and regression algorithm. The tournament types vary between knockout competition and Round-robin tournament. The methods of knockout competition are summarized in an article by Diego Kuonen (1996). At this article, he claimed that logistic models are better in

https://doi.org/10.26219/heal.aueb.5673

prediction than others are. He also used the assumption of independency between matches and his goal was to determine the strength of every team using Bernoulli distribution.

2.2.2. Round-Robin Tournament

However, more research has been made in Round-robin tournament methods and especially in time dependent Poisson regression models. According to Maher (1982), they used two independent Poisson distributions (the first one for the goals of the home team and the second one for the goals of the away team) with different means. Thus, the joint probability of the home team scoring x and the away team scoring y was the product of the two independent probabilities.

However, the development of betting made this model less effective. Therefore, Lee (1997) took Mayer's assumptions about Poisson but he used generalized linear models to estimate the mean score of home team and the mean score of the away team. The link function that he used was the logarithm. Concerning the parameters, he proposed using a constant in both models, which measured the average score in a game. In addition, he proposed three more parameters in order to estimate the mean score of the home team. The first one measured the home-team advantage, the second one measured the offensive power of the home team, and the third one measured the defensive power of the away team. Concerning the model that was used in order to estimate the mean score of the away team, he proposed two more parameters except of the constant. The first one was used to measure the offensive power of the away team, and the second one was used to measure the defensive power of the home team. In order to estimate these parameters, he used maximum likelihood method.

During the time, Lee (1997) made the previous model Dixon and Cole (1997) also created a new type of model which goal was to make better prediction than Maher's one. In particular, they accepted Mayer's assumptions about independent Poisson distributions for the mean score of the home team and the mean score of the away team. Furthermore, they claimed that the mean score of the home team was consisted to a parameter for the home advantage, one parameter for the home team offensive power, and one parameter for the away team defensive power. Concerning the mean score of the away team, they used one parameter for the away team offensive power and one parameter for the home team defensive power. In addition to this, Dixon and Cole (1997) realized that these models were static. This means that they did

https://doi.org/10.26219/heal.aueb.5673

not take into consideration the time the match took place. In order to face this difficulty, they used one index in the likelihood, which was affected by the time the match took place.

A very important factor that previous models did not take into consideration was the covariance between the two opposing teams. All previous models advocated that the expected number of goals of the home and the away team followed independent Poisson distribution. Therefore, these models made poor prediction as for draws and low scores. One model that was used to face that problem was introduced by Karlis and Ntzoufras (2003). At this model, they presented the Bivariate Poisson for X and Y, where X denoted the number of goals for the home team with mean $\lambda_1 + \lambda_3$ and Y denoted the number of goals for the away team with mean $\lambda_2 + \lambda_3$. The parameter λ_3 is the one that took into consideration the correlation between the two opposing teams. If $\lambda_3 = 0$, then X and Y follows independent Poisson distribution. In particular, λ_3 denoted these factors that influenced both home and away team. Factors like weather condition or field condition or time that the match took place. For the i^{th} observation this model takes the form:

$$(\boldsymbol{X_i}, \boldsymbol{Y_i}) \sim \boldsymbol{BP}(\boldsymbol{\lambda_{1i}}, \boldsymbol{\lambda_{2i}}, \boldsymbol{\lambda_{3i}})$$

and

$$\mathbf{log}(\boldsymbol{\lambda_{\kappa i}}) = \boldsymbol{w_{\kappa i}}\boldsymbol{\beta_\kappa},$$

with $\kappa = 1, 2, 3$ and $i = 1, \ldots, n$ which denotes the observation number, $\boldsymbol{w}_{\kappa i}$ denotes a vector of explanatory variables for the i^{th} observation, $\lambda_{\kappa i}$ and $\boldsymbol{\beta}_\kappa$ denotes the corresponding vector of regression coefficients. However, the number of explanatory variables are not the same. Therefore, the estimation of regression coefficients are not straightforward. In order to estimate then we used the EM algorithm.

Despite the fact that Poisson distribution is the most widely known for prediction in sports, its use is questionable because of the overdispersion that is appeared (sample variance exceeds sample mean) in certain leagues. In addition, many times correlation between goals is low.

In order to eliminate correlation, Ntzoufras and Karlis (2007) introduced a new method with the difference of home and away goals. This method can conclude to good prediction about the result, but not for the exact score. Respecting the assumptions of their previous model they resulted in Skellam distribution. Moreover, this model does not assume that the goals scored by each team are marginally Poisson distributed. Finally yet importantly, an advantage of this model is that it can predict the outcome of the game as well as for betting purposes related to the Asian handicap. However, the model cannot predict the final score as we

https://doi.org/10.26219/heal.aueb.5673

previously mentioned. Afterwards, they proceeded by using Bayesian approach concerning estimation of model's parameters.

The Bayesian approach is suitable for modelling sports outcomes in general, since allows to the analyst to use any further knowledge about each game via prior distribution. Further information that could be used in the model should be based on historical knowledge or data such as weather conditions or fitness of the team. Finally, the Bayesian approach allow to us predicting future games via the predictive distribution. Predictive distribution could be also used to produce a posterior predictive distribution for future scores, outcomes or even reproduce the whole tournament and produce quantitative measures concerning the performance of each team. When no information is available, they proposed to use normal prior distributions for the parameters of the Poisson Difference (i.e. Skellam) model with mean equal to zero and large variance (e.g. 104) to express prior ignorance. They also used Markov Chain Monte Carlo (MCMC) algorithms to generate values from the posterior distribution and hence estimate the posterior distribution of interest and their corresponding measures of fit. This model was used in Premier League of 2006-2007 with fairly good results.

Another model that used the same logic was proposed by Shahtahmassebi and Moyeed (2016), where they used Generalized Poisson Difference Distribution (GPDD). Its advantage over Poisson difference model is the flexibility in tails, considering them shorter as well as longer than Poisson difference. Like in Poisson difference model, they used Bayesian approach. In particular, they included information from previous year's tournament as prior. The previous season was considered baseline for the following season's results, as the teams in general intend to keep their current position at the table. Concerning home and team ability, they used as prior $Normal\ (\mu, \sigma^2)$ where μ and σ^2 were equal to the mean and variance of the goal difference of a specific team at the previous year. For teams who were promoted and were playing for the first time, they assigned non-informative normal distribution with mean zero and large variance (i.e. 10^4) because, as they claimed, team abilities by definition were measured in relation to other teams in the same league. Therefore, using this approach, they modeled both mean and variance. As for the positive constant for the variance, they used non-informative gamma distribution. As long as for the posterior distribution, it is known for normalizing the constant and it is not analytically tractable. In order to generate samples from the posterior distribution, they used Markov Chain Monte Carlo method and more specifically

https://doi.org/10.26219/heal.aueb.5673

the random walk Metropolis-Hastings (M-H) algorithm. This model was fitted to the Italian Serie A 2012-2013 with fairly good predictions.

2.3. Sports Economics

Sports economics is another important field in Statistics of Sport, which also knows huge growth through years. At the present thesis, I am going to talk about football. Football is the most popular team sport in the world. Moreover, European professional leagues show a considerable growth in both revenue and spectators. Despite that growth, there are some important issues that sports league have to take into consideration in order to ensure their long-term success. One of those issues is competitive balance, which represents the uncertainty of the outcome. Many sports economists have tried to create indices which measures competitive balance. However, these indices' effectiveness is a matter of debate.

In order to acquaint their approach, the researchers mentioned at the subsequent sections, suggested three essential components. First, they calculated the winning percentages of each team in the league. Secondly, they took into consideration the case of perfectly balanced league where all teams had the same percent of win record, and the case of total unbalanced league where the stronger team wins the weaker team. Finally, they normalized these indices in order to make them comparable among leagues. The values range from one to zero, which refer to total imbalanced league and perfect balanced league, accordingly.

I am going to present the most important indexes in what follows.

2.3.1. National Measure of Seasonal Imbalanced

The most known index called National Measure of Seasonal Imbalanced ($NAMSI$), was first mention by Goosens (2006). $NAMSI$ is defined as the ration of the observed deviation of teams' winning percentages (STD) and the standard deviation in a total unbalanced league (WSD):

$$NAMSI = \frac{STD}{WSD}$$

2.3.2. Herfindahl-Hirschman Index

Another index that measures the competitive balance in a league is the Herfindahl-Hirschman Index (HHI) presented by Depken (1999). In particular, this index measures the market shares

https://doi.org/10.26219/heal.aueb.5673

of all teams, which in professional sports are the winning shares in the league. This index is defined as the quadratic summation of the market shares of all team:

$$HHI = \sum_{i=1}^{N} (MS_i)^2$$

Owens et al. (2007) introduced a normalized version of this index, which took into consideration both lower and upper bounds of HHI:

$$HHI^* = \frac{HHI - HHI_{ideal}}{HHI_{ub} - HHI_{ideal}}$$

where HHI_{ub} is the HHI index for total unbalanced league and HHI_{ideal} is the HHI index for a perfect balanced league.

As I mentioned above, there are many indices that quantify competitive balance. On the other hand, a few of them are properly designed in order to quantify the competitive balance for the more complex structured European leagues. For instance, $NAMSI$ treat teams equally at the top and at the bottom of the ladder, while HHI rates teams according to their winning share. Therefore, the design of special indices using a suitable weighting pattern is required when measuring competitive balance in European football.

2.3.3. Championship levels and indices between them

Manasis et al (2012) presented a helpful approach in order to make domestic European leagues less complex. In particular, European championships were divided in three levels in which teams compete for the corresponding ordered set of prizes or punishments. Subsequently:

a) The first level refers to the competition for the championship title, which is considered the most important prize in any league. For some countries, the champion team has more chances of competing in next year's most prestigious European tournament, Champions league than the rest league teams. Champions League offers reputation and high prizes for both participation and successful results.
b) The second level refers to the remaining qualifying places for European tournaments of the following season. Except for Champions League, there is a recently restructured new European tournament called Europa League. Europa League also offers decent prizes for both participation and successful results. Therefore, over and above the

https://doi.org/10.26219/heal.aueb.5673

championship title, teams also compete for any of the remaining pre-determined top places.

c) Finally, the third level refers to the relegation places. European leagues are open tournaments. This means that due to their poor performance, teams that occupy the last league positions are relegated to lower leagues (divisions). Such relegation can hurt team's financial status and reputation. Therefore, some teams have as their main goal to avoid relegation.

One of this championship's goals is to attract fans into attending or watching as many matches as possible. However, in many European's leagues a small number of teams dominate. In this kind of structure, if one team dominates the championship run (level one) it is more likely for the remaining levels to be more competitive. This means that, the overall competitive balance is affected by the degree of competitiveness in each level.

In the same paper, they also tried to calculate the most known indices that measure competitive balance using the approach mentioned previously. Overall competitive balanced is determined by the corresponding degrees at the three levels. They also took into consideration that competition for the title is more attractive than the battle in order to avoid relegation. Therefore, they weighed the levels depending on the importance that fans show. In particular, the first level was weighted higher than the other two levels. In the second level, which included the qualifying places, the weights were decreasing as the ranking was increasing. Finally yet importantly, in the third level, which included the fight to avoid relegation, indices received lower weights than the corresponding ones for the qualifying places and higher than the corresponding one for the positions in the middle of the league. To sum up, they need to quantify the competitiveness at each level separately and weight each ranking according to their importance. Therefore Manasis et al (2011), having as target the construction of these indices, they adopted the Normalized Concentration Ratio (NCR). This index is used in order to measure the strength of the top K teams in comparison to the remaining ones. NCR_K index is defined as:

$$NCR_K = \frac{\sum_{i=1}^{K} P_i - 2K(N-1)}{2K(N-K)} = \frac{1}{2(N-K)}\left(\frac{1}{K}\sum_{i=1}^{K} P_i\right) - \frac{N-1}{N-K}$$

where N is the total number of teams participating in the league, K is the total number of top league positions considered for this index and P_i is the number of points collected by the i^{th} ranked team. This index effectively measures the competitiveness for the first level.

https://doi.org/10.26219/heal.aueb.5673

Concerning the second level, the design of an index is more complicated since the performance of each team depends on the champion's performance. Therefore, they introduced the Adjusted Concentration Ration (ACR), which was created under the assumption that higher ranking places are more important for fans, thus ranking places must be weighted accordingly. ACR ranges from zero to one, with zero denoting the absence of domination; it means that competitiveness between top K teams is high. One disadvantage of this index is that it does not provide information for the competition introduced by teams after K_{th} position. ACR_K is defined as:

$$ACR_K = \frac{\sum_{i=1}^{K} NCR_i}{K}$$

Concerning the third level, they introduced an index for the third level, which is also important in European league where the promotion-relegation rule exists. Therefore, they presented the Normalized Concentration Ratio for Relegated (NCR^I), which measures the degree of weakness of the I relegated teams when we compare them with the remaining ones. NCR ranges from zero to one, with zero denoting that the I teams are strong enough to collect the maximum attainable points. In that case, the league is perfectly balanced.

Finally, they introduced the Special Concentration Ratio (SCR), which captures all of the three levels mentioned above, embodied in the European multi-prized leagues. SCR rates levels and ranking positions are conducted according to their significance and could be considered as a custom build index, which ca could n be adapted according to the particular interest generated by a league. For the development of this index, they used the previously mentioned (ACR and NCR) indices in order to capture the first, second and third level. SCR_K^I is defined as:

$$SCR_K^I = \frac{\sum_{i=1}^{K} NCR_i + NCR^I}{K+1}$$

Manasis et al (2010) also presented a dynamic version of $NAMSI$ and HHI indices. Despite the fact that in ''static'' approach is simple calculated the ''dynamic'' approach, it is more complicated, because it should include both the case of perfectly balanced and the case of completely unbalanced league. In static approach, the calculation is simple by allocating each team 50 percent of the points or wins for every round. However, in dynamic approach its value depends on the order of competing teams. For instance, the second strongest can lose only if it plays with the best team without knowing in which round this match will take place.

https://doi.org/10.26219/heal.aueb.5673

In dynamic approach, they computed the previous indices for each round. Those indices are refered to seasonal information since all rounds are included in the measurement.

As an application of the above, they tested those indices in Greek and German league. One problem was the high degree of noise that appeared in early season because teams' winning percentages were influenced by the opponents they had faced so far. Last but not least, according to the round-based indices competitions, they were less balanced than the conventional indices would suggest. This result could be explained by the fact that a stronger team employ most of its resources at the beginning of the season to gain competitive advantage and after securing its position, it tends to slow down towards the end of a season.

2.4. Player Performance

Player performance is another very important field in Statistics of Sports, which also knows great development through years. In team sports, player performance metrics help coaches to value their players. In addition, team's manager uses these metrics in order to estimate how beneficial a wanted player can be for their team. Particularly, low budget teams try to find player that are undervalued by their team and take them as a transfer. This makes them more competitive against wealthier teams. Bookmakers also use player performance indices in order to determine the odds on bets. People usually bet on match result, tournament result, and individual awards such as Player of the Year etc. Fans also use player performance metrics to identify how well their favorite player performs. In addition, many analysts use player performance indices as explanatory variable in order to estimate other variables such as transfer fees, team's chances of winning the championship, etc.

Some individual sports such as tennis, use player performance metrics to rank all the individuals in a list, according to their performance in the last 52 weeks. If a player plays well enough to make it at the top 104 in the world, the player will be able to gain direct entry into the Grand Slam tournament. This is how most players do it and is the best way to guarantee a spot in the main draw. Sixteen tickets go to individuals that won the qualifying tournament that takes place before the Grand slam and eight tickers go to organizers' choices. However, our main interest is on team sports player performance indices. More details about these indices are presented at the next chapter.

In what follows, I am going to categorize these models according to each sport.

https://doi.org/10.26219/heal.aueb.5673

2.4.1. Soccer

At soccer, most indices use goals to estimate how well a player perform since goals are responsible for the outcome. Most known index is the EA Sports Player Performance Index, which is used at the top 2 tiers of English football the Barclay's Premier League and Championship. This index is divided in 6 subindices. The first subindex captures the in-match contributions. It is calculated using a model with goals as response variable and many explanatory variables for both attacking and defensive team. The second subindex represents the contribution of each player on the points that each team win on each match according to the number of minutes he spend in the field. The third subindex also represents the performance of losing team's players according to the number of minutes they spend in the field. The fourth subindex represents the number of points that every goal gives to the team that scored. The fifth subindex represents the number of points that every assist gives to the team that scored. The sixth subindex represents the number of points that every clean sheet gives. The final index is a weighted combination of all subindices.

McHale and Szczepanski (2014) presented another approach in estimating player's performance. This approach was based on Bayes theorem. Instead of modeling the number of goals, they tried to estimate first the number of shots via a Poisson mixed effect model. As explanatory variables they used both team and player information. Then, they proposed a Binomial mixed effects model where the response variable was the number of goals given the number of shot (shot accuracy). As explanatory variables, they used information about the player's ability to convert shots into goals. Model selection was carried out via Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

2.4.2. Basketball

At basketball, most known metrics are plus-minus indices that estimate a basketball's player performance based on box score statistics such as points, rebounds, etc. However, these metrics mislead us since they do not take into consideration the situation under which players perform. Moreover, plus-minus metrics depend on the contributions of teammates and opponents. Deshpande and Jensen (2006) proposed a Bayesian linear regression model in order to construct an index based on team's winning chances instead of points. They also divided the game into shifts to face the problem of plus-minus indices that was mentioned above. Player performance was estimated in each shift. They used indicators for home team's

https://doi.org/10.26219/heal.aueb.5673

players and away team's players as explanatory variables. In order to fit their model, they considered empirical estimates of win probability given the number of points that home team was in front at the start of each shift. Variable selection was carried out via Bayesian approach. In particular, they proposed Laplace centered at zero prior distribution on player partial effect. This approach uses a penalty term, which pulled the small partial effects of the player towards zero. This method is equivalent to Lasso shrinkage method. Moreover, this method helped them do comparisons between players via posterior samples that obtained with the method of Gibbs sampling. More details are presenting in the next chapter.

2.4.3. Hockey

The approach that was followed in hockey's metrics was similar to basketball's approach. However, measuring player's performance in hockey is more difficult due to big number of players, frequent changes in lines and rareness of goals. Many indices have the form of plus-minus. They count the number of goals that scored by the player's team minus the number of goals that the opposing team scored while he was on ice. Other indices use further information, but some analysts do not think that is needed. Gramacy et al (2013) proposed a logistic regression model in order to estimate the partial effect of each player. They used the logarithm of odds that the goal was scored by the home team as response variable. As explanatory variables they proposed indicators for the opposing teams and for the players that were on ice during the goal. The variable selection was also carried out via Bayesian approach. They proposed a Laplace centered at zero prior distribution on player's partial effect to pull the small towards zero. They also proposed a Normal centered at zero prior distribution, which uses a penalty term. This approach equals with ridge shrinkage method. Similar to the basketball approach, they tried to make comparisons between players. Furthermore, they obtained data about player's salary and they identified if highly paid players worth their money.

2.5. Training

Training is another important field of sports. Training and performance are highly related. Therefore, coaches try to determine the best possible training load in order to improve their players' performance. Studies have described the training-performance relationship with the dose-response relationship reported in pharmacological studies. Coaches' goal through training is to maximize performance potential and minimize negative consequences such as

https://doi.org/10.26219/heal.aueb.5673

injury, illness, fatigue, etc. Many researchers have tried to find the best training program with the use of statistical methods.

I am going to divide these researches in two categories; Individual Sports and Team Sports.

2.5.1. Individual Sports

Banister et al (1999) tried to estimate if different formats of taper could benefit physical performance more than the standardized period of taper. As they mentioned, taper is an effective way to ensure recovery from heavy training and is generally accepted as an integral part of optimal preparation for competition. They simulated four different taper profiles in a model to predict performance using standard square-wave quantity of training for 28 days. These profiles were:

- Step reduction versus exponential decay
- Fast versus slow exponential decay

They also determined physical tests in order to identify the effectiveness that each profile had in the physical improvement. These tests were taken above pre-taper criterion standards, included running and cycle ergometry. This simulation showed that exponential taper was better than step reduction. Furthermore, fast decay concluded in better results than slow decay. Then, they tested these results in field trials to identify the correspondence between simulation and real-training criterion physical tests in triathlon athletes. The results showed that exponential taper group made a significantly greater improvement than step reduction taper group in cycle ergometry. Exponential taper group was also better than step reduction group in 5-km run but not significantly. Fast exponential taper group was significantly better than slow exponential taper group in cycle ergometry. Last but not least, fast exponential taper group was also better than slow exponential taper group but not significantly.

Avalos et al (2003) tried to estimate the relationship between training and performance in 13 competitive swimmers over 3 seasons and how swimmers responded individual and as a group to training. They used linear mixed effects models. The effect of training was studied over three periods: short-term, which included the average of training load during the first two week before each performance; mid-term which included the average of training load during weeks 3,4,5 before each performance; and long-term which includes weeks 6,7 and 8. Using cluster analysis, they identified four groups according to their reaction to training. The first group was about swimmers who responded well to long-term training period. The second

https://doi.org/10.26219/heal.aueb.5673

group was about swimmers who responded well to long-terms and mid-terms periods of training. The third group was about swimmers that responded well to short-term and mid-term period of training. Finally, the forth group was about swimmers who responded well to combined periods.

The differences between swimmers and their improvement over the three seasons were statistically significant for the identified group of swimmer. Short-term training had negative influence on performance in the four groups. On the other hand, mid-term and long-term training had a positive influence in the three out of groups. To sum up, the mixed model described a significant relationship between training and performance both for individuals and for group of swimmers.

2.5.2. Team Sports

In contrast to individual sports, team sports are more physical demanding due to the big number of physical collisions such as tackles, change of direction, rapid accelerations, decelerations and the ability to produce high muscular force rapidly. Therefore, team sports demand to have well developed speed, strength, muscular impulse, agility and maximal aerobic power. However, many studies also report a relationship between training and injuries. This means that if players train very hard, it is possible for them to be injured during the season.

Caraffa et al (1996) tried to identify the benefits of proprioceptive training. Proprioceptive training was shown to reduce the incidence of ankle sprains. It could also improve rehabilitation after Anterior Cruciate Ligament (ACL) injuries. ACL injuries could lead to long absence from sports or permanent sports disability. They used 600 soccer players in 40 semi-professional or amateur teams in order to identify the possible preventive effect that proprioceptive training had on four types of wobble-broads during three soccer seasons. Half of these players were assigned in group A and they had to train for 20 minutes per day with 5 different phases of increasing difficulty. The first phase included balance training without any balance board. The second phase of training included balance on a rectangular balance board. The third phase of training included balance on the round board. The fourth phase of training included balance on a combined round and rectangular board. The last phase of training included balance on a so-called BABS board. The other half players were assigned in group B and they were trained normally without any special balance training.

https://doi.org/10.26219/heal.aueb.5673

Both groups were observed for three seasons. Possible ACL lesions were diagnosed by clinical examinations such as KT-1000 measurements, magnetic resonance imaging or computes tomography and arthroscopy. They found an incidence of 1.15 ACL injuries per team per year in group B and 0.15 injuries per team per year in group A. This means that proprioceptive training could significantly reduce the incidence of ACL injuries in soccer players.

The previous research does not include any advanced statistical methods and the conclusion came from the descriptive analysis. More detailed research has been conducted from Gabbet and Domrow (2007). Their purpose was to develop statistical models that estimate the influence of training load on training injury and physical fitness. They used for this purpose 183 rugby league players over two competitive seasons. The players were assessed for their height, body mass, skinfold thickness, vertical jump, 10-m, 20-m and 40-m sprint time, agility and estimated maximal aerobic power. The season lasted from December to September, so the tests performed in December (off-season), March (pre-season), May (mid-season), and August (end-season) as part of the training program.

The observed data from training load and injuries were summarized into pre-season, early-competition and late-competition training phases. Individual data for training loads, fitness and injury were modeled using logistic regression with binomial distribution (injured or no injured) and logit link function. On the other hand, team's data for training load and injury were modeled using linear model.

While physical fitness improved along with training, there was no association between training load and physical fitness during any of the training phases. However, any increase in training load during early-season training led to decrease on agility performance. A relationship was observed between the log of training and odds of injury incidence during each training phase. In particular, for every increase of training load by one arbitrary unit, the odds of injury incidence also increased by 1.50-2.85. Furthermore, during pre-season training phase a relation between training load and injury incidence was observed within the training long range of 155 and 590 arbitrary units. During early competition training phase increases in training load of 175-620 arbitrary units resulted in no further increase in injury incidence. During late-competition training phase, increases in training load of 145-410 arbitrary units also resulted in no further increase in injury incidence. These findings proved that increase in training load, especially during the pre-season training, increase the odds of injury in physical

https://doi.org/10.26219/heal.aueb.5673

sports such as rugby, soccer, basketball, etc. Furthermore, despite the fact that increases in training load during early-competition training phase resulted in no further increase injury incidence, they could affect the agility performance. Last but not least, they suggested reducing the training load during early-season in order to reduce players' the odds of injury without affecting their agility performance.

https://doi.org/10.26219/heal.aueb.5673

https://doi.org/10.26219/heal.aueb.5673

3. Player Performance Analysis

3.1. Introduction

In this chapter I present player performance models and metrics more detailed. Statistical models are divided in four sections based on the sport for which they were created. The first section refers to models that estimate player's performance in soccer. The second section refers to basketball and the third section refers to ice hockey. Fourth section refers to further conclusions that can be obtained by player performance metrics. It is important to mention that all player performance models use the number of goals or points as response variable.

3.2. Models and Metrics

Player performance is an important field in Statistic of Sports which knows great development through years. Many managers use player performance indices in order to decide if the traders really helped their team. Those indices have the form of ratings where players earn points by their performance. By these indices aid, coaches will be able to clarify the true value of a player and judge whether the relevant soccer player is worthy to be acquired or not. By signing players that are undervalued by the market, many teams are more competitive against wealthier teams. Fans also use player performance indices to see how their favorite player or team performs.

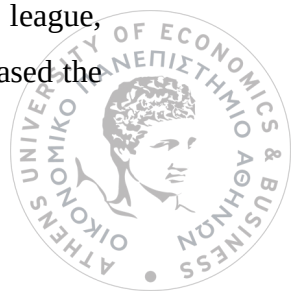
Another reason that makes player performance indices so important is that in many sports ranking is vital. Individual sports, like tennis, already has rating system based on the results in tournaments from the previous 52 weeks. This rating system determines the seeds and pairings at every major tennis tournament in the world. Only a few top players are invited to end-of-season events like ATP World Tour Finals (Men) and the WTA Tour Championships (Women). Golf uses similar rating system. However, we are going to study indices only in team sports. Assessing player's performance in team sports is more complicated due to the structure of the team. Many players could play in different positions. Furthermore, especially in soccer, players could change positions during the match. These changes create interactions between players. Therefore, the construction of player performance index in team sports is more complex than in individual sports.

https://doi.org/10.26219/heal.aueb.5673

3.2.1. Player Performance Metrics for Soccer

According to McHale et al (2012) an index which measures player's performance was presented. This index was a weighted combination of six subindices. The first subindex captures match contributions. If goals for and against are known then the result of the match is also known. Therefore, they modeled the number of shots for each team as independent Poisson distribution. As mean of the distribution they used shot effectiveness that they estimated as a linear combination of variables, based on attacking team such as crosses, dribbles, passes, etc., but also variables based on the defending team such as tackles, clearances cards, etc. Results showed that the variable "cards" was the most important factor that created more shots, with less important being the variable clears. The second subindex was calculated as the product of points which one team won on each match and the ratio of the number of minutes played by one player to the total number of minutes played by all team including substitutes. This subindex measured the performance of one player despite of his position, since it takes into consideration only the minutes that every player participated. However, this subindex undervalued the performance of losing team since points that were gained in this match were zero. The third subindex shared the total points won by all teams according to the proportion of games they played during the season. The average points that one team won per game was 1.34. This subindex was calculated as the product of 1.34 and the ratio of the number of minutes played by one player to the total number of minutes played by the entire team including substitutes. This subindex measured the performance of losing's team players better than the previous one. However, it undervalued the performance of the top player due to the rotation that many coaches do. This rotation was the result of the large number of games that highly profiled teams played during the season. The fourth subindex was the ratio of the number of goals that were scored during the season to the number of league points were won. This subindex estimated how many points every goal contributed to the team that scored.
The fifth subindex was based on the assists in a same way the forth subindex was based on goals. The sixth subindex was based on the total points earned by clean-sheets. This subindex was calculated as a weighted sum of defensive action such as blocks and tackles won, since goalkeeper's saves were more important than striker's tackle.To conclude, the final index was a weighted sum of the points achieved on each subindex.

Therefore, they decided to calculate those indices for season 2008-2009 of Premier league, despite the fact that the first subindex was calculated based on variables which increased the



https://doi.org/10.26219/heal.aueb.5673

winning chances, such as goals, assists and clean sheets. According to fans and media, winning determines the best players. However, results showed that the top 20 players consisted of 7 goalkeepers, 11 defenders, 1 midfielder and 1 striker. These results were due to the reduction of points for off-target shots. Another important result was the absence of top-teams players on the top 20 players according to the first subindex. On the other hand, the final index included a nice mix of players from different positions with 6 strikers, 5 midfielders, 4 defenders and 5 goalkeepers. This index also included players from top teams. According to this index, Nicolas Anelka was the best player of Premier league 2008-2009.

According to McHale and Szczepanski (2014) a model in order to identify the goal scoring ability of footballers was presented. As I mentioned above goals are responsible for the outcome. Therefore, teams have to find players with great ability in scoring goals. This need makes the estimation of goal scoring ability very important. They also estimated shot effectiveness with mixed effect models. Firstly, they denoted the number of shots and the number of goals with n and y respectively. Then, they splitted the process in two components. The first component was the number of shots denoted with n and the second component was the number of goals given the shots $y|n$. This approach was based on Bayes theorem, where the distribution of goals $p(y)$ can be calculated as:

$$\mathrm{p(y)} = \sum_{n} p(n)p(y|n)$$

Therefore, instead of finding a model for the distribution of $p(y)$, they found two models. The first model was for the distribution of the number of shots and the second for the distribution of the number of goals given the number of shots.

Model for the Number of Shots

Concerning the model for the number of shots, it is divided in the basic model, which includes only team information such as home's advantage, team's attacking ability, etc., and the extended model, which includes information about player position or the time he spends in the pitch.

Concerning the basic model, they assumed that the number of shots made by a player in a game, depends on player and team abilities through the linear predictor

$$\mathbf{n_i} = \boldsymbol{\gamma}^{(\mathbf{n})} + \boldsymbol{\nu}^{(\mathbf{n})}\mathbf{h_i} + \boldsymbol{\beta}^{(\mathbf{att})}_{\mathbf{l(i)}} + \boldsymbol{\beta}^{(\mathbf{def})}_{\mathbf{m(i)}} + \mathbf{b_{j(i)}}$$

https://doi.org/10.26219/heal.aueb.5673

Where $\gamma^{(n)}$ is the intercept term, $\nu^{(n)}$ represents the home advantage with h_i being a dummy variable; 1 if the shot was taken by a home team player and 0 otherwise. The parameter $\beta_{l(i)}^{(att)}$ represents the attacking ability of the team the player plays for. The parameter $\beta_{m(i)}^{(\mathrm{def})}$ represents the defensive ability of the team the player plays against. The parameter $b_{j(i)}$ represents the player abilities and they assumed that they come from a multivariate normal distribution, i.e. $b \sim N(0, \sigma I)$, where σ is a variance parameter to be estimated. As for the number of shots n, they assumed that it follows Poisson distribution. The reason for using Poisson was the histogram of the number of shots. It showed that a single-mean Poisson distribution with mean 0.75 fits satisfying. They chose single-mean Poisson distribution because they wanted to ignore the overdispersion that data showed. Finally yet importantly, in order to ensure that the model is identifying well they set

$$\beta_1^{(att)} = \beta_1^{(\mathrm{def})} = 0$$

where 1 is the index of Manchester United who were the champions of the previous years of the data. This set left $2(T-1)$ parameters to estimate, where T denotes the number of teams. Number 2 represents the round-robin construction of the championship. In the extended model, they also included the time that the shooting player spent in the pitch, divided by 100, and his position in team's formation at the game.

Model for the Number of Goals Given the Number of Shots

Concerning the model for the number of goals given the number of shots, it is also divided into the basic model and the extended model. For the basic model, they used only these observations that provided information about the player's abilities to convert shots into goals. Therefore, they used Binomial mixed effects models for the numbers of goals with linear predictor:

$$\boldsymbol{\psi_\kappa} = \boldsymbol{\gamma^{(y)}} + \boldsymbol{\nu^{(y)} h_k} + \boldsymbol{a_{l(k)}^{(att)}} + \boldsymbol{a_{m(k)}^{(\mathrm{def})}} + \boldsymbol{a_{j(k)}}$$

where $\gamma^{(y)}$ is the intercept term and $\nu^{(y)}$ is the home advantage parameter as previously mentioned. The parameter $a_{j(k)}$ represents the ability of each player to convent shots into goals. The parameter $a_{l(k)}^{(att)}$ represents the effect that the player's team had. The parameter $a_{m(k)}^{(\mathrm{def})}$ represents the effect of the opposing team. They assumed that the player parameters α are the random effects and they follow multivariate Normal distribution, i.e. $\alpha \sim N(0, \varphi I)$,

https://doi.org/10.26219/heal.aueb.5673

where φ is a variance parameter to be estimated. Last but not least, they assumed that the number of goals given the number of shots $y|n$ follows Binomial distribution. In the extended model, they also included the number of shots that made by the player in the given game.

In order to estimate the fixed parameters σ and $\beta = (\gamma^{(n)}, \nu^{(n)}, \beta_{l(i)}^{(att)}, \beta_{m(i)}^{(\mathrm{def})})$ for the number of shots, likelihood function was used. In particular:

$$L(\sigma, \beta|n) = \int p(n, b|\sigma, \beta) = \int p(n|b, \beta) f(b|\sigma) db$$

where $p(n|b, \beta)$ is the probability mass function of the multivariate Poisson distribution and $f(b|\sigma)$ is the density function of the multivariate Normal distribution as was previously defined. They approximated this integral by using Laplace approximation.

Furthermore, they fitted several versions of this model and they chose the best according to AIC and BIC. The best basic model for the number of shots included all the variables, with the exception of the attacking ability of the player's team. The best extended model included the variable of the best basic model plus the time the player spent in the pitch and the position of the player. The best basic model for the number of goals given the number of shots included the ability of each player. Despite the fact, that home advantage was statistically significant, it remained because it gave better AIC and BIC. The rest of the variables were removed from the model. The extended model had as explanatory variables the same as the basic plus the number of shots.

3.2.2. Player Performance Metrics for Other Sports

Basketball

Soccer is not the only sport that statistics of sports shows development regarding to assess player's performance. Deshpande and Jensen (2006) estimated the impact that any National Basketball Association (NBA) player had on his team's winning chances.

Other most common metrics that estimates a basketball's player performance are based on box score statistics like points, rebounds, assists, steals, etc. However, these metrics usually mislead us since they ignore the situation in which players perform. In particular, these kinds of metrics assess equally the performance of a player whose team is 30 points ahead with 2 minutes left and the performance of a player whose team points are equal to the opposing team's with 2 minutes to go. At the first situation, the player's performance does not affect his

https://doi.org/10.26219/heal.aueb.5673

team's chances of winning. At the second situation, from the other hand, player's performance highly affects team's chances of winning.

In order to assess how much each player's performance affects his team's winning chances, they proposed a win probability framework and Bayesian linear regression model. This also allowed them to identify if the highly paid players help their teams win the game more than the lower paid. Concerning the use of win probability models to evaluate the performance of professional athletes it dates back to Mills and Mills (1970), who evaluated Major league baseball players. As Studerman (2004) noticed, this methodology has been used several times, with most notably being WPA. In order to compute WPA, first they estimated the win probability for every state of the game. In particular, for each state they credited the pitcher and the batter with the change of their team's win probability respectively. Secondly, they summed these contributions over season to determine how much involved a player was in his team's wins or losses.

Deshpande and Jensen (2006) used an extension of WPA. This extension is similar to the plus-minus statistic, besides that it is computed in the win probability scale instead of points. One disadvantage of the plus-minus statistic is that estimating the performance of a player depends on the contributions of his teammates and opponents. The problem is that the number and quality of teammates and opponents can vary dramatically. To overcome this problem they broke the game into "shifts". As shift, they assumed the period of play between substitutions. They tested player's performance valuation metric in data from 2006 to 2014 NBA championship. They proposed to regress the change ατ the home team win probability during a shift. Therefore, they used indicators to represent the effect of the five home team's players and the five away team's players. This approach contributed in order to estimate the partial effect that each player has on his team's chances of win. They denoted the change of the win probability at the i^{th} shift as y_i. This change in win probability could be directly validated from the performance of the 10 players during the swift. In order to measure each player's impact on the change in the win probability they regressed y_i into indicator variables. Those indicator variables indicated which of the 488 players were on the court during the i^{th} shift. So they modeled:

$$\boldsymbol{Y_i|h_i, a_i} = \boldsymbol{\mu} + \boldsymbol{\theta_{h_{i1}}} + \ldots + \boldsymbol{\theta_{h_{i5}}} - \boldsymbol{\theta_{a_{i1}}} - \ldots - \boldsymbol{\theta_{a_{i5}}} + \boldsymbol{\tau_{H_i}} - \boldsymbol{\tau_{A_i}} + \boldsymbol{\sigma\varepsilon_i}$$

where $\theta = (\theta_1, \ldots, \theta_{488})$ represents the partial effects for the 488 players. The parameter $\tau = (\tau_1, \ldots, \tau_{30})$ represents the partial effects of the 30 teams. The parameters

https://doi.org/10.26219/heal.aueb.5673

$h_i = (h_{i1}, \dots, h_{i5})$ and $a_i = (a_{i1}, \dots, a_{i5})$ are indices on θ corresponding to the home team and away team players, respectively. The parameters H_i and A_i are indices on τ corresponding to teams which were playing during shift i and the parameter ε_i represents the independent standar normal random variables. The intercept μ is the average advantage of the home team. Last but not least, the parameter σ represents the variability of y_i which comes from the uncertainty of measuring y_i and the variability at win probability that cannot be explained from the performance of the players on the court.

In order to fit this model, they had to estimate the probability that the home team wins the game after leading with L points after T seconds ($p_{T,L}$). Stern et al (1994) introduced a probit model in order to estimate $p_{T,L}$ and Bashuk et al (2012) considered empirical estimates of win probability. The probit model, introduced by Stern et al (1994), produced smooth estimates of the win probability based on all games form 2006-2007 to 2012-2013 regular seasons. The empirical estimates of $p_{T,L}$ are given by the proportion of times that the home team won after leading by L points after T seconds. One problem that the empirical estimates showed was the fact that they were very sensitive to small sample size and therefore leading to extreme values. In order to face this problem they let $n_{T,L}$ be the number of games that the home team won after leading with L points after T seconds. Subsequently, they modeled $n_{T,L}$ as a Binomial ($N_{T,L}, p_{T,L}$) random variable, with $N_{T,L}$ representing the number of games that the home team led by L points after T seconds. Afterwards they set Beta ($a_{T,L}, b_{T,L}$) as prior on $p_{T,L}$ and estimated it as the posterior mean given by:

$$p_{T,L} = \frac{n_{T,L} + a_{T,L}}{N_{T,L} + a_{T,L} + b_{T,L}}$$

Based on this expression they added wins (a) and losses (b) to the observed counts. These wins and losses were referred as "pseudo-wins" and "pseudo-losses", respectively. Moreover, these pseudo-games tend to shrink the original empirical estimates of $p_{T,L}$. Due to the large number of variables they resulted in high level of collinearity. Therefore, they used prior distribution in order to shrink every component of θ towards zero. They chose Laplace prior which equals to Lasso shrinkage method. The reason for choosing Laplace prior was because it pulled the small partial effects of the player towards zero faster than the Normal prior. They let P^i be the vector that indicates which players were on the court during shift i. So, P_j^i is equal to 1 if the player j is on the court and 0 otherwise. In the same way, T^i is a vector that

https://doi.org/10.26219/heal.aueb.5673

indicates which teams were on the court during shift i. So, T_k^i is equal to 1 if the team k is the home team, -1 if team k is the away team and 0 otherwise. Then they modeled:

$$\boldsymbol{Y_i|P^i, T^i \sim N(\mu + P^i \theta + T^i \tau, \sigma^2)}$$

Subsequently they placed the Laplacian priors on the components of θ and τ. In order to calculate the posterior distribution of $(\mu, \theta, \tau, \sigma^2)$ they used Gibbs sampling due to the hierarchical structure of the model. To determine this model they made two strong assumptions. First they assumed that y_i's are independent since some player are not substituted at the end of a shift. The second assumption is that, conditional on (P_i, T_i), the y_i's are Gaussian with constant variance.

Next, they used Gibbs sampler function to obtain 1000 independent samples in order to estimate the marginal posterior density of each player's partial effect. Then, they compared players according to their partial effect on team's winning chances. However, this comparison was complicated since many players perform under different circumstances. Therefore, they defined which players were most comparable. This definition was based on the number of shifts that player performed, his team's average win probability at the start of the shift, the average duration of these shifts and the average length of each shit. This information was named leverage profile. To continue with, they calculated Mahalanobis distance between the leverage profiles of each pair of players. When Mahalanobis distance between two players was very small, their leverage profiles were very similar. So, after these computations, they could compare their partial effect on their team's win probability. In a similar way, they estimated the team effect τ because they wanted to ensure that the partial effects of players were not overly undervalued if they played on a bad team or overly overrated if they played on a good team.

Impact Ranking

Due to the fact that player's partial effects resulted according to their value to their team, they made their estimation more valuable. Thus, they thought to create ranking for every team that showed which player were more impactful for their team's winning probability. In particular, they could rank the partial effects of all players of each team and identify which rank order had the highest posterior rank-order. However, the average NBA team had 15 players which made over 1 trillion combinations. Therefore, they proposed to identify player ranking as an average over 1000 posterior samples for each team. This ranking was called Impact Ranking.

https://doi.org/10.26219/heal.aueb.5673

By adding the salary for every player, they determined which players were undervalued or overvalued by their teams.

Impact Score

Partial effects could also rank all players in the league. One natural approach would be to rank players according to their posterior mean of their partial effects. However, some players with large posterior mean partial effects had also large posterior standard deviation. Therefore, they proposed to rank players according to their Impact Score. As Impact Score was defined the ratio between the posterior mean and the posterior standard deviation of player's partial effect. As they denoted, Impact Score was viewed as balance between player's estimated "risk" and player's estimated "reward".

Player Evaluation Rating (PER)

As we previously stated, most basketball metrics used box-score statistics to evaluate player's performance. One of them was introduced by Hollinger (2004) and was called PER. PER sums up all the positive accomplishments and subtracts the entire negative in order to rank the performance per minute. However, PER did not take into consideration under what circumstances the player performed. Whereas Impact Score included this information, since its calculation based on player's partial effects on team winning chances changes from shift to shift.

To continue with, they also compared how different five-man line-ups match ups against each other. Therefore, they simulated the posterior distribution of those line-up's effect on its team's win chances. Then, they summed the corresponding rankings of each posterior sample. The approach of estimating the Impact Score of the line-up is similar to the one they had followed for the player's Impact Score. In particular, they divided the posterior mean effect of the line-up by the posterior standard deviation of its effect.

Hockey

Similar techniques with basketball were used for hockey from Gramacy et al (2013). However, the tradition metric for hockey is the plus-minus despite the fact that it measures only the marginal effect of the player, do not account for sample sizes and provide very noise estimates of performance.

Because of the large numbers of players, the frequent changes of lines and the rareness of goals, measuring the player performance in hockey is a difficult task. The most known player

https://doi.org/10.26219/heal.aueb.5673

performance plus-minus metric counts the number of goals scored by the player's team minus the number of goals scored by the opposing team while the player is on the ice. More complex indices, including more information than goal scoring have been presented. However, analysts do not agreed with the importance of this added information.

Back in the article, they used logistic regression model to estimate the partial effect of individual players on goal scoring in hockey. They also used two techniques for variable selection and for further analysis. These techniques were based on Bayesian approach like the one that was used on basketball.

Their data included 1467 players and 18.154 goals without overtime goals. Their data were fixed into a response variable and a design matrix with two parts $X_T X_P$. The first part X_T had indicator variables corresponding to team and the second part X_P had indicator variables for player identities. For every goal i that scored, the response variable $y_i = 1$ if the goal was scored by the home team or $y_i = -1$ when the goal was scored from the away team. In the design matrix, the i^{th} row indicates the teams played and the players on ice when the goal is scored. In addition, $X_{T_{ij}} = 1$ for the home team and $X_{T_{ij}} = -1$ for the away team in the match that the goal scored. Furthermore $X_{P_{ij}} = 1$ and $X_{P_{ij}} = -1$ for the players of home and away team, respectively. The rest indicators of i^{th}row are equal to zero. This means that the design matrix's dimensions are 18.154×1497 and every row will have 1485 zeros.

Their logistic regression equation is:

$$\log\left(\frac{q_i}{1-q_i}\right) = x_{Ti}^{'}\alpha + x_{Pi}^{'}\beta$$

Where $q_i = p(y_i = 1)$, X_{P_i} is the vector corresponding to the players in i^{th} row and X_{T_i} is the vector corresponding to the 30 teams in the i^{th} row. Moreover, α and β were vectors of teams and players effects, respectively.

Therefore, they used Bayesian approach in order to shrinkage their regression coefficients. In particular, Bayesian approach uses the likelihood $p(y|\alpha, \beta)$ of the model that describes data's function with the unknown parameters α and β and prior distribution $p(\alpha)$, $p(\beta)$ for each unknown parameter, respectively. Then, Bayesian approach is based on the posterior distribution:

$$p(\alpha, \beta|y) \propto p(y|\alpha, \beta) \times p(\alpha, \beta)$$

https://doi.org/10.26219/heal.aueb.5673

This posterior distribution is usually estimated via simulation methods using MCMC. However, they limited their analysis into maximum a posteriori (MAP) estimates. Bayesian approach helped them to guard their model from overfitting and resulted in steady estimates of individual player effects.

Using a (centered at zero) prior on player's parameter β makes each coefficient to converge in the central value of zero. This prior distribution is the same as putting a penalty term when $\beta_j \neq 0$, which can be used to give us an estimate of β_j. Different types of prior concluded in different type of penalty. They finally chose Laplace prior because it gives penalty equal to $\lambda_j |\beta_j|$ in the objective function. Putting Laplace prior, centered to zero, is the same as using Lasso in order to make variable selection. Choosing the value of λ_j was very demanding. For instance, a very low value, like $E[\lambda_j] = \frac{1}{10}$ gives a 98% of players with non-zero β_j. By finding the ideal value of λ_j, they could identify which players were above average. As for team's parameter α, they used normal, centered at zero, distribution which put $\lambda_j \beta_j^2$ as penalty term. The variance term in normal prior is fixed at $\sigma = 1$.

Finally, they considered two models. The first was mentioned above and was used to give point estimates of player's coefficient. The second was called player only model and included only the player's parameter β. The second model showed how far from team's average the player's performance was. Only non-zero players were in both models. This way they could identify which player really helped their team and which player's performance was absorbed by their respective teammates.

One difference of MAP comparing to other plus-minus metric is that partial player estimates is in log odds. In other word e^{β_j} symbolized the ration of for-versus-against for every goal score when player j was on ice. On the other hand, traditional plus-minus metrics is calculated as the difference of for-versus-against.

The point estimates and samples from full posterior distribution allowed them to do further analysis which could not be conducted before. Such analysis was pairwise comparisons. Moreover, by obtaining more data sources such as salary information and constraints thereon (salary budgets), their approach offered potential for explanatory analysis and decision making by coaches, general managers, etc.

https://doi.org/10.26219/heal.aueb.5673

Finally, their analysis showed that Pavel Datsyuk dominated over other stars such as Sidney Crosby and Alex Ovechkin. They also found several players, such as Evgeni Malkin, who did not have significant player effect.

3.2.3. Other Uses of Player's Performance Estimates

Saebo and Hvattum (2017) introduced another use of player's performance estimates. They first estimated player's performance via a regression model and then they used these ratings as variables in a standard transfer fee regression model together with other explanatory variables.

They used the traditional plus-minus metric, which measures the number of goals scored minus the number of goals conceded when the given player was used by the team. But first, they divided each match in parts. Each part is defined as a period of time where players had not change, similar with the approach that followed in basketball metric. For each part i, they defined $a_{ij} = 1$ if the player j played for the home team during segment, $a_{ij} = 0$ if player j did not play in segment and $a_{ij} = -1$ if player j played for the away team during segment i. They defined as β_{ij} the number of goals scored minus the number of goals conceded. The model takes the form:

$$\alpha x = \beta$$

where x is a vector of player's ratings.Ratings were found by minimizing model errors, given as the sum of squared differences between the actual goal differences, β and model predictions $\hat{\beta} = \alpha x$. So, vector x that minimized this error resulted from the equation:

$$x = (\alpha^T \alpha)^{-1} \alpha^T \beta$$

However, due to high errors, they used ridge regression which is the same as using Normal, centered at zero, prior distribution on the explanatory variables, which was mentioned above. This approach helped them battle the noise that appeared before. They also made some improvements on this model by adding the duration D_i in minutes, since the duration of each part is not the same. They also took into consideration the home advantage of the team as a dummy. Standard plus-minus ratings did not reward past performances. By making ratings using time-independent data there was not margin for dynamic performance standards. So, they proposed to exponentially down-weight all past observations, depending on age observations t, and a discounting parameter k. Thus, the variable $a_{ij} = e^{-kt}$ if the player

https://doi.org/10.26219/heal.aueb.5673

plays for home team in the segment i, $a_{ij} = 0$ if the player did not play in the segment and $a_{ij} = -e^{-kt}$ if the player played for the away team in the segment i. This way they designed a more dynamic rating metric that could change more quickly. The definition of goal difference β_{ij} also changed. Because considering that H_i and A_i is the number of goals for home and away team, respectively. After scaling up to 90 minutes, β_i is defined as:

$$\beta_i = \frac{90(H_i - A_i)e^{-kt}}{D_i}$$

Results shown that out of the 20 nominees for the FIFA 2014 Ballon d'Or, eight also featured in the list of top 20 players of the model, including the three finalists. Seven out of the eleven of FIFA Team of Year 2014 also appeared on the list of top 20 players according to their model.

We have to mention that these ratings were calculated before the transfer and is included as a performance measure in the transfer fee models. They proposed two transfer fee models. Concerning the first model, with exception of ratings, they used 25 other variables. Variables such as player's foot (with two dummies), main position (5 dummies), height (2 dummies), nationality (7 dummies), appearances with youth or senior national team (2 dummies), one factor that scaled the total performances with the national team until the date of transfer assuming that the player debut at the age of 18, another factor with similar approach about the injuries, player's age (2 dummies), player's month of birth and the year of transfer. The second model also included the remaining contract duration and was used for the English clubs.

To conclude, players who used both feet tend to increase their transfer fees. Wingers were under-valued and height did not seem to affect transfer fees. Nationality affected transfer fees since European players tend to have higher value than Africans, etc. Also, international experience is a significant variable. Players that played in the senior national team were greatly rewarded. However player with only youth international experience were undervalued. In addition, age variables shown that transfer fees were smaller for very young players due to the uncertainty as well as older players due to lack of potential and risk of injuries. At the second model the variable of contract duration was significant. When a contract is about to expire the value of transfer decreases because soon the player would be free to sign in any team he wants.

https://doi.org/10.26219/heal.aueb.5673

https://doi.org/10.26219/heal.aueb.5673

4. Analysis of Player Contribution Using Participation Data

4.1. Introduction

In this chapter I start by presenting the dataset which is based on Leicester's course in Barclay's Premier League of 2015-2016. This chapter is divided in 3 sections. First section refers to descriptive analysis of all variables. Descriptive analysis includes summary statistics, barplots and correlation coefficients for the number of minutes every player played in each match. Second section refers to player performance estimation via descriptive analysis. Third section refers to inferential analysis, which includes 22 Poisson models one for each player in order to estimate player's performance. Last section contains the comparison between the results of these two methods.

4.2. Descriptive Analysis

My data includes goals that scored from every team on each game for the English Barclays Premier League of 2015-2016. At the end of this season Leicester was crowned the champion team for the first time in its history. Therefore, I also obtained the number of minutes that every player of Leicester City played on each game. If Leicester is not taking part in a game, then the number of minutes that every player played is zero. My data also includes the home effect, which is equal to 1 or 0 for home or away team, respectively. The last two variables are the attacking team and the defensive team. All this information is obtained from the official site of BBC. My data includes not only the number of minutes that the player of the starting line-up spend in the field, but also the number of minutes that players who came as a substitution spend in the field.

In my descriptive analysis I used only the games that Leicester City played. Descriptive analysis refers to the number of minutes played by a Leicester player. Therefore, having the games that Leicester is not taking part will only cause trouble and false conclusion. At the first section, I provide descriptive analysis for the time that every Leicester player played during season 2015-2016 of English Barclays Premier League. At the second section, I present barplots in order to identify how Leicester's coach Ranieri shared the time through season. At the last section of statistical analysis I present correlation table for each player's minutes so that I can understand how the appearance of one player in the first team affects his

https://doi.org/10.26219/heal.aueb.5673

teammates. Descriptive analysis, barplots and correlation plots are made in the statistical program R-studio.

Players	Vars	N	Mean	Sd	Min	Max	Range	Skew	Kurtosis
K.Schmeichel	1	38	90	0	90	90	0	-	-
R.D.Laet	2	38	17.24	33.11	0	90	90	1.52	0.49
W.Morgan	3	38	90	0	90	90	0	-	-
R.Huth	4	38	82.89	24.59	0	90	90	-3	7.19
D.Amartey	5	38	2.66	12.76	0	78	78	5.43	29.04
D.Simpson	6	38	68.76	36.74	0	90	90	-1.25	-0.34
M.Wasilewski	7	38	7.95	24.89	0	90	90	2.82	6.35
C.Fuchs	8	38	71.32	36.06	0	90	90	-1.37	-0.1
Y.Benalouane	9	38	1.68	5.97	0	25	25	3.22	8.93
D.Drinkwater	10	38	77.61	29.53	0	90	90	-2.08	2.51
A.King	11	38	27.58	36.38	0	90	90	0.91	-0.96
M.Albrighton	12	38	72.82	22.71	8	90	82	-1.59	1.82
NG.Kante	13	38	79.47	26.02	0	90	90	-2.13	2.87
J.Schlupp	14	38	36.37	37.97	0	90	90	0.41	-1.59
D.Gray	15	38	4.63	9.74	0	45	45	2.48	6.26
N.Dyer	16	38	5.55	12.81	0	45	45	2.33	4.21
R.Mahrez	17	38	80.45	17.56	0	90	90	-2.95	9.58
G.Inler	18	38	5.11	16.26	0	67	67	3.06	8.01
J.Dodoo	19	38	0.5	3.08	0	19	19	5.69	31.16
J.Vardy	20	38	82.63	21.02	0	90	90	-3.23	9.67
A.Kramaric	21	38	0.58	2.5	0	12	12	3.91	13.77
S.Okazaki	22	38	54.47	25.19	0	90	90	-0.8	-0.17
L.Ulloa	23	38	25.42	27.53	0	90	90	1.16	0.27

Table 1: Descriptive Statistics

In table 1 are presented the descriptive statistics for the minutes that every Leicester player spent in the field. Rows represent all Leicester players. First column is the name of the player and second column is the serial number of every player. Third column represents the number of games Leicester played during season 2015-2016, which is 38. The most important column is the forth, which represents the mean number of minutes that every player played in each game. As we clearly see, goalkeeper Kasper Schmeichel has mean number of minutes 90, which means that he played in all of Leicester games. The same goes for the central defender Wes Morgan, who also has mean number of minutes 90. Another central defender, Robert Huth, has large mean number of minutes (82.89), which means that he played in most of Leicester games. This leads us to the conclusion that coach Ranieri used the same central defender duo in most of Leicester's games. Next in order is the forward Jamie Vardy with mean number of minutes 82.63. Fifth and last player with mean number of minutes over 80 is

https://doi.org/10.26219/heal.aueb.5673

Riyad Mahrez with mean number of minutes 80.45. It is important to mention that K.Schmeichel kept a clean sheet in 15 matches. In addition, W.Morgan, R.Mahrez and J.Vardy were named in the Professional Association (PFA) team of the year. J.Vardy also named PFA player's player of the season and Football Writer's Association (FWA) footballer of the year. Last but not least PFA Player's Player of the year was awarded to R.Mahrez.

Standard deviation in the sixth column shows how wide every player's data spreads around his mean. Goalkeepers and central defenders have fairly small standard deviation. On the other hand, players like Jeff Schulpp and Andy King who are midfielders or Ritchie De Laet and Danny Simpson who are right defenders, present greater standard deviation. Left defender Christian Fuchs also appears big standard deviation. To sum up, it seems that coach Ranieri shared his team's game time among his roster fairly balanced with the exception of goalkeeper and the central defenders. In section 2 we will see more with the help of barplots.

Sixth column is the minimum number of minutes for each player and the seventh column is the maximum number of minutes for each player. Eighth column is about range, which is calculated as the maximum value minus the minimum. Ninth column represents the skewness, which measures the asymmetry of the probability distribution of the data. Tenth column represents kurtosis, which similarly to skewness, measures the shape of the probability distribution.

4.3. Plots

In this section we are going to understand how coach Ranieri shared the match time among his roster. It is known that Barclay's Premier league is very demanding tournament because teams used to compete every weekend from August to May without Christmas break. Therefore, rotation is a very important factor of keeping the team at the top through season. I used barplots to identify this rotation because of the poor variety of the number of minutes that each player spend in the field. I chose to divide every player's barplot according to his main position.

https://doi.org/10.26219/heal.aueb.5673

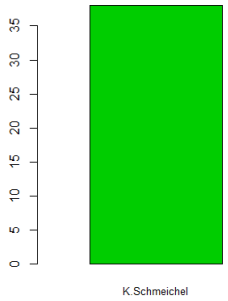


Figure 1: Goalkeepers Green:whole, Red:none, Blue: as substitution Light Blue: being substituted

At first we start with the position of goalkeeper. As we can clearly see at Figure 1, Kasper Schmeichel played in all 38 league matches during season 2015-2016. So, coach Ranieri did not make any rotation in the position of goalkeeper. From these 38 matches Schmeichel managed to keep his clean sheet in 15 matches, which is the second best behind Arsenal's Peter Cech, who kept his clean sheet in 16 matches.

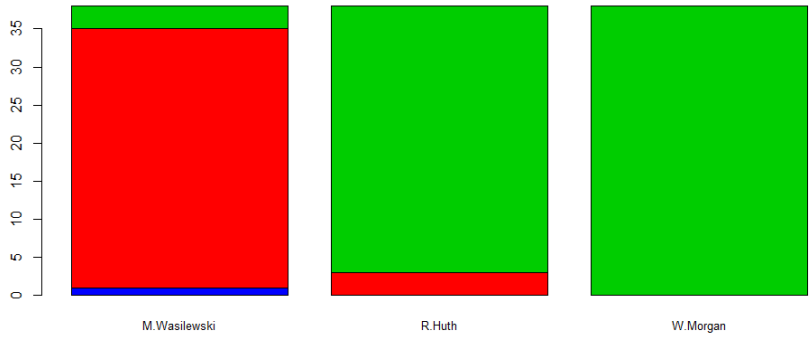


Figure 2: Central Defenders Green:whole, Red:none, Blue: as substitution Light Blue: being substituted

Figure 2 presents the position of the central defender. As we clearly see at Figure 2, W.Morgan played in all 38 matches of Leicester city without being substituted. R. Huth has also played in 35 out of 38 games without being substituted. The rest 3 games, W.Morgan played them with Marcin Wasilevski as his partner. M.Wasilevski also has 1 appearance as left defender for 32 minutes. The English youngster Liam Moore had 0 appearances through season. To sum up coach Ranieri used the same central defender duo almost through the whole season.

https://doi.org/10.26219/heal.aueb.5673

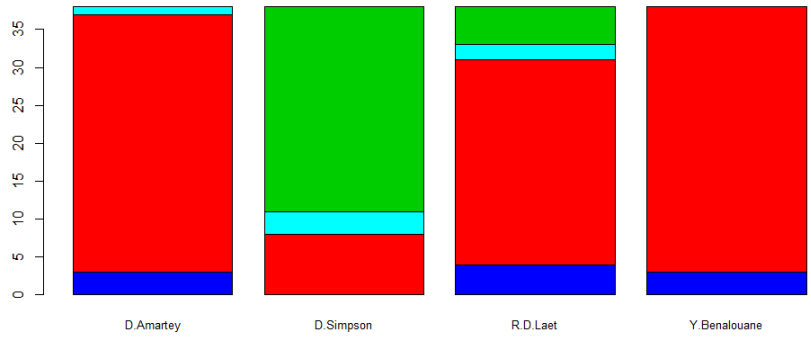


Figure 3: Right Defenders Green:whole, Red:none, Blue: as substitution Light Blue: being substituted

Figure 3 presents the position of right defender. As we clearly see at Figure 3 we have 4 players for one position spot. This makes coach's rotation wider than central defender's and goalkeeper's rotation. This plot show to us that Danny Simpson should be considered the main right defender because he played in 30 out of 38 matches, where at the 26 of them without being substituted. R.D.Laet played in 11 matches, where 5 of them had duration 90 minutes. Yohan Benalouane played in all 4 matches as a substitution. Daniel Amartey made 4 appereances; one of them in the starting line-up.

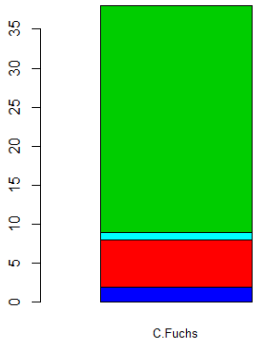


Figure 4: Left Defenders Green:whole, Red:none, Blue: as substitution Light Blue: being substituted

Figure 4 presents the position of left defender. Christian Fuchs should be considered the main left defender of the team since he played in 32 games, where 26 of them were without being substituted. However, English youngster Ben Chilwell did not make any appearance. That means that coach Ranieri used another player to cover Fuchs's absence, despite the fact that left defender was not his main position.

https://doi.org/10.26219/heal.aueb.5673

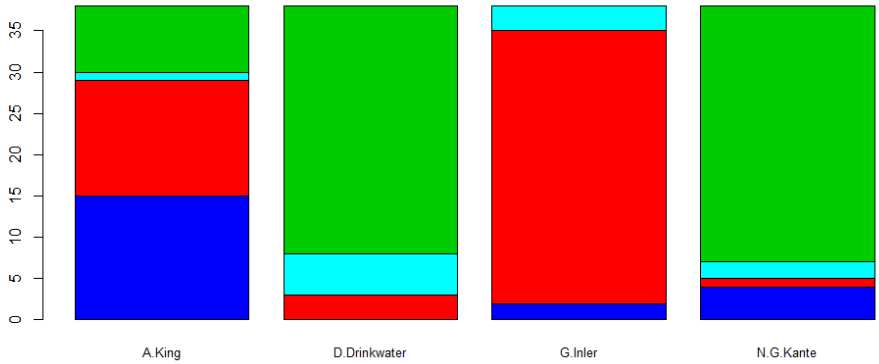


Figure 5: Central Midfielders Green:whole, Red:none, Blue: as substitution Light Blue: being substituted

Figure 5 presents the position of central midfielder. Here we have 4 players for 2 spots. N'Golo Kante played in 37 matches, where at 31 of them he participated in the whole match and at 2 of them as substitution. Danny Drinkwater made 34 appearances all in the starting line-up, where 28 of them were in the whole match. Andy King played in 17 matches, where 8 of them in the whole match. Swiss Gohan Inler played in 5 matches, where 3 of them were in the starting line-up.

As it seems, the competition for the central midfielder spot is greater than the previous mentioned positions. Despite the fact that N'Golo Kante and Danny Drinkwater were considered the main central midfielders, Andy King also played in many matches. This is natural since central midfielders connect defenders and forwards. This connection is very demanding because it needs stamina and fairly good offensive and defensive skills. Therefore, coach Raniery made greater rotation for this position in order to keep his team's stamina in high levels.

https://doi.org/10.26219/heal.aueb.5673

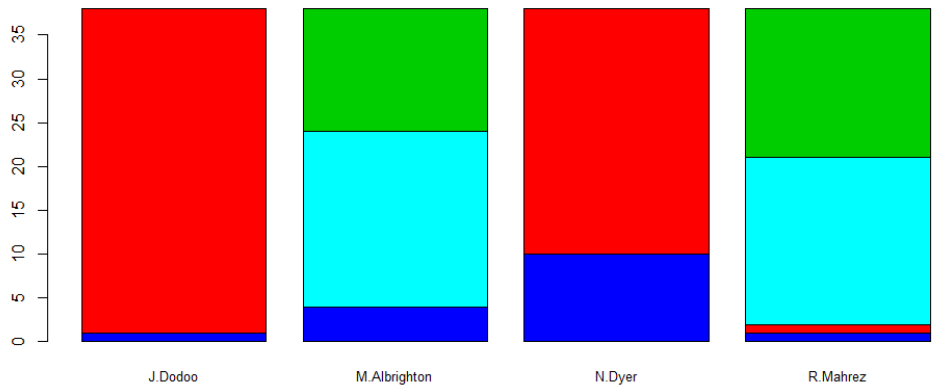


Figure 6: Right Midfielders Green:whole, Red:none, Blue: as substitution Light Blue: being substituted

Figure 6 presents the position of right midfielder. Riyad Mahrez made 37 appearances, where 36 of them were in the starting line-up. Marc Albrighton played in all 38 matches, where 34 of them were in the starting line-up. Joe Dodoo made only one appearance for 19 minutes. Nathan Dyer played in 12 matches, all as a substitution.

The competition for the spot of right midfielder was mainly between Mahrez and Albrighton. However, Mahrez and Albrighton can also play as central midfielder or left midfielder. This gave the coach the flexibility of using both players in the same time but in different positions.

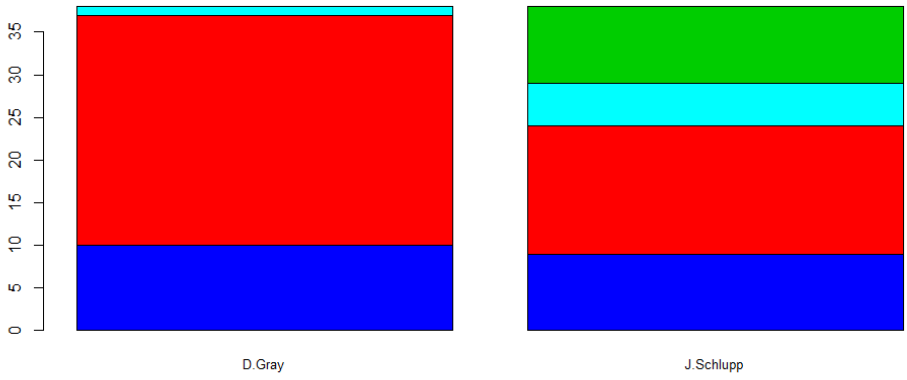


Figure 7: Left Midfielders Green:whole, Red:none, Blue: as substitution Light Blue: being substituted

Figure 7 presents the position of left midfielder. Jeff Schlupp made 24 appearances, where 14 of them were in the starting line-up. Demarai Gray made 12 appearances but only one was in the starting line-up. As I previously mentioned, probably Albrighton and Mahrez also filled

https://doi.org/10.26219/heal.aueb.5673

the position of left midfielders because Gray and Schlupp combined made less than 38 appearances.

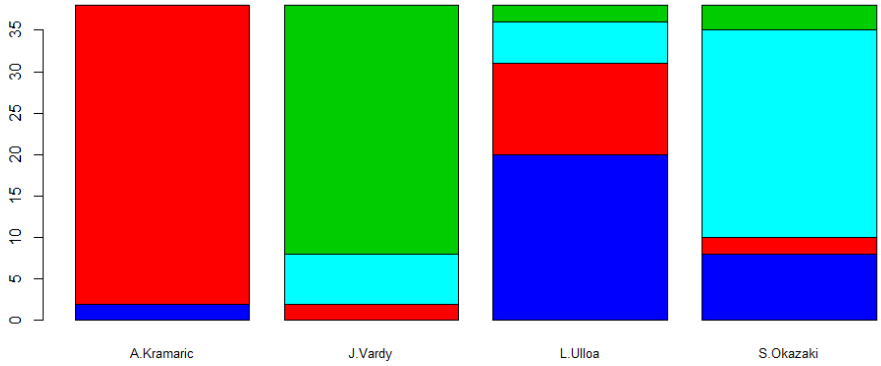


Figure 8: Forwards Green:whole, Red:none, Blue: as substitution Light Blue: being substituted

Figure 8 presents the forwards. Here we have 4 players for 2 spots since Leicester usually plays the traditional $4-4-2$ system. Andrej Kramaric made only 2 appearances, which leave us with 3 players. Jamie Vardy played in 36 matches all in the starting line-up with 26 goals (first scorer of Leicester), which made him the owner of the one spot. Okazaki played 36 matches, where 28 of them in the starting line-up and he scored 5 goals. The Argentinean striker Leonardo Uloa also made 29 appearances but only 7 in the starting line-up and scored 6 goals. Despite the fact that Vardy and Okazaki were the main offensive duo, L.Ulloa was more productive than the Japanese striker.

https://doi.org/10.26219/heal.aueb.5673

4.3.1. Correlation

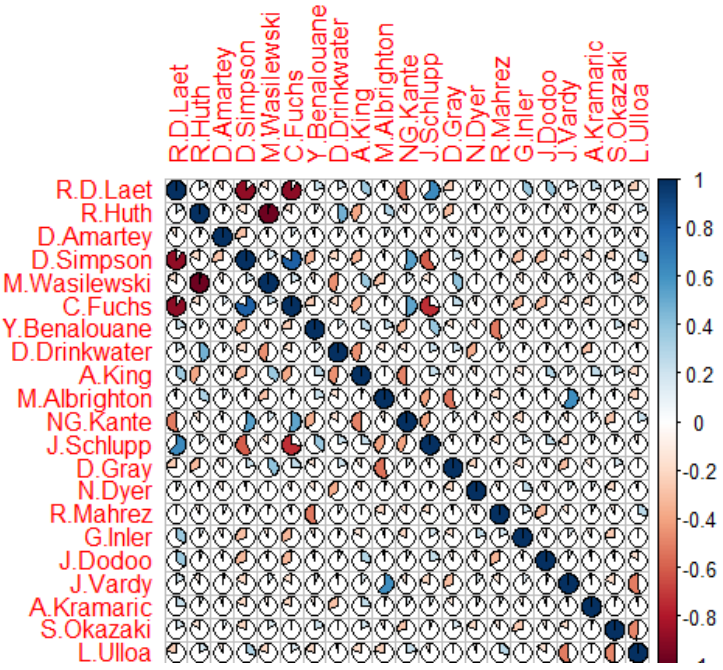


Figure 9: Correlation

Figure 9 presents the correlation coefficients of Leicester's players. Correlation will help us identify how one's player appearance in the pitch affected his teammates. We exclude from our analysis players that played in all matches and players that did not played in any of the 30 matches.

R.D.Laet has a very strong negative correlation with D.Simpson (−0.881) and C.Fuchs (−0.893). As a matter of fact, when Laet's game time was increased, the game time of Simpson and Fuchs was decreased. This can be explained from the fact that Laet was used either as right defender like Simpson or as a left defender like Fuchs. R.Huth and M.Wasilewski also has a very strong negative correlation (−0.978), since the competition for the spot next to W.Morgan in the center of defense was between those two. On the other hand, D.Simpson and C.Fuchs has a very strong positive correlation (0.811), which means that when Simpson was in the field probably Fuchs was in the field too. Strong positive correlation appeared on the couples of R.D.Laet - J.Schlupp (0.636) and J.Vardy – M.Albrighton (0.610), which means that when one of them is in the field, the other probably is in the field too. On the other hand, strong negative correlation appeared on the couples of D.Simpson – J.Schlupp (−0.608) and C.Fuchs – J.Schlupp (−0.738) which means that when one of them used to play, the other was not. The rest of the correlation coefficients are not so strong (<0.6).

https://doi.org/10.26219/heal.aueb.5673

4.4. Player's Contributions based on Descriptives

In this section I present my player performance analysis based on descriptives. First I calculate the number of minutes that each player played. Then, I divide this number with 90 to estimate the number of games he played. Secondly, I calculate the number of goals his team scored and conceded during his appearance in the pitch. Lastly, I divide the number of goals with the estimated number of games. Thus I identify the number of goals that Leicester scores or suffer when each player is in the pitch. The following table shows the results.

Player	Attack	Defense
K.Schmeichel	1.79	0.95
R.D.Laet	2.06	2.20
W.Morgan	1.79	0.95
R.Huth	1.74	0.91
D.Amartey	1.78	1.78
D.Simpson	1.79	0.65
M.Wasilewski	2.09	1.79
C.Fuchs	1.66	0.66
Y.Benalouane	0.00	1.41
D.Drinkwater	1.72	0.83
A.King	2.15	1.46
M.Albrighton	1.76	0.72
N.G.Kante	1.76	0.86
J.Schlupp	1.89	1.56
D.Gray	1.02	1.53
N.Dyer	2.56	1.71
R.Mahrez	1.83	0.77
G.Inler	0.46	2.32
J.Dodoo	4.74	0.00
J.Vardy	1.72	0.92
A.Kramaric	4.09	8.18
S.Okazaki	1.70	0.87
L.Ulloa	2.42	1.21

Table 2: Estimated Number of Goals in Each Game

In Table 2 we can see that Kramaric and Dodoo have the greater numbers of goals in favor of their team. However, these players played only in two and one matches, respectively. Players such as Inler, Gray, Benalouane, Amartey, and Ritchie de Laet have greater number of goals against Leicester than in favor. On the other hand, players such as Vardy, Mahrez, Ulloa, Okazaki, Kante, Albrighton, Drinkwater, Fuchs, Schmeichel, Morgan, and Huth have almost

https://doi.org/10.26219/heal.aueb.5673

double number of goals in favor of Leicester than against. It is no accident that 10 out of 11 players are in the first lineup in the most matches.

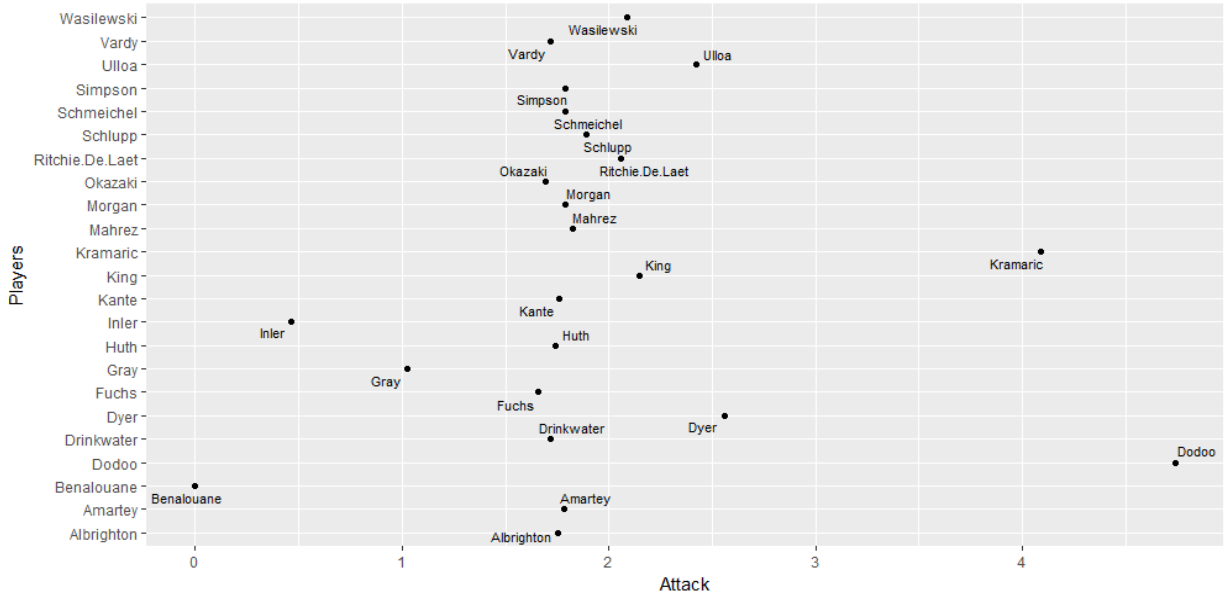


Figure 10:Attacking Impact Based on Descriptive Statistics

In Figure 10 I show a visual presentation of player's attacking ability. As we can see Dodoo and Kramaric have the biggest attacking impact. Ulloa and Dyer also have fairly good attacking impact. On the other hand, Benalouane, Inler, and Gray have the smallest attacking impact. Furthermore, starters seem to have equal level of attacking impact.

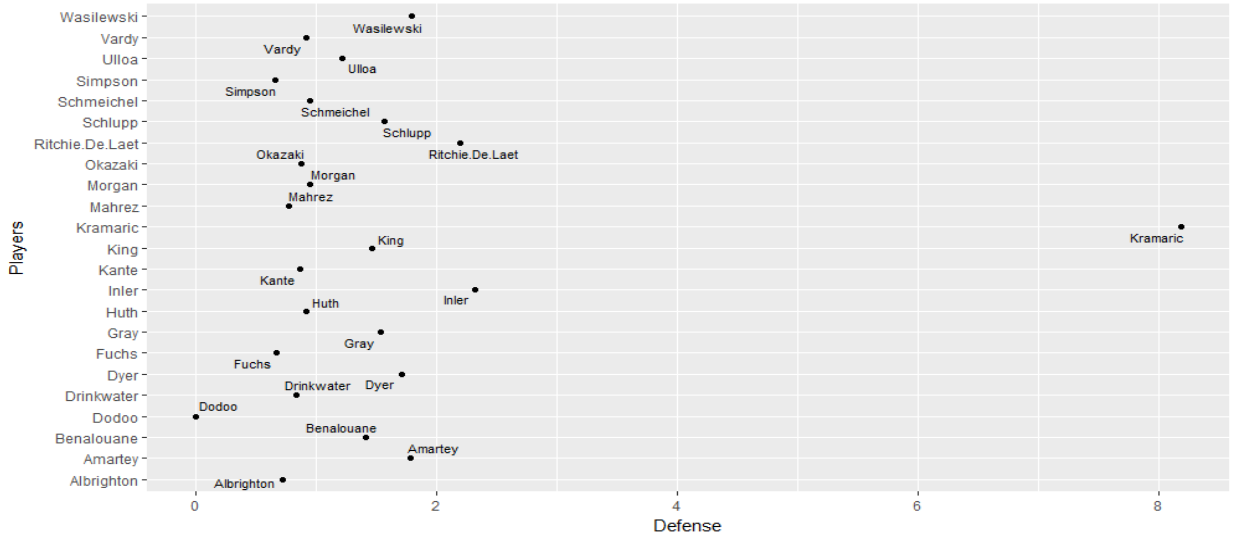


Figure 11: Defensive Ability Based on Descriptive Statistics

In Figure 11 we follow a different approach than in Figure 10. In particular, players with big defensive impact hurt Leicester because they increase the number of goals. Kramaric has the

https://doi.org/10.26219/heal.aueb.5673

worst defensive impact. On the other hand Dodoo has the best defensive impact. However Dodoo only played in one match as a substitution. Similarly, starters seem to have equal level of defensive impact.

4.5. Inferential Analysis

In this section we will try to understand how Leicester's players effect the number of goals that any Barclay's Premier League's team scores. We count their effect based on the number of minutes they spend in the field. The method I used is generalized linear model. In particular, we will try to understand which players were more decisive for the number of goals in all games. As for our data, every player is divided into two sub-variables, one for their defensive ability and one for their offensive ability. As a number in their defensive ability, they have the number of minutes they played divided by the match duration of 90 minutes when Leicester was in defense. The same approach has been followed for their offensive ability.

4.5.1. Generalized Linear Models

As I previously mentioned, I am going to use generalized linear models to estimate the effect of every Leicester's player on the number of goals. Generalized linear models (GLMs) are defined as a flexible generalization of linear models, where the response variable follows a distribution other than normal. The response variable is assumed to be generated from a distribution that belongs to the exponential family. GLMs allow the linear model to be related with the response variable via a link function. At first, the linear predictor $\boldsymbol{\eta = X * \beta}$ is estimated from the exploratory variables (X_i). To continue with, the link function is used to model the relationship between the linear predictor and the mean of the distribution function. In particular, if we symbolize the mean of the distribution function μ, then $\mu = g(\eta)$, where g is the link function. Most popular link functions are log for Poisson distribution (i.e. $\eta = \ln \mu$), logit for Binomial or Multinomial distribution ($\eta = \ln \frac{\mu}{1-\mu}$), identity for Normal distribution ($\eta = \mu$) and inverse for Gamma distribution ($\eta = \mu^{-1}$).

https://doi.org/10.26219/heal.aueb.5673

4.5.2. Response Variable Distribution

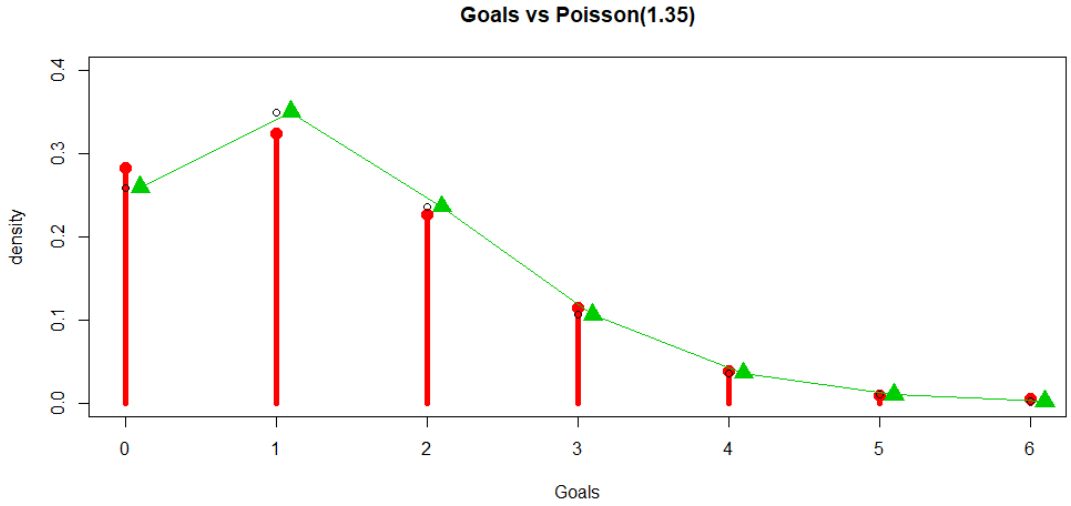


Figure 12: Plot of goals versus Poison (1.35)

In order to construct the model, I used the variable goals as dependent variable and all the sub-variables would be the explanatory variables. First in generalized linear models, we have to identify the distribution that the response variable goals follows. In Figure 12 I present a plot of 760 counts of goals and the density line of Poisson with, the expended number of goals, 1.35 as mean. As we clearly see, the density line describes very well the values that are over one. Moreover, Poisson values can take only non-negative integers. Our response variable satisfies this assumption. Therefore, I am going to use Poisson distribution generalized linear models to estimate the effect of each player on the number of goals.

Variable	Vars	N	Mean	Sd	Min	Max	Range	Skew	Kurtosis
Goals	1	760	1.35	1.21	0	6	6	0.87	0.62

Table 3: Descriptive Statistics for Goals

Goodness-of-fittest for Poisson distribution			
	X^2	df	$P(> X^2)$
Likelihood Ratio	7.17	5	0.21

Table 4: Distribution Test for Goals

In table 3 mean and variance of goals' are close enough which strengthens our hypothesis. In table 4, I used the function *goodfit* from package "*vcd*". This function computes the fitted values of a discrete distribution to count the data given in Goals. The distribution parameters are estimated via maximum likelihood. The null hypothesis of this test is that our variable

https://doi.org/10.26219/heal.aueb.5673

follows Poisson distribution. The test statistic shows that $X^2 = 7{,}17$ with 5 degrees of freedom. P-value equals to 0.20, which indicates that we do not reject the null hypothesis.

4.5.3. Poisson Generalized Linear Model

According to the previous information, the model that I proposed is Poisson generalized linear model. This means that we have to use the exponential of the observed coefficient in order to identify each variable's effect on mean. Our model selection will be based on AIC. We can also use the deviance of the model in order to check if our model fits with "saturated" model. Saturated is the model with all variables and interactions between them. This model provides all the information that can be obtained from our data. However, due to the big number of parameters it is not recommended. We could also use Wald test to make variable selection in our model.

4.5.4. Akaike Information Criterion (AIC)

AIC was developed by Akaike (1973) and is used to compare models given the outcome. AIC is considered as a penalized-likelihood criterion. In few words, AIC estimates the distance between the true likelihood function of the data and the likelihood function given by the model. The selection of a model is very important because an under-fitted model may ignore important information about the variability of response variable, while an over-fitted model makes poor prediction. Therefore, AIC tries to balance these disadvantages and is defined as:

$$\boldsymbol{AIC = 2K - 2log(Likelihood)}$$

where $\boldsymbol{K}$ is the number of parameter the model estimates including the intercept. The constant "$\mathbf{-2}$" stays for historical reason. The term $2K$ denotes the penalty of including many parameters in the model. The $log(Likelihood)$ term denotes the probability of taking the data given the model. Likelihood is obtained as statistical output and measures model's fit. The model with the higher likelihood fits better. By multiplying likelihood with "$\mathbf{-2}$" we make the model with the least AIC to fit better. The more parameters the model has, the greater its likelihood will be. However, the penalty term $\boldsymbol{2K}$ increases as more parameters we add in the model.

https://doi.org/10.26219/heal.aueb.5673

4.5.5. Deviance

Deviance counts the distance between candidate and saturated model. As I previously mentioned, saturated is the model with all explanatory variables and all-way interactions between them. Deviance is defined as:

$$\boldsymbol{Deviance} = -\mathbf{2}[\boldsymbol{log}(\boldsymbol{likelihood})_{\boldsymbol{M}} - \boldsymbol{log}(\boldsymbol{likelihood})_{\boldsymbol{S}}]$$

where $log(likelihood)_M, log(likelihood)_S$ is the maximized log-likelihood value for the candidate model and saturated model, respectively. Deviance approximately follows chi-squared distribution with degrees of freedom (df) equal to the number of saturated's parameters minus the number of candidate's model parameters. Deviance also provides goodness of fit test because it tests the hypothesis that all the parameters that are not included in the candidate model equals zero. P-value for that hypothesis is calculated via the right tail above the statistic (deviance) value, from chi-squared distribution given the degrees of freedom. Law value of deviance compared to df and large value of p-value ($p > 0.05$) are evidence of good fitting.

Deviance can be also used to compare different models. Let's consider two models, the first is a simple model denoted by M_0, and the second is a more complex model that includes all parameters of M_0 plus some additional parameters, denoted by M_1. Given that complex model M_1 fits with saturated, the likelihood ratio test to identify if the simple model M_0 fits with M_1 is:

$$-2\left[\log \text{likelihood}_{M_0} - \log \text{likelihood}_{M_1}\right]$$

This can be also calculated as:

$$-2\left[\log \text{likelihood}_{M_0} - \log \text{likelihood}_{M_1}\right] = \text{Deviance}_{M_0} - \text{Deviance}_{M_1}$$

Thus, we can compare models by comparing their deviances. This difference usually follows chi-squared distribution with df equal to the difference of df values for the separate models. In particular, this test's degrees of freedom df show the number of parameters that are in the complex but not in the simple model. Low test statistic value compared to df and large value of p-value are evidence of good fitting between these models.

https://doi.org/10.26219/heal.aueb.5673

4.5.6. Wald Test

Wald test is used in every parameter and tests if the value of each parameter equals zero. The test statistic is:

$$\mathbf{z} = \frac{\hat{\boldsymbol{\beta}}}{\mathbf{SE}}$$

Where $\hat{\beta}$ is the estimated value of the parameter, and SE is the standard error of $\hat{\beta}$. The test statistic approximately follows normal distribution. Similarly, z^2 approximately follows chi-squared distribution with 1 degree of freedom.

4.6. Player contribution based on models

I propose 22 Generalized linear models similar with chapter 4. Each models has Goals as response variable. As explanatory variables I used Home advantage, Attacking Team, Defensive team, and one player's attacking and defensive ability.

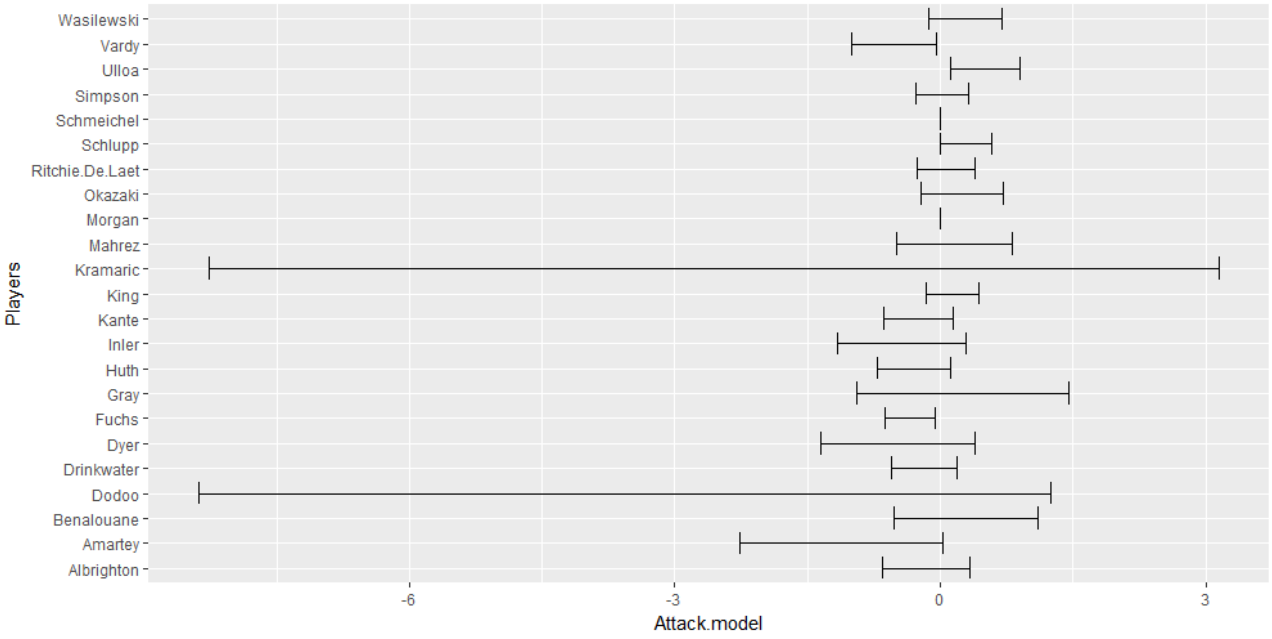


Figure 13: Errorbars for Attacking Ability Based on Model

In Figure 13 we see every player's attacking impact based on model. Dodoo and Kramaric who had the best attacking impact in descriptive, have the worst attacking impact. They also have big intervals. This is normal because they played in few matches. Striker Ulloa has the greatest attacking impact. Ulloa is the player with the most appearances as substitution. Mahrez and Okazaki who also scored many goals during season, have positive attacking

https://doi.org/10.26219/heal.aueb.5673

impact. On the other hand, Vardy who was the first scorer and main striker of the team, has negative coefficient. This fact questions the efficiency of our model.

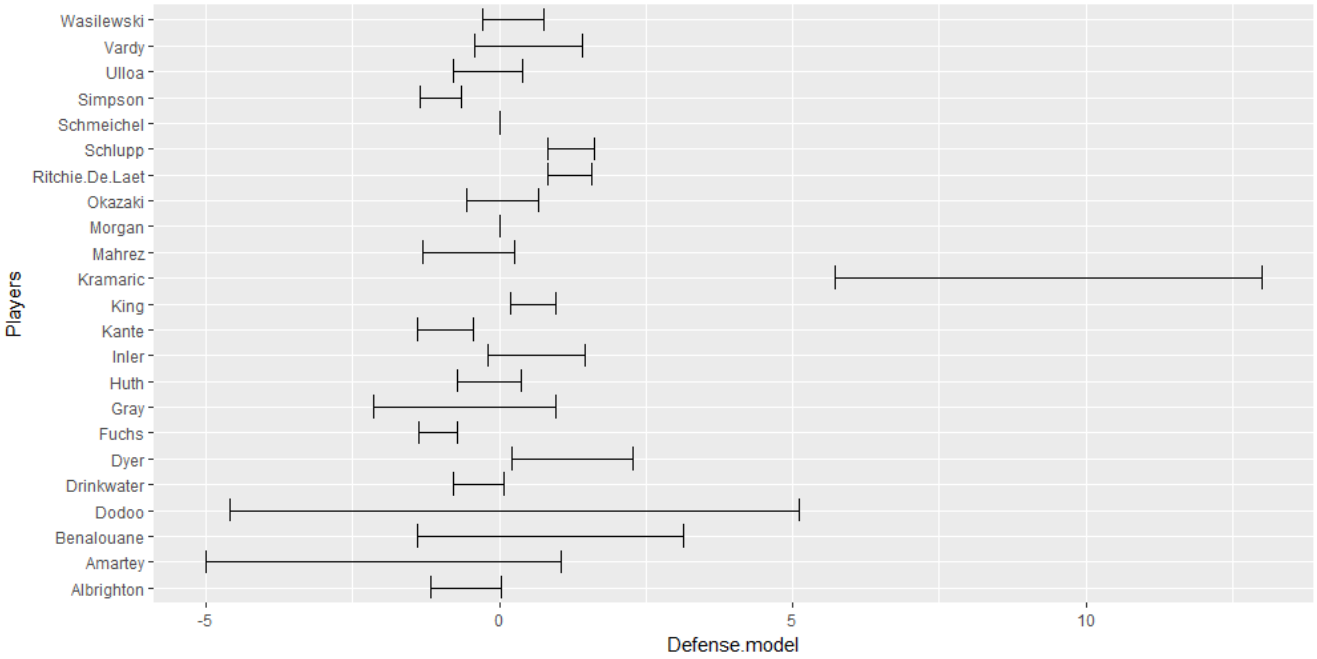


Figure 14: Errorbars for Defensive Ability Based on Model

In Figure 14 we can see every player's defensive ability based on model. Here, the approach is similar to defensive ability based on descriptives. Kramaric has the worst defensive impact. Ulloa, Simpson, Huth, Mahrez, Kante, Fuchs, Drinkwater, Albrighton, Gray, and Amartey have negative coefficient which means that they help their teams winning chances when they are in the pitch. However, Vardy who has positive coefficient, seems to have negative effect on his team's winning chances.

4.7. Descriptive Contributions vs Model Contributions

In this section I compare player's contribution based on descriptives and model. Models seem to behave better on players with little time in a match whereas descriptives overestimate their appearance. On the other hand, models seem to underestimate players with many appearances while descriptives gives better estimations.

https://doi.org/10.26219/heal.aueb.5673

4.7.1. Attack

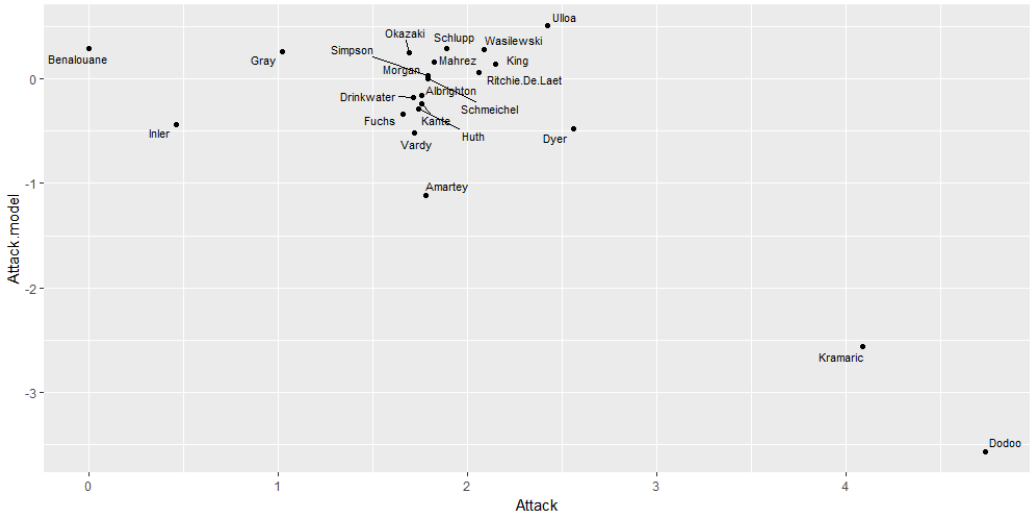


Figure 15:Comparison Between Descriptive and Model in Attack

In Figure 15 we can see that with the exception of Kramaric and Dodoo, most players have fairly even impact. However, starters such as Vardy, Drinkwater, Kante, Fuchs, and Huth have less impact based on model but higher based on descriptives than Benalouane and Gray who played in a few matches.

4.7.2. Defense

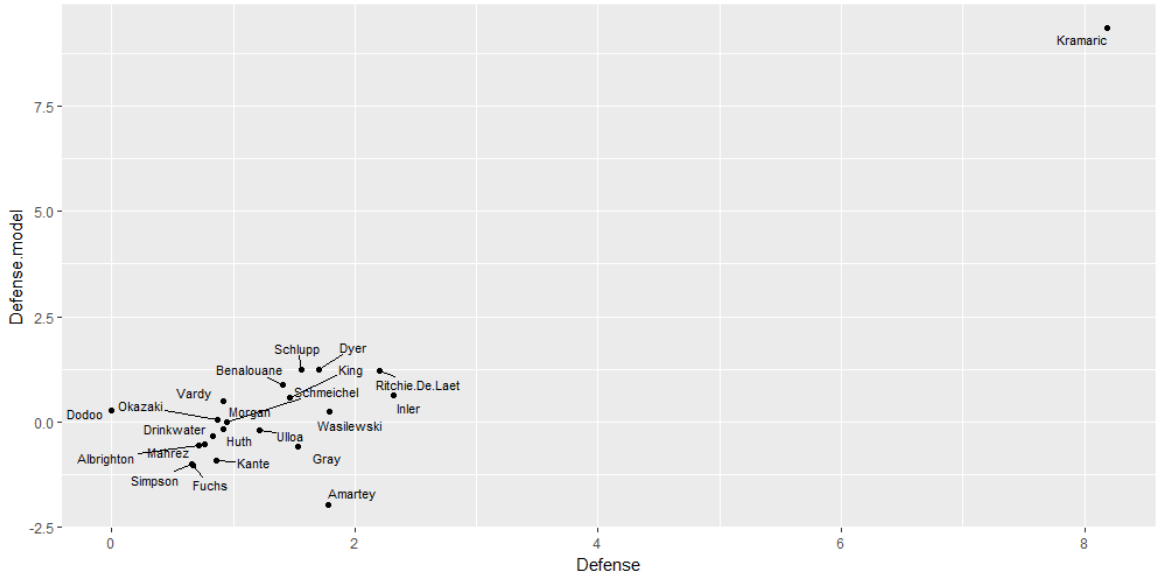


Figure 16: Comparison Between Descriptives and Model in Defense

https://doi.org/10.26219/heal.aueb.5673

In Figure 16 we can see that model and descriptives present similar results. In particular, with the exception of Amartey and Dodoo, most players that have big level of impact through model also have big level of impact through descriptives.

4.8. Teams

Last section is about teams' contribution is in goals.

4.8.1. Attack

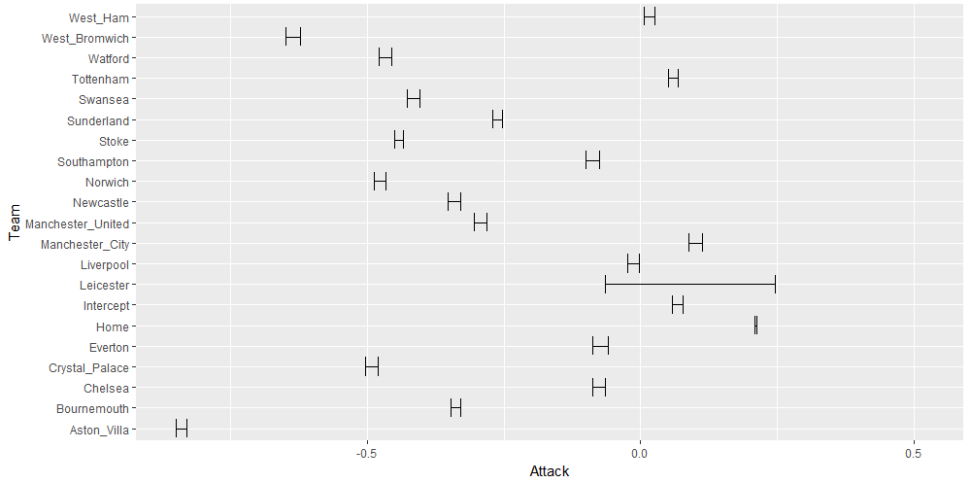


Figure 17: Errorbars for Teams Attacking Ability

In Figure 17 we can see Error-bars for the attacking ability of each Team and home advantage. Leicester shows big Intervals due to the fact that I used its' players as explanatory variables. Home advantage seems to have the biggest effect in the number of goals. As for teams, Manchester City has the best attacking impact, with Tottenham in the second place. Intercept is in the third place but includes Arsenal both attacking and defensive ability. Leicester also has fairly good attacking impact. Worst values are shown by Aston Villa, West Bromwich, Crystal Palace, Watford, and Norwich.

https://doi.org/10.26219/heal.aueb.5673

4.8.2. Defense

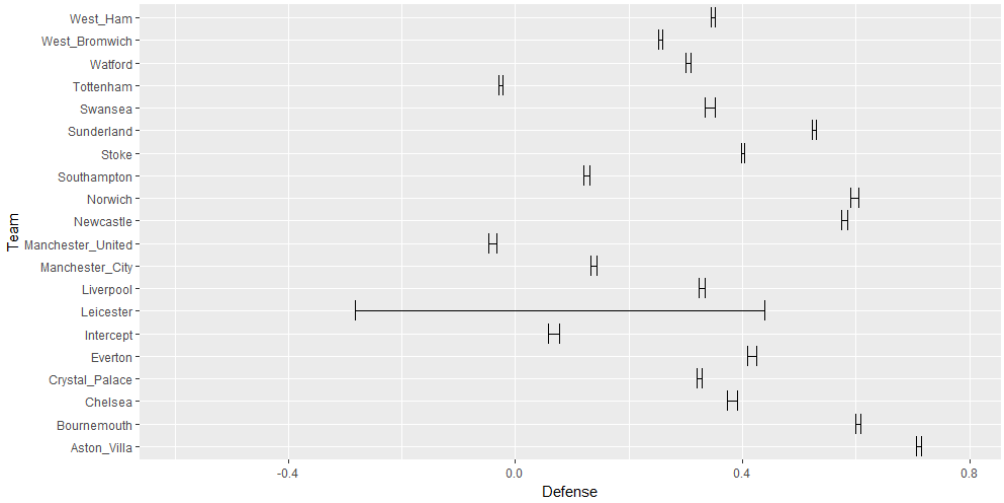


Figure 18: Errorbars for Teams Defensive Ability

In Figure 18 we can see Error-bars for the defensive ability of each Team. Leicester again shows big Intervals due to the fact that I used its' players as explanatory variables. As for the rest of the teams, Manchester United has the lowest and best defensive impact, with Tottenham in the second place. Intercept is in the third place but includes Arsenal both attacking and defensive ability. Leicester also shows fairly good defensive impact. Worst values are shown by Aston Villa, Bournemouth, Norwich, Newcastle, and Sunderland.

Next I present the final championship at Table 5, that confirms my previous Figures.

We can see Aston Villa, Norwich, Newcastle, Bournemouth, and Sunderland seem to have the worst defense. On the other hand, Manchester United, Tottenham, Arsenal, and Leicester seem to be the best defensive teams in this season. As for the number of goals scored, Manchester City, Tottenham, Leicester, Arsenal, and West Ham have the best value while Aston Villa, Norwich, West Bromwich, and Crystal Palace have poor attacking ability.

https://doi.org/10.26219/heal.aueb.5673

Pos	Team	Pld	W	D	L	GF	GA	Pts
1	Leicester City (C)	38	23	12	3	68	36	81
2	Arsenal	38	20	11	7	65	36	71
3	Tottenham Hotspur	38	19	13	6	69	35	70
4	Manchester City	38	19	9	10	71	41	66
5	Manchester United	38	19	9	10	49	35	66
6	Southampton	38	18	9	11	59	41	63
7	West Ham United	38	16	14	8	65	51	62
8	Liverpool	38	16	12	10	63	50	60
9	Stoke City	38	14	9	15	41	55	51
10	Chelsea	38	12	14	12	59	53	50
11	Everton	38	11	14	13	59	55	47
12	Swansea City	38	12	11	15	42	52	47
13	Watford	38	12	9	17	40	50	45
14	West Bromwich	38	10	13	15	34	48	43
15	Crystal Palace	38	11	9	18	39	51	42
16	AFC Bournemouth	38	11	9	18	45	67	42
17	Sunderland	38	9	12	17	48	62	39
18	Newcastle United (R)	38	9	10	19	44	65	37
19	Norwich City (R)	38	9	7	22	39	67	34
20	Aston Villa (R)	38	3	8	27	27	76	17

Table 5:Championship Table

https://doi.org/10.26219/heal.aueb.5673

https://doi.org/10.26219/heal.aueb.5673

5. Model for Leicester's Players

My dataset includes all 38 fixtures of Barclays Premier League. Each fixture includes ten matches and each match is divided in two parts. One part contains the number of goals that were scored by home team and the other part contains the number of goals that were scored by away team. Each part also divided in shifts according to in match events such as substitutions and red cards. In total, my dataset includes 952 rows. Each row also contains the home advantage, the name of the attacking team, the name of the defensive team and the number of minutes that each one of the 24 Leicester players participated in the match. Home advantage takes value 1 when the home team is attacking and 0 when the away team is attacking. I divided the number of minutes that every player participated in each shift of the match by 90. This scaling represents the proportion of the match in which each player played. As I previously mentioned, each player's column has at least 722 zero values because Leicester's players played only in the 38 out of 380 matches.

My purpose is to analyze player's contribution on goals using their participation data. I used Goals as response variable and the rest as explanatory variables. Poisson distribution seems to fit in my response variable, as I mentioned above. So, I will symbolize my response variable Y_i, where $i = 1, \dots 760$ is the number of rows. Thus:

$$\boldsymbol{Y_i \sim Poisson(\lambda), i = 1, \dots, 760}$$

$$\boldsymbol{E[Y_i] = Var[Y_i] = \lambda}$$

where the parameter λ is the mean number of goals that every team scores in a match. As link function η we use the logarithm. Therefore,

$$\boldsymbol{\eta = \log(\lambda)}$$

We know that the values of Y_i are positive integers. Thus, parameter λ will be positive, which agree with logarithm's domain. Therefore, the general model is defined as:

$$\boldsymbol{\log(\lambda_i) = \mu_i + \mathrm{Home}_i + \mathrm{Att.team}_i + \mathrm{Def.team}_i + \overrightarrow{X_{att_i}} + \overrightarrow{X_{def_i}}} \qquad \textbf{(4.1)}$$

where parameter μ_i represents the intercept term for the i^{th} row of the matrix. The parameter Home_i represents home advantage for the i^{th} row of the matrix. The factor $\mathrm{Att.team}_i$ represents the team that scores in the i^{th} row of the matrix. The factor $\mathrm{Def.team}_i$represents the team that is conceded goals in the i^{th} row of the matrix. $\overrightarrow{X_{att_i}}$ is a vector representing the

https://doi.org/10.26219/heal.aueb.5673

proportion of the match that each player played for the attacking team. Similarly, $\overrightarrow{X_{def_1}}$ is a vector representing the proportion of the match that each player played for the defensive team.

Model (4.1) has $AIC = 2436.2$ and $Deviance = 913.56$ with 870 degrees of freedom. This model fitted with saturated because $p - value > 0.05$.

At the subsequent sections, we check if the model satisfies some requirements. To begin with, we test for multi-collinearity problems. Secondly, we check normality and homogeneity of residuals.

5.1.1. Collinearity

Collinearity is the high linear relationship between one explanatory variable and the rest explanatory variables. Two or more variables may carry the same information, so there is no reason of keeping them. Multi-collinearity can also make significant effects to appear as non-significant. Furthermore, effects are very difficult to be estimated because multi-collinear variables are very close connected. Moreover, multi-collinearity creates high standard errors of coefficients, unsteadiness of estimators and distortion of the effects. Mathematical explanation to understand the problem of multi-collinearity is presented. In order to estimate model's parameters, we use the data matrix X. In particular, parameter's estimation is defined as:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y$$

However, if one column of data matrix X is a linear combination of the rest variables then the inverse $(X^T X)^{-1}$ does not exist since $|X| = 0$. Despite the fact that multi-collinearity rarely express perfect linear relationship, when variables are highly connected they have unsteady estimates and high standard errors. One way to face the problem of multi-collinearity is by removing the variable with the highest collinearity coefficient.

Collinearity test for model 4.1 is summarized in Table 6. In this Table, we clearly see that collinearity values for most variables, for the model (4.1), are below 10, meaning that the most explanatory variables are not highly associated. However, 14 out of 21 players have big collinearity coefficient in attack, where 9 of them are starters. This means that first team players have similar impact on the number of goals that Leicester scores. In defense Simpson and Huth have big collinearity coefficient.

https://doi.org/10.26219/heal.aueb.5673

	GVIF	Df	GVIF^(1/(2*Df))
Att.Ritchie.De.Laet	239.48766	1	15.475389
Def.Ritchie.De.Laet	95.177352	1	9.755888
Att.Huth	25968.46518	1	161.14734
Def.Huth	491.309172	1	22.165495
Att.Amartey	6.97362	1	2.640761
Def.Amartey	1.983793	1	1.408472
Att.Simpson	1136.312661	1	33.709237
Def.Simpson	117.169916	1	10.824505
Att.Wasilewski	2934.807954	1	54.173868
Def.Wasilewski	60.571887	1	7.782794
Att.Fuchs	409.076356	1	20.225636
Def.Fuchs	30.618367	1	5.533387
Att.Benalouane	1.000001	1	1.000001
Def.Benalouane	1.2649	1	1.124678
Att.Drinkwater	483.503085	1	21.988704
Def.Drinkwater	49.663483	1	7.047232
Att.King	180.775661	1	13.445284
Def.King	27.426051	1	5.236989
Att.Albrighton	505.425118	1	22.481662
Def.Albrighton	32.315248	1	5.68465
Att.Kante	438.916138	1	20.950325
Def.Kante	28.838928	1	5.370189
Att.Schlupp	187.305212	1	13.685949
Def.Schlupp	23.46262	1	4.843823
Att.Gray	1.760088	1	1.326683
Def.Gray	5.440995	1	2.332594
Att.Dyer	13.468308	1	3.669919
Def.Dyer	1.291675	1	1.136519
Att.Mahrez	488.284116	1	22.097152
Def.Mahrez	31.06362	1	5.573475
Att.Inler	1.673772	1	1.293743
Def.Inler	4.055061	1	2.013718
Att.Dodoo	2.135891	1	1.461469
Def.Dodoo	1	1	1
Att.Vardy	495.760359	1	22.265677
Def.Vardy	68.160491	1	8.255937
Att.Kramaric	1.346349	1	1.160323
Def.Kramaric	1.29386	1	1.137479
Att.Okazaki	343.998752	1	18.547203
Def.Okazaki	23.512649	1	4.848984

https://doi.org/10.26219/heal.aueb.5673

Att.Ulloa	178.341307	1	13.354449
Def.Ulloa	8.989565	1	2.99826

Table 6: Collinearity coefficients for players

5.1.2. Checking Residuals

We continue with checking model's requirements for the residuals. At first, we check if standardized errors follow Normal distribution. Standardized errors are defined as:

$$e_i = \frac{y_i - \widehat{\mu_i}}{\sqrt{\widehat{\mu_i}}}$$

Also, we used QQ-plots and Kologorov-Smirnov test in order to test the standardized errors.

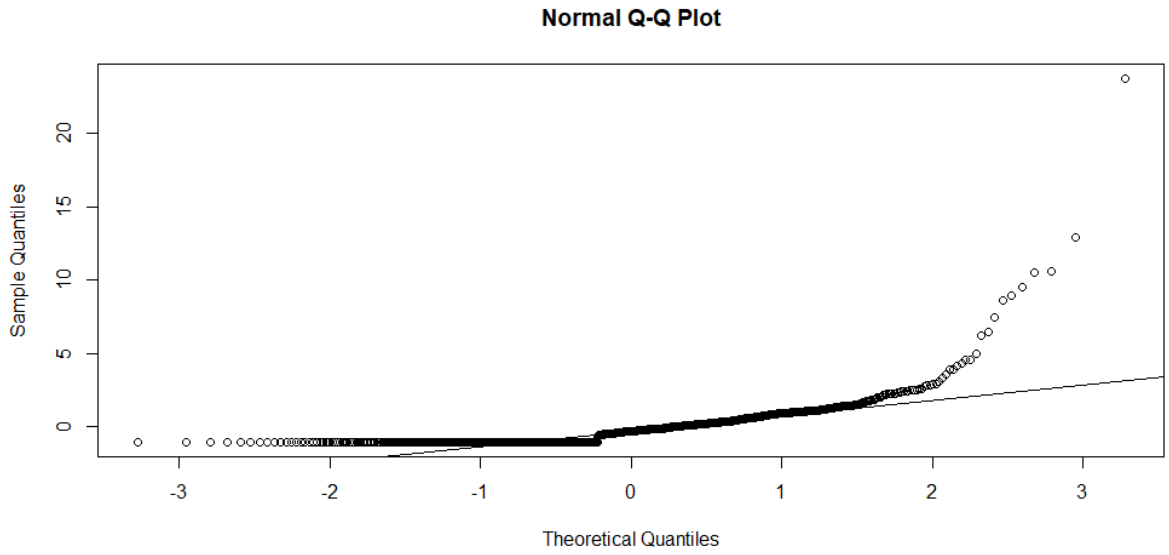


Figure 19: QQ-plot for Standardized Errors

At Figure 19 we clearly see that the points are not near the line, which means that our standardized errors are probably not Normally distributed. Kolmogorov-Smirnov test rejects the null hypothesis, which claims that model's (4.1) standardized errors were generated from Normal distribution, with $p - value < 2.2 * 10^{-16}$. However, we can overlook the fact that our errors are not normally distributed because of the large number of observations in dataset.

Next requirement we have to check is homoscedasticity of errors. Absence of homoscedasticity may cause biased estimators of coefficients; standard errors and their variance are not estimated appropriately. However, GLMs usually complies with homoscedasticity.

https://doi.org/10.26219/heal.aueb.5673

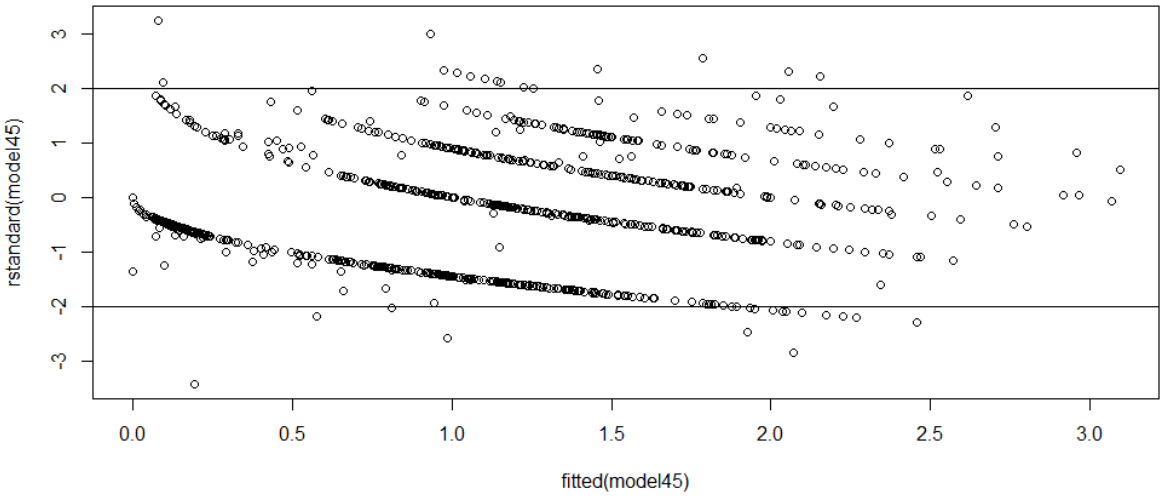


Figure 20: Fitted vs Standardized Errors

Figure 20 confirms the assumption that model's (4.1) errors show homoscedasticity, since the distance between each group of points seems to be the same. We also use Barlett test which does not reject the null hypothesis, which claims that the variance between the values is the same, with $p - value = 0.4206$.

5.1.3. Contribution Analysis for Players

Last section includes the statistical analysis of player's contribution according to model (4.1) and are presented at the next table. As we can see, Huth and Wasilewski showed extreme values in both attacking and defensive coefficients. This is resulted by multi-collinearity. In particular, Huth played in 35 whole games, so his performance is underestimated in order to have better estimations from the rest of the team. The same approach is followed for Wasilewski who played in the rest 3 games. Benalouane has extreme negative value in attack due to the fact that Leicester did not score any goal while he was in the pitch. Similarly, Dodoo has extreme negative defensive coefficient because Leicester did not conceded any goal while Dodoo was in the pitch. However, Dodoo played only in one match as a substitution. Scmeichel and Morgan played in all matches. So, my model could not estimate their performance and I decided to remove them from the explanatory variables.

https://doi.org/10.26219/heal.aueb.5673

Players	Attack	SE.Attack	Defense	SE.Defense
Ritchie.De.Laet	0.374	7.16	7.29	4.80
Huth	-9.04	32.9	14.8	7.57
Amartey	-1.79	7.64	8.52	4.92
Simpson	1.13	7.18	8.27	4.63
Wasilewski	-7.71	32.9	14.7	7.48
Fuchs	0.382	4.42	-4.66	2.50
Benalouane	-505	17100	1.38	14.4
Drinkwater	2.97	4.59	-5.16	2.54
King	2.58	4.49	-2.65	2.28
Albrighton	1.58	4.63	-2.37	1.94
Kante	1.98	4.47	-2.86	2.15
Schlupp	1.01	4.32	-0.782	1.94
Gray	-4.27	7.94	-5.44	9.03
Dyer	5.12	4.40	-2.07	3.91
Mahrez	0.823	4.37	-3.88	1.88
Inler	-4.04	6.81	0.0892	2.03
Dodoo	6.00	6.93	-78.6	16400
Vardy	1.09	4.55	-2.44	2.75
Kramaric	9.51	9.14	9.27	6.25
Okazaki	2.00	4.27	-1.72	1.74
Ulloa	2.52	4.30	-2.25	2.25

Table 7: Contribution Analysis for Players

The following Figure shows the attacking impact of Leicester's players. With the exception of Huth, the rest of the starting team has almost the same attacking impact in Leicester's

https://doi.org/10.26219/heal.aueb.5673

attacking ability. Kramaric, Dyer, Dodoo, Ulloa and King seem to have the best attacking impact on their team. However, differences between most of attacking coefficients are little.

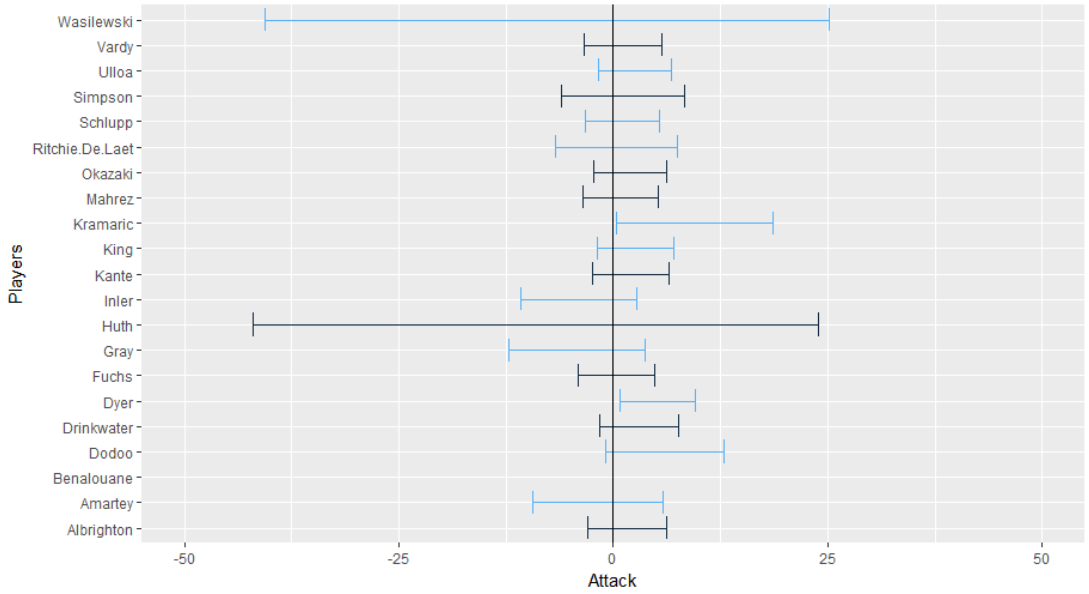


Figure 21: Attacking impact of Leicester's players.

Next Figure shows the defensive impact of Leicester's players. As we can see all 6 players with the worst defensive coefficient are defenders. Moreover, starters seem to have the best defensive coefficient. This means that Leicester's defense was the key factor of the championship run.

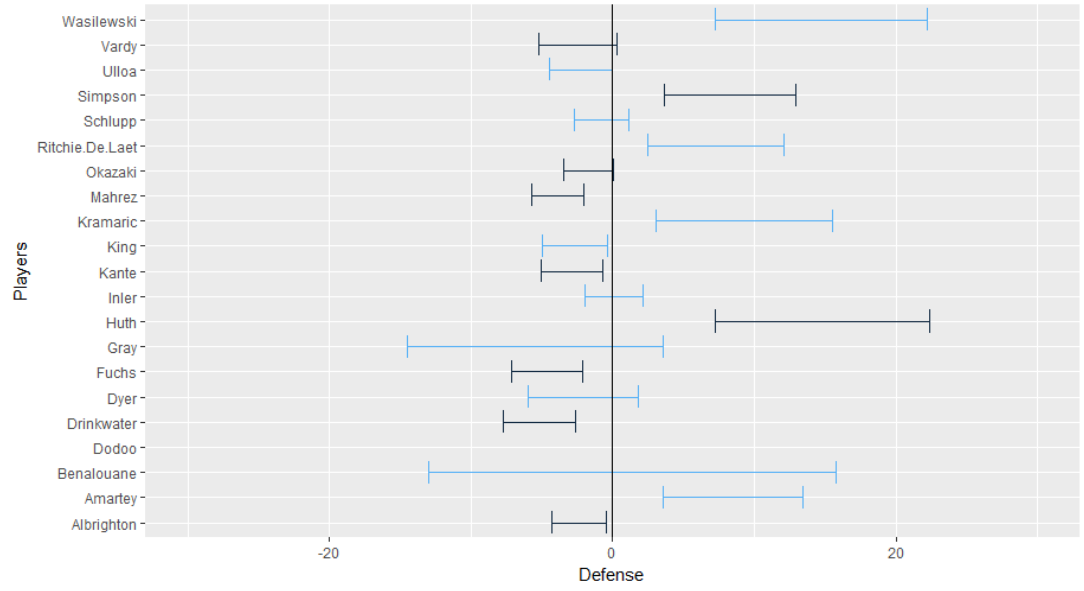


Figure 22: Defensive impact of Leicester's players

https://doi.org/10.26219/heal.aueb.5673

5.1.4. Contribution of Teams

Last I present the coefficients for the Teams that play in each match. I used the sum to zero method which helped us identify the impact for all Teams. As base, Team I used West Ham. The following Table shows each Team's attacking and Defensive coefficient.

Teams	Attack	SE.Attack	Defense	SE.Defense
Arsenal	-0.0185	0.178	-0.348	0.219
Aston_Villa	-0.880	0.233	0.368	0.182
Bournemouth	-0.319	0.196	0.234	0.188
Chelsea	-0.0626	0.181	0.0385	0.198
Crystal_Palace	-0.473	0.204	-0.0243	0.199
Everton	-0.0884	0.183	0.0140	0.200
Leicester	-2.54	0.355	-2.41	0.412
Liverpool	-0.0184	0.179	-0.0422	0.201
Manchester_City	0.113	0.173	-0.198	0.211
Manchester_United	-0.253	0.191	-0.414	0.221
Newcastle	-0.337	0.196	0.201	0.188
Norwich	-0.475	0.205	0.269	0.188
Southampton	-0.0721	0.181	-0.255	0.213
Stoke	-0.457	0.203	0.0622	0.195
Sunderland	-0.282	0.191	0.161	0.190
Swansea	-0.402	0.199	-0.0794	0.202
Tottenham	0.0544	0.174	-0.381	0.221
Watford	-0.449	0.202	-0.0312	0.200
West_Bromwich	-0.650	0.215	-0.141	0.206
West_Ham	-0.401	0.202	-0.156	0.212

Table 8: Contribution of Teams

The following Figure compares the attacking impact for all Teams according to model (4.1). As we can see Manchester City, Tottenham and Arsenal scored more than the average number of goals. On the other hand, Aston Villa, West Bromwich, and Crystal Palace seem to be the least productive Teams of the league. It is important to mention that Leicester coefficient is the worst among Teams due to the fact that Leicester's players are also included in the model.

https://doi.org/10.26219/heal.aueb.5673

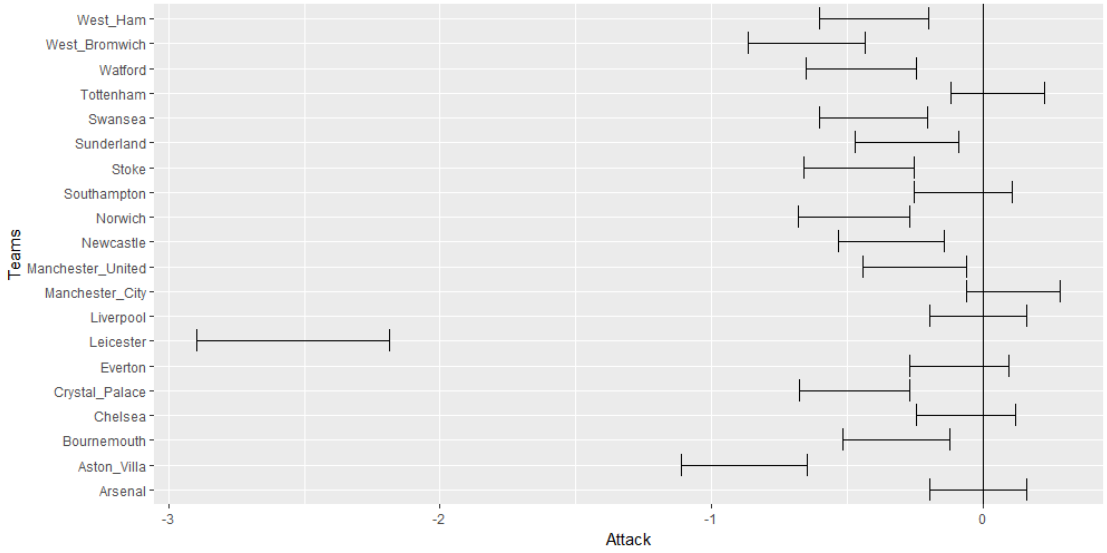


Figure 23: Attacking coefficient for Teams

The following Figure compares the defensive impact for all Teams according to model (4.1). As we can see Manchester United, Tottenham, Arsenal, Southampton, and Manchester City conceded less than the average number of goals. On the other hand, Aston Villa, Bournemouth, Norwich, and Newcastle conceded more than the average number of goals. It is important to mention that Leicester coefficient is the best among Teams due to the fact that Leicester's players are also included in the model.

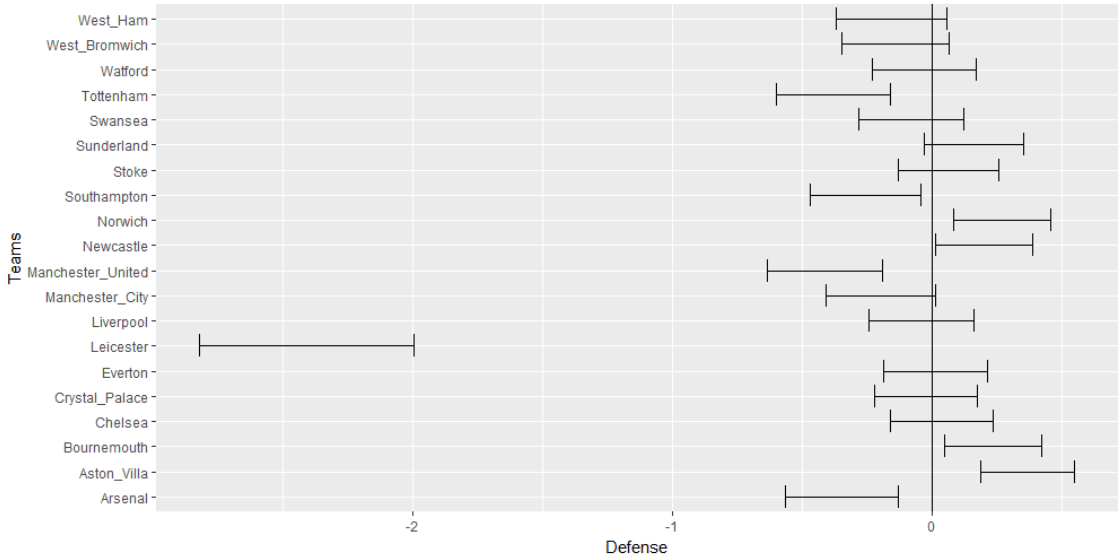


Figure 24: Defensive coefficient for Teams

https://doi.org/10.26219/heal.aueb.5673

https://doi.org/10.26219/heal.aueb.5673

6. Conclusions and Future Research

6.1. Conclusions

In my thesis, I tried to evaluate player's contribution using Poisson generalized models. First, I presented sport fields where statistic has developed. In prediction and performance analysis most of the models are used to model goals. However, in sport performance we care more about player's contribution on goals than to estimate the number of goals. I also present papers about sports economics. In this field, researchers develop indices to estimate the competitive balance for each robin-round tournament. Training is also a field that statistic knows great development since high load of training may cause serious injuries to players.

Next, I present information about the dataset using R. Leicester's coach used close rotation in the positions of goalkeeper and central defender. However, the rotation was wider in the rest positions.

First I estimated the attacking and defensive ability through descriptives. Results shown that Kramaric and Dodoo who played in few matches, have the greatest attacking and defensive ability. Last I proposed 22 models for each player to estimate each player's and team's attacking and defensive impact. Ulloa, Wasilewski, and Schlupp had the biggest attacking ability while Amartey, Fuchs, and Simpson had the best defensive ability. As for the teams, Leicester, Tottenham, and Manchester City were the best attacking teams while Leicester, Tottenham, and Manchester United were the best defensive teams.

Next I divide each game into shift according to events such as substitutions and red cards. Each player that played on each shift, had as count the the number of minutes of the shift divided by 90. This way my dataset had further information about the impact of each player in Leicester's goals. I presented a model for all players where I used sum to zero method in order to obtain coefficient for all Teams. Then, I checked model requirements using plots and tests. Due to high level of collinearity, some results where extreme. Moreover, the model fits with the saturated. Last, the results about Teams' impact on the number of goals were similar with the results of the previous models.

6.2. Future Research

As an improvement of this work, someone can try to spend more time in explanatory variables' distribution using Bayesian approach. This will help them find a more suitable

https://doi.org/10.26219/heal.aueb.5673

model and face collinearity without removing variables. This way, they can estimate the contribution of all players and not the most significant according to other methods. With the use of variables' distribution they can broaden their limits by using other data such as wages etc. Then, they can estimate if the high paid players worth their money or overpaid. Similarly, they can estimate if low paid player are undervalued by their team. This is more important for the less challenging team because by obtaining undervalued players they became more challenging.

Someone can also change the distribution of the response variable. Negative binomial can also describe variable goal. This distribution has the advantage of flexible variance. This approach can lead to different deviance values and, fit better with the saturated model. Further investigation also could be done according to the level of interaction that usually is shown in team sports. As we saw some players have high levels of correlation. This means that models that include two-way or three-way interactions might lead in better results than my model. Last but not least someone can use a different type of count that will be more fair for striker since number of minutes seem to undervalue their performance. One type of count could be the number of shots or the number of minutes that each player has the ball etc.

Someone can also include in their data the exact score when a player enters as a substitution. This way they can identify player's effect on his team chances of winning. This approach is similar with the one followed in basketball's metrics.

https://doi.org/10.26219/heal.aueb.5673

Appendix A

I am going to present frequency tables for Leicester's player except of Schmeichel and Morgan who played in all matches.

Ritchie De Laet				
value	N	raw %	valid %	cumulative %
0	27	71.05	71.05	71.05
7	1	2.63	2.63	73.68
10	1	2.63	2.63	76.32
16	1	2.63	2.63	78.95
31	1	2.63	2.63	81.58
66	1	2.63	2.63	84.21
75	1	2.63	2.63	86.84
90	5	13.16	13.16	100
missings	0	0		

Table 9: Frequency table of Ritchie De Laet

Huth				
value	N	raw %	valid %	cumulative %
0	3	7.89	7.89	7.89
90	35	92.11	92.11	100
missings	0	0		

Table 10: Frequency Table of Huth

Amartey				
value	N	raw %	valid %	cumulative %
0	34	89.47	89.47	89.47
5	1	2.63	2.63	92.11
6	1	2.63	2.63	94.74
12	1	2.63	2.63	97.37
78	1	2.63	2.63	100
missings	0	0		

Table 11: Frequency table of Amartey

https://doi.org/10.26219/heal.aueb.5673

Simpson				
value	N	raw %	valid %	cumulative %
0	8	21.05	21.05	21.05
54	1	2.63	2.63	23.68
65	1	2.63	2.63	26.32
74	1	2.63	2.63	28.95
80	1	2.63	2.63	31.58
90	26	68.42	68.42	100
missings	0	0		

Table 12: Frequency table of Simpson

Wasilewski				
value	N	raw %	valid %	cumulative %
0	34	89.47	89.47	89.47
32	1	2.63	2.63	92.11
90	3	7.89	7.89	100
missings	0	0		

Table 13: Frequency table of Wasilewski

Fuchs				
value	N	raw %	valid %	cumulative %
0	6	15.79	15.79	15.79
8	1	2.63	2.63	18.42
13	1	2.63	2.63	21.05
79	1	2.63	2.63	23.68
90	29	76.32	76.32	100
missings	0	0		

Table 14: Frequency table of Fuchs

Benalouane				
value	N	raw %	valid %	cumulative %
0	35	92.11	92.11	92.11
15	1	2.63	2.63	94.74
24	1	2.63	2.63	97.37
25	1	2.63	2.63	100
missings	0	0		

Table 15: Frequency table of Benalouane

https://doi.org/10.26219/heal.aueb.5673

Drinkwater				
value	N	raw %	valid %	cumulative %
0	4	10.53	10.53	10.53
17	1	2.63	2.63	13.16
78	1	2.63	2.63	15.79
80	2	5.26	5.26	21.05
86	1	2.63	2.63	23.68
88	1	2.63	2.63	26.32
90	28	73.68	73.68	100
missings	0	0		

Table 16: Frequency table of Drinkwater

King				
value	N	raw %	valid %	cumulative %
0	14	36.84	36.84	36.84
2	1	2.63	2.63	39.47
3	2	5.26	5.26	44.74
4	1	2.63	2.63	47.37
7	1	2.63	2.63	50
10	1	2.63	2.63	52.63
11	1	2.63	2.63	55.26
13	3	7.89	7.89	63.16
20	1	2.63	2.63	65.79
21	1	2.63	2.63	68.42
45	3	7.89	7.89	76.32
73	1	2.63	2.63	78.95
90	8	21.05	21.05	100
missings	0	0		

Table 17: Frequency table of King

https://doi.org/10.26219/heal.aueb.5673

Albrighton				
value	N	raw %	valid %	cumulative %
8	1	2.63	2.63	2.63
10	1	2.63	2.63	5.26
13	1	2.63	2.63	7.89
45	2	5.26	5.26	13.16
54	1	2.63	2.63	15.79
63	2	5.26	5.26	21.05
64	2	5.26	5.26	26.32
65	1	2.63	2.63	28.95
67	1	2.63	2.63	31.58
70	1	2.63	2.63	34.21
71	2	5.26	5.26	39.47
75	1	2.63	2.63	42.11
76	1	2.63	2.63	44.74
78	1	2.63	2.63	47.37
79	1	2.63	2.63	50
83	1	2.63	2.63	52.63
84	1	2.63	2.63	55.26
85	1	2.63	2.63	57.89
86	1	2.63	2.63	60.53
88	1	2.63	2.63	63.16
90	14	36.84	36.84	100
missings	0	0		

Table 18: Frequency table of Albrighton

Kante				
value	N	raw %	valid %	cumulative %
0	1	2.63	2.63	2.63
8	1	2.63	2.63	5.26
12	1	2.63	2.63	7.89
26	1	2.63	2.63	10.53
28	1	2.63	2.63	13.16
70	1	2.63	2.63	15.79
86	1	2.63	2.63	18.42
90	31	81.58	81.58	100
missings	0	0		

Table 19: Frequency table of Kante

Schlupp

https://doi.org/10.26219/heal.aueb.5673

value	N	raw %	valid %	cumulative %
0	15	39.47	39.47	39.47
4	1	2.63	2.63	42.11
20	1	2.63	2.63	44.74
21	1	2.63	2.63	47.37
23	1	2.63	2.63	50
25	1	2.63	2.63	52.63
27	1	2.63	2.63	55.26
36	1	2.63	2.63	57.89
45	4	10.53	10.53	68.42
77	2	5.26	5.26	73.68
82	1	2.63	2.63	76.32
90	9	23.68	23.68	100
missings	0	0		

Table 20: Frequency table of Schlupp

Gray				
value	N	raw %	valid %	cumulative %
0	27	71.05	71.05	71.05
2	1	2.63	2.63	73.68
5	1	2.63	2.63	76.32
9	1	2.63	2.63	78.95
11	3	7.89	7.89	86.84
13	1	2.63	2.63	89.47
17	1	2.63	2.63	92.11
23	1	2.63	2.63	94.74
29	1	2.63	2.63	97.37
45	1	2.63	2.63	100
missings	0	0		

Table 21: Frequency table of Gray

https://doi.org/10.26219/heal.aueb.5673

Dyer				
value	N	raw %	valid %	cumulative %
0	28	73.68	73.68	73.68
1	1	2.63	2.63	76.32
4	1	2.63	2.63	78.95
6	1	2.63	2.63	81.58
11	1	2.63	2.63	84.21
14	1	2.63	2.63	86.84
19	1	2.63	2.63	89.47
21	1	2.63	2.63	92.11
45	3	7.89	7.89	100
missings	0	0		

Table 22: Frequency table of Dyer

Mahrez				
value	N	raw %	valid %	cumulative %
0	1	2.63	2.63	2.63
45	2	5.26	5.26	7.89
58	1	2.63	2.63	10.53
69	1	2.63	2.63	13.16
77	2	5.26	5.26	18.42
78	1	2.63	2.63	21.05
79	2	5.26	5.26	26.32
80	2	5.26	5.26	31.58
81	1	2.63	2.63	34.21
82	2	5.26	5.26	39.47
83	1	2.63	2.63	42.11
84	1	2.63	2.63	44.74
85	1	2.63	2.63	47.37
86	1	2.63	2.63	50
88	1	2.63	2.63	52.63
89	1	2.63	2.63	55.26
90	17	44.74	44.74	100
missings	0	0		

Table 23: Frequency table of Mahrez

https://doi.org/10.26219/heal.aueb.5673

Inler				
value	N	raw %	valid %	cumulative %
0	33	86.84	86.84	86.84
8	1	2.63	2.63	89.47
10	1	2.63	2.63	92.11
45	1	2.63	2.63	94.74
64	1	2.63	2.63	97.37
67	1	2.63	2.63	100
missings	0	0		

Table 24: Frequency table of Inler

Dodoo				
value	N	raw %	valid %	cumulative %
0	37	97.37	97.37	97.37
19	1	2.63	2.63	100
missings	0	0		

Table 25: Frequency table of Dodoo

Vardy				
value	N	raw %	valid %	cumulative %
0	2	5.26	5.26	5.26
56	1	2.63	2.63	7.89
69	1	2.63	2.63	10.53
71	1	2.63	2.63	13.16
77	1	2.63	2.63	15.79
82	1	2.63	2.63	18.42
87	1	2.63	2.63	21.05
88	1	2.63	2.63	23.68
90	29	76.32	76.32	100
missings	0	0		

Table 26: Frequency table of Vardy

Kramaric				
value	N	raw %	valid %	cumulative %
0	36	94.74	94.74	94.74
10	1	2.63	2.63	97.37
12	1	2.63	2.63	100
missings	0	0		

Table 27: Frequency table of Kramaric

https://doi.org/10.26219/heal.aueb.5673

Okazaki				
value	N	raw %	valid %	cumulative %
0	2	5.26	5.26	5.26
2	2	5.26	5.26	10.53
17	1	2.63	2.63	13.16
25	1	2.63	2.63	15.79
27	1	2.63	2.63	18.42
45	7	18.42	18.42	36.84
59	2	5.26	5.26	42.11
60	1	2.63	2.63	44.74
61	1	2.63	2.63	47.37
62	4	10.53	10.53	57.89
63	1	2.63	2.63	60.53
64	2	5.26	5.26	65.79
65	1	2.63	2.63	68.42
67	1	2.63	2.63	71.05
69	2	5.26	5.26	76.32
70	1	2.63	2.63	78.95
73	1	2.63	2.63	81.58
76	1	2.63	2.63	84.21
77	1	2.63	2.63	86.84
81	1	2.63	2.63	89.47
87	1	2.63	2.63	92.11
90	3	7.89	7.89	100
missings	0	0		

Table 28: Frequency table of Okazaki

https://doi.org/10.26219/heal.aueb.5673

Ulloa				
value	N	raw %	valid %	cumulative %
0	11	28.95	28.95	28.95
3	1	2.63	2.63	31.58
9	1	2.63	2.63	34.21
12	1	2.63	2.63	36.84
14	1	2.63	2.63	39.47
15	1	2.63	2.63	42.11
19	1	2.63	2.63	44.74
20	1	2.63	2.63	47.37
21	2	5.26	5.26	52.63
23	1	2.63	2.63	55.26
26	4	10.53	10.53	65.79
27	1	2.63	2.63	68.42
28	3	7.89	7.89	76.32
30	1	2.63	2.63	78.95
31	1	2.63	2.63	81.58
45	1	2.63	2.63	84.21
69	1	2.63	2.63	86.84
73	1	2.63	2.63	89.47
79	1	2.63	2.63	92.11
87	1	2.63	2.63	94.74
90	2	5.26	5.26	100
missings	0	0		

Table 29: Frequency table of Ulloa

https://doi.org/10.26219/heal.aueb.5673

	R.D.Laet	R.Huth	D.Amartey	D.Simpson	M.Wasilewski	C.Fuchs	Y.Benalouane	D.Drinkwater	A.King	M.Albrighton
R.D.Laet										
R.Huth	0.134									
D.Amartey	-0.111	0.062								
D.Simpson	-0.881	-0.172	-0.281							
M.Wasilewski	-0.15	-0.978	-0.068	0.156						
C.Fuchs	-0.893	-0.154	0.111	0.811	0.17					
Y.Benalouane	0.22	0.084	-0.06	-0.342	-0.093	-0.242				
D.Drinkwater	0.152	0.478	0.09	-0.2	-0.458	-0.169	0.122			
A.King	0.356	-0.387	-0.058	-0.342	0.363	-0.375	0.258	-0.459		
M.Albrighton	-0.041	0.307	0.097	-0.048	-0.288	0.029	0.219	-0.143	0.02	
NG.Kante	-0.517	-0.12	-0.04	0.549	0.133	0.52	-0.356	-0.147	-0.499	-0.076
J.Schlupp	0.636	0.107	-0.077	-0.608	-0.139	-0.738	0.371	0.206	0.227	-0.41
D.Gray	-0.241	-0.316	-0.09	0.203	0.399	0.244	-0.138	0.196	-0.104	-0.549
N.Dyer	0.081	-0.018	-0.093	-0.017	0.003	-0.08	-0.126	-0.362	-0.111	0.043
R.Mahrez	0	-0.066	0.089	0.068	0.02	-0.046	-0.522	0.087	-0.005	-0.197
G.Inler	0.37	0.093	-0.067	-0.298	-0.103	-0.327	-0.091	0.059	-0.138	-0.035
J.Dodoo	0.366	0.048	-0.035	-0.312	-0.053	-0.33	-0.047	0.07	0.286	-0.013
J.Vardy	0.163	-0.09	0.034	-0.183	0.101	-0.165	0.076	-0.035	0.11	0.61
A.Kramaric	0.229	0.069	-0.05	-0.18	-0.076	-0.201	-0.067	-0.282	0.248	0.031
S.Okazaki	0.15	-0.132	0.1	-0.201	0.139	-0.007	0.198	-0.005	0.176	-0.065
L.Ulloa	-0.235	0.163	-0.084	0.291	-0.193	0.123	-0.185	-0.011	-0.189	-0.223

https://doi.org/10.26219/heal.aueb.5673

	NG.Kante	J.Schlupp	D.Gray	N.Dyer	R.Mahrez	G.Inler	J.Dodoo	J.Vardy	A.Kramaric	S.Okazaki
R.D.Laet										
R.Huth										
D.Amartey										
D.Simpson										
M.Wasilewski										
C.Fuchs										
Y.Benalouane										
D.Drinkwater										
A.King										
M.Albrighton										
NG.Kante										
J.Schlupp	-0.394									
D.Gray	0.092	-0.067								
N.Dyer	-0.053	0.051	-0.2							
R.Mahrez	-0.134	-0.171	-0.035	-0.068						
G.Inler	-0.193	0.16	-0.153	0.234	0.169					
J.Dodoo	0.067	0.235	-0.079	-0.072	-0.336	-0.052				
J.Vardy	-0.113	-0.206	-0.304	0.105	-0.126	0.112	0.058			
A.Kramaric	0.096	0.08	-0.113	0.074	0.068	-0.075	-0.039	-0.025		
S.Okazaki	-0.274	0.042	0.151	-0.186	-0.091	-0.257	-0.063	-0.178	0.014	
L.Ulloa	0.216	-0.082	0.005	-0.046	0.263	0.01	-0.154	-0.518	-0.015	-0.488

Table 30: Correlation for player's participation time

https://doi.org/10.26219/heal.aueb.5673

https://doi.org/10.26219/heal.aueb.5673

Bibliography

Avalos, M., Hellard, P., and Chatard, J. C. (2003). Modeling the training – performance relationship using a mixed model in elite swimmers, *Medicine and Science in Sports and Exercise*, 35,838-846.

Bannister, E. W., Carter, J. B. and Zarkadas, P.C. (1999). Training Theory and Taper: Validation in Triathlon Athletes, *Eur J Appl Physiol*, 79, 182-191.

Bashuk, M. (2012). Using Cumulative Win Probability to Predict NCAA Basketball Performance, *Sloan Sports Analytics Conference*, 1–10.

Caraffa, A., Cerulli, G., Projetti, M., Aisa, and G., Rizzo, A. (1996). Prevention of anterior cruciate ligament injuries in soccer, *Knee surgery Sports Traumatol, Anthroscopy*, 4, 19-21.

Clarke, R., and Norman, J.M. (1995). Home ground advantage of individual clubs in English soccer, *The Statician*, 44, 509-521.

Depken, G.A., II (1999). Free-Agency and the Competitiveness of Major League Baseball, *Review of industrial Organization*, 14, 205-217.

Deshpande, S.K., and Jensen, S.T. (2016). Estimating an NBA player's impact on his team's chances of winning, *Journal of Quantitative Analysis in Sports*, 12, 51-72.

Dixon, M.J., and Coles, S.G. (1997). Modelling Association Football Scores and Inefficiencies in the Football Betting Market, *Appl. Stat*, 46, 265-280.

Gabbett, T.J., Domrow, N. (2007). Relationships between training load, injury, and fitness in sub-elite collision sport athletes, *Journal of Sports Sciences*, 25, 1507-1519.

Goossens, K. (2006). Competitive balance in European football: comparison by adapting measures: national measure of seasonal imbalance and TOP 3, *Rivista di Diritto ed Economia Dello Sport*, **II**, 77–122.

Gramacy, R.B., Jensen, S.T., and Taddy, M. (2013). Estimating player contribution in hockey with regularized logistic regression, *Journal of Quantitative Analysis in Sports*, 9, 97-111.

Hollinger, J. (2004). Pro Basketball Forecast, *Washington, DC: Brassey's Sports*.

Karlis, D., and Ntzoufras, I. (2000). On Modelling Data, *Student*, 3 (3), 229-244.

https://doi.org/10.26219/heal.aueb.5673

Karlis, D., and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models, *The Statician*, 52, 381-393.

Karlis, D., and Ntzoufras, I. (2005). Bayesian modelling of football outcomes:Using the Skellam's distribution for the goal difference, *IMA Journal of Numerical Analysis*.

Karlis, D., and Ntzoufras, I. (2006). Bayesian analysis of the differences of count data, *Statistics in Medicine*, 25, 1885-1905.

Karlis, D., and Ntzoufras, I. (2011). Robust fitting of football prediction models, *IMA Journal of Management Mathematics*, 22, 171−182.

Kuonen, D. (1997). Statisitcal Models for Knock-out Soccer Tournaments.

Lee, A. J. (1997). Modeling Scores in the Premier League: Is Manchester United Really the Best?, *CHANCE*, 10:1, 15-19.

Maher, M. J. (1982). Modelling Association Football Scores, *Statistica Neerlandica*, 36, 109-118.

Manasis, V., Avgerinou, V., and Ntzoufras, I. (2012). A New Approach for the Meassurement of Seasonal Competitive Balance, *Series of Sports Economics*, 13.

Manasis, V., Avgerinou, V., Ntzoufras, I., and Reade, J.J. (2011). Measurement of competitive balance in professional team sports using the Normalized Concentration Ratio, *Economics Bulletin*, (3), 31.

Manasis, V., Avgerinou, V., Ntzoufras, I., and Reade, J.J. (2012). Quantification of competitive balance in European football: development of specially designed indices, *IMA Journal of Management Mathematics*, 1-13, 117-134.

Manasis, V., Ntzoufras, I. (2013). Between-seasons competitive balance in European football: review of existing and development of specially designed indices *De Gruyter* 139-152.

McHale, I.G., Szczepanski, L. (2014). A mixed effects model for identifying goal scoring ability of footballers, *Journal of the Royal Statistical Society*, 397-417.

McHale, I.G., Scarf,P.E., Folker,D.E. (2012). On the Development of a Soccer Player Performance Rating System for the English Premier League, *Interfaces*, 42, 339-351.

Mills, E. G. and Mills, H. D. (1970). Player Win Averages: A Complete Guide to Winning Baseball Players, *The Harlan D. Mills Collection*.

https://doi.org/10.26219/heal.aueb.5673

Owen, P.D., Ryan, M., Weathrston, C.R. (2007). Measuring Competitive Balance in Professional Team Sports Using the Herfindahl-Hirschman Index, *Rev. Ind. Org.*, 31, 289-302.

Saebo, O.D., Hvattum, L.M. (2017). Evaluating the efficiency of the association football transfer market using regression based player ratings.

Shahtahmassebi, G., Moyeed, R., (2016). An application of the generalized Poisson difference distribution to the Bayesian modeling of football scores, *Statistica Neerlandica.*

Stern, H. (1994). A Brownian Motion Model for the Progress of Sports Scores, *Journal of the American Statistical Association*, 89,1128–1134.

Studeman, D. (2004). The One About Win Probability, (http://www.hardballtimes.com/the-one-about-win-probability/).