Strathmore
UNIVERSITY

STRATHMORE UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

BBS-FINANCIAL ECONOMICS

084541 ANDREW ODERA ODUOR LUJAN

OPTIMIZATION OF SOCCER SUBSTITUTIONS; A PROPOSED DECISION RULE.

A Case of the Kenyan Premier League

Strathmore
UNIVERSITY

)PTIMIZATION OF SOCCER SUBSTITUTIONS; A Proposed Decision Rule

1 ***Case Study on the Kenyan Premier League***

)84541 Andrew Odera Oduor Lujan

Submitted in partial fulfillment of the requirements for the Degree of

BBS- Financial Economics at Strathmore University

Strathmore Institute of Mathematical Science

Strathmore University

Nairobi, Kenya

June, 2017

DECLARATION

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the Research Proposal contains no material previously published or written by another person except where due reference is made in the Research Proposal itself.

ANDREW ODERA ODUOR ..... [*Name of Candidate*]

..... [*Signature*]

30TH NOV 2017 ..... [*Date*]

This Research Proposal has been submitted for examination with my approval as the Supervisor.

Paul Ochiang ..... [*Name of Supervisor*]

..... [*Signature*]

30th November 2017 ..... [*Date*]

STRATHMORE UNIVERSITY
30 NOV 2017
DEAN OF STUDENTS

Strathmore Institute of Mathematical Sciences

Strathmore University

DEDICATION

This project is dedicated to Andrew Lujan and my family, who without their constant encouragement, splendid example and utter belief in me, this and many other of life's projects would not have even started. Thank you for your faith in me, your providence and the sense of direction that you readily give.

ACKNOWLEDGEMENTS

An undertaking of this nature is not done alone and while it is impossible to individually thank everyone let me nevertheless try to acknowledge a few special personalities.

All is vain without thanking the Almighty God for giving me the health, strength and brilliance. I acknowledge my supervisor, Dr. Paul Ochieng', Dean of Students, Strathmore University. This research is a reflection of not only my work but has the indelible hand print of your care, concern and utter professionalism. Thank you.

Liz Kalu, Edwin Munyui, Julie Songok and Henry Odera you all helped me in ways and means that are immeasurable. Thank you for your prayers as well as the small things that you did for me during the course of my studies. My mother, Jane Wangeci, thank you for the support and care you have shown during my undergraduate studies. Lastly my classmates with whom we shared all the joys and tears of the undergraduate program I acknowledge your contribution to this final project.

Table of Contents

ABSTRACT

Managers and coaches make the tactical decisions and substitutions during a football match. Formation rearrangements, positional changes and substitution of players are some of key decisions taken by coaches. An independent Poisson model has been used before to predict the probability of scoring in a match. This study uses the model to determine the effect of a substitute's performance on the goal difference in a match. Substitute's performance is measured by the minutes played, goals scored and yellow cards obtained. A dummy variable measuring the effect of home advantage is also introduced into the model. The paper uses observations of substitutes coming on to the field from 2012-2015 seasons of the KPL. An integrated independent Poisson regression model is used to estimate the optimal effect the substitution has in a match by estimating the parameters using a maximum likelihood function. The paper concludes that an away substitute introduced has a significant effect if introduced in the first, second and last period while there was no significant difference in the third period.

Key words; *Substitution, Kenyan Premier League, period.*

1. Introduction

1.1 Background

In association football (this paper further refers to it simply as football and also known as soccer in North America) tactical changes occur throughout the match. A typical football match consists of 11 players; 10 outfield players and a goalkeeper. The outfield players are classified as either defenders, midfielders or strikers. Tactical decisions may involve formation changes or positional changes during the match or a substitution change during stoppage of play. According to FIFA, the governing body of football around the world, a maximum of three substitutes may be used in any match played in an official competition organized under the auspices of FIFA, confederations or national football associations. The competition rules must state how many substitutes may be named, from three to a maximum of twelve. We have to note that once a substitution occurs, the player is not allowed back onto the field of play. (IFAB, 2017)

Managers and coaches have to make tactical decisions on when, which and how substitutes are introduced to the match. According to (Corral, Barros, & Rodríguez, The Determinants of Soccer Players Substitutions; A Survival Analysis of the Spanish League, 2007) substitutions are costly since the team must have a large number of quality players to make more efficient substitutions. Therefore, managers and coaches need to find an optimal strategy to decrease the cost of making substitutions. Injuries may occur due to extreme work rate on the pitch when the intensity of working out is increased from normal level (Gulhane, 2015). To prevent or reduce these injuries, substitutions have to be made during the match hence a need to find the best time and strategy to make the substitution.

1.1.1 Managers Role in Tactical Decisions and Substitutions

Managers tend to spread out their substitutions in the second half of the match. Some make the substitutions early in the match while others prefer introducing players later in the closing stages of the match. Making the three substitutions too early may cause a strain in the team if an infield player is injured. This will cause the team to finish the match with less players if all substitutes were used. Making all the substitutions in the very last minutes of the match can at times be too

late for a team in a losing position due to time constrains. But for a winning team, this can add the needed energy to hold on and actually win the match. In the Kenyan Premier League, the managers and coaches make these tactical decisions hence the need to find an optimal time to make the substitution and the optimal effects of substitutes on influencing the outcome (win rate) of the match.

Managers on the losing side have a higher incentive to make substitutions in a push to change the outcome of the match. but studies in (Corral, Barros, & Rodríguez, 2007) shows The score of the match not to be determinant in determining the optimal time a substitute is to be made . The paper also shows no correlation between the goals scored by substitutes and the score of the match at time of substitution. This is an assumption used in determining the factors used in this model.

1.2 The Kenyan Premier League

According to the Kenyan Premier League website[1], the league consists of 16 teams (recently increased to 18) who play against each other over two legs; once at their home stadium and also in their opponent's home stadium (away). The KPL season starts in February and ends in November with a total of 30 games being played by a team. The team with the most points wins the league title at the end of the season. In case the leading teams have same points, the team with the highest goal difference wins the league. The winning team also gets the opportunity to go to the playoffs of the Confederation of African Football (CAF) Champions League while the second team goes to the playoffs of the CAF Confederations' Cup. Two teams with the lowest points in the league are relegated to the National Super League. A tie in points means the team with lower goal difference is relegated. Top two teams in the National Super League are promoted to the KPL. The competitiveness and rewards shows the need for coaches to optimize their assets (players) capability during the season, both the starters and substitutes.

1.3 Effects of Substitutions on Football Matches

Substitutions are likely to happen depending on the current goal scores of the match. The team losing has more incentive to make a substitution than the winning team. The winning team makes more defensive substitutions to cement their win position while the losing team makes more offensive substitutions so as to win the match (Corral, Barros, & Rodríguez, 2007).Using an inverse Gaussian hazard model to analyze the first substitutions on each team taking place either

[1] Kenya Premier League Website- KPL.co.ke. Some of the data was collected directly from their offices.

at halftime or in the second half of the game, (Corral, Barros, & Rodríguez, 2007)argued that the most important factor explaining the timing of the first player substitutions on a team is the score as it stands prior to the substitution. Substitutes can either be an offensive tactic or defensive tactic by the manager. An offensive tactic occurs when a midfielder replaces a defender or a striker replaces a midfielder. (Maher , 1982).

This paper uses a multivariate independent Poisson regression model to optimize the value of a substitution by regressing the time a substitute is introduced in the match, goals scored by the substitute, yellow cards obtained on the goal difference after a substitution. The paper also includes the home advantage effect on the scores after a substitution. Data is obtained from the 526 matches played from 2012-2015 in the Kenyan Premier League (KPL). Other factors affecting player substitutions include tiredness of the player, tactical changes imposed by the coach, injuries and a red card to a player. Some of these factors are uncontrollable and forces the manager to make a substitution. These factors are not included in the model due to difficulty in obtaining data in the said period. Substitution made in the first half of the match are not significant in the research. This is due to many managers making 'forced substitutions' during the first half.

1.4 Research Problem

Football is evolving to be a more tactical and strategic game rather than a purely talent based sport. Most teams focus on bringing in quality players into their team while under-utilizing the current crop of players they have (Myers, 2012). For many years teams in the KPL have tried to formulate an optimal tactic to optimize substitutions. In the KPL, most goals scored in the KPL come in the second period of the match. In this period most team players are fatigued and hence performance drops. Managers need to determine an optimal time to introduce a substitute in order to optimize the substitutions made during this period. This is done so as to increase their score in the match. For instance, Gor Mahia, the 2013 KPL champions, had an average of 2.4 substitutions in the second half of matches played in the league. In the second half, Gor Mahia scored more goals after substitutions than the other teams. This can be seen to be a tactical approach to optimizing

substitutions in the league. The problem that KPL managers face is determining the optimal time these substitutions should be made. Prior to this research, no study to my knowledge has been carried out in Africa regarding this topic. This paper determines the optimal range of time to make a substitute, the impact substitutes add to their teams and the role of home advantage in the substitutes' performance.

1.5 Research Questions

1. What is the optimal time to introduce a substitute in a league match in the Kenyan Premier League?
2. What is the impact of home advantage on the substitutes' performance in determining the goal difference in the match?

1.6 Research Hypothesis

1. There is not a fixed minute to optimize the introduction of a substitute during the second half of the match but a range of time.
2. There is no discernible impact of home advantage on to introduce a substitute during the second half of the match.

2. Literature Review on the Optimization of Substitutions

Sports economics in Africa is an interesting topic of research to many authors. Prior to this paper, I have no knowledge of a research published on the optimization of football players in Kenya. Studies on other areas of the sport and other aspects have been carried out. For example, (Mählmann, , 1992) studied roles that sports carries out in affecting modernization in Kenya. The paper focuses on football (most popular sport) and athletics, which is Kenya's most successful sport collectively. (Mmbaya, 2013) Analyzed the financial performance of clubs in the Kenya Premier League using analytical regression on net earnings ratio, ratio of cash flows, assets owned and ratio of current assets to liabilities. Strategic planning and the challenges affecting strategic planning in Kenya through analysis of the management, board members and players in the KPL. (Omondi, 2008) . Substitution and optimization strategies in higher leagues such as in Europe, Major League Soccer and cup tournaments have been published elsewhere.

2.1 Theoretical review on Substitution

(Myers, 2012) Proposed the optimal time to make the three substitutions in a match using decision trees and data from the 2009/2010 seasons of the top four leagues in Europe[2], Major League Soccer-North American 2010 season- and 2010 FIFA™ World cup in South Africa. Optimality was based on improving the goal differential. The paper analyses the optimal periods a player should be introduced in a match when it is in a losing position. The research proposed the best times to make the substitutions is in the second half. A 58-67-79 minutes rule was formulated. The paper proposed the first substitution be made in the 58th minute, the second in the 67th minute and the final substitute to be made in the 79th minute. However, the research failed to account for the relative strength of the losing team. This paper focuses on the quality of the substitutes on the bench hence despising the notion that all substitutes are equal. The rule does not apply to winning or tied teams substitutions, who the managers may substitute at will. The decision rule also may not be applicable to laggard substitutions[3], substitutions due to injury and substitutions to increase morale in the team. The study also ignores the impact of a substitute when they enter the field which determines the possible timing of the next substitution. Goal differential was omitted as one

[2] According to the Union of European Football Association, the top four leagues in 2009 were the Spanish La Liga, English Premier League, German Bundesliga and Italian Serie A.

[3] Substitutions made by managers to decrease the momentum or waste time.

of the key factors to be considered in a substitution. This paper highlights the impact of substitutes in determining the optimal substitution techniques.

The proposed substitution rule by Myers was revised by (Silva & Swartz, Analysis of Substitution Times in Soccer, 2016) citing the flaws by Myers. Authors focused on the goal differential prior to the substitution and the quality of players in the lagging teams. The paper proposed an alternative analysis based on the Bayesian logistic regression. By using the same data as (Myers, 2012) and discrepant statistical methods and explanatory variables the results obtained differed. The model incorporated the strength of the trailing team, the time of the match and extra substitutions[4] who by the underlying assumptions, infuse energy to a team in the same way across all teams. For a trailing team with a strength of 0.2, the probability of the trailing team scoring the next goal was 55%. The authors concluded that for any two evenly matched teams, the trailing team in the second half will have a higher probability to score. They also observed that there is no discernible time during the second half when there is a benefit due to substitution. The authors suggested that leading teams should not play as if they are leading. They deduced from their study that, defensive substitutions made by the leading team gives the trailing team a greater probability to score next. The paper focuses their research on the substitution timing. It is not conclusive by the fact that some substitutions are made by managers focusing on prior performance of the player off the bench. This paper shows the effect a substitute brings on to his team using an integrated Bayesian Regression model which factors in the prior performance in a match where at least one substitution was made in the match. This enables this research to make more realistic predictions on how managers make substitutions in the KPL.

(Corral & Barros, 2007) Estimated a model for the first strategic change (substitution) in a football match using the inverse Gaussian models. The authors analyzed the Spanish La Liga using data obtained over the 2004-2005 season. The paper focused on four key variables to estimate the models. The first was pegged on being the home team, which the author showed had no effect on substitution. Second, the result of the match and the home team result. The observation made was winning teams make their substitutions later in the match and the result of the home team had no significant effect on substitution. The paper also included a substitution strategy component

[4]The ranking of coefficients for extra substitutions is as follows;

- 1 trailing team has made more substitutions than the leading team
- -1 trailing team has made fewer substitutions than the leading team
- 0 trailing team has made the same number of substitutions as the leading team

categorized in three; offensive, defensive or neutral. These substitution strategies involved players brought on. Defenders for midfielders or strikers is a defensive strategy, strikers for midfielders or defenders an offensive strategy while players substituted for the same position are neutral strategies. The final variables used were the strength component which was measured using points gained in the last four matches played. The authors concluded that goal differential is the key in the timing of substitutions. They showed there is evidence that teams that are behind substitute earlier than those that are tied or ahead. In addition, the study used the survival analysis to show that goal differential also affected the defensive and offensive substitutions. They concluded that defensive substitutions were made later in the match as compared to the offensive ones which were made earlier. Teams that were winning or tied also made more defensive substitutions as teams which are behind made more offensive substitutions. Research bases on the time elapsed until the first substitution is made, as proposed by (Green, 2003) in parametric duration analysis. The study ignores the effect of the second and third substitution. By integrating the second and third substitution in the models, this study finds the optimal time elapsed until the first, second and third substitutions are made.

(Bradley, Peñas, & Rey, 2014) Evaluated the match performances of substitutes in a match. The study focused on substitute's data collected in the FA English Premier League in the 2013 season. Statistical analysis using analysis of variance (ANOVA) was carried out on a large sample of substitute players. An independent-measure analysis compared the substitutes, substituted players and players completing the entire match using independent-measures analysis. Using a repeated-measure analysis, they compared players introduced as substitutes with previous performances across identical time periods. Finally the authors also quantified the physical and technical indicators across various playing positions. The performance variables used in the study was the speed thresholds, total distance covered and the pass completion rate. Results from the study showed that most substitutions occur at halftime and the latter part of the second half (Silva & Swartz, Analysis of Substitution Times in Soccer, 2016). The study revealed also showed substitutions become more offensive as the half progressed i.e. strategies increasing the attacking positions in the match. This is contrary to (Corral & Barros, 2007) results that showed defensive substitutions were more probable as the match wears out for winning sides. The analysis showed that substitutes had greater high-intensity-running than substituted players or players completing the match. Focus on how the substitutes performed after coming on compared to an equivalent

period in the second period highlighted that the same players covered more high intensity running when they were introduced as substitutes compared with the equivalent period of the second half. Attacking substitutes showed a greater intensity for running after coming on as substitutes. The study concluded that there is no significant difference in the pass completion rate between substitutes and players being replaced or completing the match. An independent-measures analysis does not factor in the abilities of the players being compared. A more skillful or experienced player has a greater influence in determining the outcome of the match, according to (Silva & Swartz, 2016). The research fails to account for the goals scored by a substitute or a negative effect the substitute has on a match such as conceding a penalty kick, scoring an own goal or being sent off after coming on.

Soccer Game Optimization is an algorithm developed for optimizing soccer player movements in the match (Purnomo, 2014) using a metaheuristic approach. The algorithm mimics the soccer player's movement during the soccer game and is implemented in continuous and discrete problems. The model in this paper takes into account the players within the sample space (pitch) only and does not take into account the dynamics of the match such as substitution. An introduction of a substitute or change in the formational tactics is not factored in the paper.

Analysis on the match performance of substitutions on English Premier League players carried out by (Bradley, Peñas, & Rey, 2014) evaluated the effect a substitute has on the match. They concluded that substitutes cover greater high-intensity-running distance but pass-completion rates did not differ for any position. The research did not take into account the overall effect a substitute has on the match factoring in key moments in the match; goals, penalties, yellow cards and red cards.

2.2 Overview of Literature Review

Research carried out in Kenyan footballing leagues do not take into account substitute's optimization at all. Most researchers focus on the value of football players and the football clubs revenue and profitability. This paper focuses on optimizing the value of substitute players in determining the key moments in a football match (scoring a goal). The paper aims at finding the value a substitute adds to the team during a KPL match by analyzing the minutes played by the substitute and the overall performance of the substitute. I am not aware of any paper published on optimization of substitutes using goals, yellow card substitutions and the home team's advantage

in the match to optimize substitutions. The research analyzes the variables through an integrated independent Poisson distribution model to find the optimal substitution in the Kenya Premier League. Through this, the optimal effect of the substitution shows the best time to carry out a substitution in the second period and the best substitution strategies coaches and managers in the KPL should focus on.

2.3 Empirical studies on Substitution

Studies and research on modelling of a football match aspect has been carried out before. Dyke and Clarke (2000) suggested a method for predicting the distribution of scores in international soccer matches. The paper assumed each team's goals scored were distributed as independent Poisson variables which depended on the match venue and ratings[5] of each team. The authors use the Poisson regression results to estimate parameters for this model. The model parameter estimates were used to simulate matches played during the 1998 World Cup tournament in France. Hirotsu and Wright (1982) used dynamic programming to analyze the optimal time to make substitutions and tactical decisions. The approach taken by the author analyzed a football match using the scoring and conceding of goals and also the gaining and losing of possession. The process termed as a four state Markov process, is estimated using the maximum likelihood function. The paper shows that although the probability of scoring may increase when a substitute is introduced, the substitute may also increase the probability of his team losing possession and conceding a goal. The authors concluded that when facing the decision of introducing two different players, the optimal time to introduce either substitute is determined by considering the current score and the time remaining in the match. The data is analyzed using a dynamic programming model. Football players have different abilities and skills which affects the performance in the match. (Corral & Barros, 2007) The four state Markov process applied in the research fails to factor in the strength of each substitute in the model. This paper further discusses the incorporated Poisson model that includes the strength component in each team.

Hirotsu and Wright (2006) offered a solution to their earlier paper by including both the strength of the teams and the tactic used by the coaches in substituting. The study analyzed the J-League Division 1 2002 data of goals scored and played time of each match. The authors proposed a

[5] Ratings are based on the FIFA rating system of each team in 1998.

theoretic approach to modelling tactical changes in the league using the Poisson distribution model on the data obtained in the league. The paper assumed probability of scoring and conceding a goal followed a Poisson distribution leaning on the work of Maher, 1982. These means represent the offensive strength for scoring a goal and defensive propensity to concede a goal in terms of a team's formation, i.e. a combination of the number of each type of outfield player on the pitch, and are estimated by means of the maximum likelihood method. The paper also discussed the possibility of either the home team or away team winning the match. The paper introduced coefficients of variables representing the offensive strength and defensive strength of each team in the match. The model factored in a four case Poisson process model in which the decisions made were either the best or worst for each team. The paper showed how the managers' decisions affect the probability of winning the match using real data of the Japan professional football league, by showing four cases of the quality of both managers' decisions, depending on whether they each use their best or worst strategies. Conclusion from the study showed that the probability of a team winning the match is affected by the manager's decisions. The model can be used to predict the winning of the home team or away team resulting to a zero-sum game but fails to account for draws brought about by the manager's tactical decisions.

(Karlis & Ntzoufras, 2003) Analyzed the estimation of tactical strategies in football using the bivariate Poisson distribution model. This study built upon the independent Poisson distribution model proposed before in research on this field. (See Lee (1997), Maher (1982) and Rue and Salvesen (2000) and others). The paper focuses on the Italian Serie A data for the 1991–1992 season and the Champions League data for the 2000–2001 season. The method proposed by the authors replaces the independent Poisson distribution assumption by considering a bivariate Poisson model and its extensions. The model introduced correlation between the events occurring between the two teams and explained the effects of any correlation. The paper concluded that using a bivariate Poisson distribution can improve model fit and also predict a draw in a football match. This observation offered a solution to Hirotsu and Wright (2006) assumption where only the probability of scoring was determined and a draw was not considered in the outcome of the match. However, the model is complex and requires high level mathematical algorithms which makes it difficult to estimate the model using normal estimators. The results obtained using the bivariate Poisson model in the paper fails to account for defensive and offensive tactics in the match. A factor accounting for the mixed strategies occurring in the match enables the model be more

effective in determining how change in tactics (from defensive to offensive and vice versa) affects the correlation between the events occurring in the match.

The paper bases its model from Lee (1997) independent Poisson distribution model of determining goals scored in a football match by integrating the tactical measures by managers in a league match. The paper analyzes Kenyan Premier League player substitutions from 2012 to 2015 to determine the optimal strategy the managers ought to take while making a substitution in a league match. The next section of this paper shows the methodology used and empirical studies used to formulate the integrated independent Poisson process model.

3. Methodology

3.1 Data Collection

Substitutes and goals data in the research is obtained from the Kenya Premier League official website and from their offices for substitutions made and goals scored in every match. Data of other variables used in the model; goals scored by substitutes, was obtained from the referees reports which are kept by the Referees Association of Kenya[3] and Football Federation of Kenya (FKF). Data is also collected on the number of goals for and goals against to measure the offensive and defensive strength of the team respectively. We ignore any substitution made in the first half of the game which may occur as a result of injury or sending off of a player. These 'forced' substitutions should not be included since they will not give a true estimate of the optimal value of the substitution (Corral, Barros, & Rodríguez, 2007). The focus of this paper is on outfield substitutions hence substitution of a goalkeeper is not considered. Managers do not intend to change their tactics by substituting a goal keeper given the constrained number of substitution hence including goalkeeper substitutions will not give the true estimate of optimal value of a substitution.

3.2 Observations

The league comprises of 16 teams who play against each other in a league system. Each team played a total of 30 games during the season. This paper uses data from the 2012-2015 KPL season giving the total games observed as 576. The number of observations on substitutions made in the observation period totaled to 4039 substitutions. This is further detailed to 1010, 1008, 1009 and 1012 substitutions made in the 2012, 2013, 2015 and 2015 season respectively. A total of 6 teams were relegated at least once during the period. A total number of 266 substitutes came on in the first half which may be due to either injury of a player or tactical changes after a red card. These may be 'forced' on the manager hence do not represent the true picture of a tactical substitution. A substitution of a goal keeper represents a forced substitution on the manager. None was made during the observation period. These 'forced' substitutions are not included in the study.

3.3 Independent Poisson Regression model

The independent Poisson distribution model suggested by Lee explains the probability of scoring a goal between teams in a view to increase their win rate. This paper chooses to follow Lee's method in determining the probability of a substitute optimizing their performance in order to

increase the win rate of their team in the match. The difference in scores before and after the substitution is made is used as a proxy to the optimal performance of a substitute. We assume that goals scored in the match are independent of each other i.e. a goal scored by a home or away team substitute does not affect goals scored by the away team or any other goal scored in the match. The model used in (Lee, 1997) estimates the average score in the game using the offensive prowess and defensive abilities and the effect of 'home advantage' on the home and away team.
This paper makes the assumption that goals scored by substitutes in a match follow a Poisson distribution. The increase in goal difference for their team after the substitution is the dependent variable in the model. The independent variables used in the model include the minutes played by the substitute, goals scored by the substitute, yellow cards and red cards obtained by the substitute and the effect of home advantage.

The mean of a home substitute optimizing his performance is, λ, while μ represents the mean of a substitute introduced by the away team to increase their teams goal differential. These respective means of the distribution reflect the strength of the substitution, the average time a substitution is made and the home advantage effect. Goals scored by a team are assumed to be independent of any other goals scored in the match.

The optimized mean of each substituting team, represented by λ or μ above, is expressed as the goal differential of an independent Poisson distributed function of the factors mentioned above in Section 3.2.

The equations below show the function of factors affecting the optimal performance of a substitute in the KPL. The function is a Poisson distributed function affected by the minutes played by the substitute, the substitute's performance and whether the substitute is playing at home or away. Equation 1 (a) and (b) shows the function:-

$$Goal\ Difference \sim \mathcal{P}\ f(minutes\ played, goals\ scored, cards\ obtained, \text{?}ome\ advantage)$$

$$\lambda = \alpha + \beta_1 T_s + \beta_2 G_s + \beta_3 Y_s + \beta_4 R_s + D_1 H$$

$$\mu = \alpha + \beta_1 T_s + \beta_2 G_s + \beta_3 Y_s + \beta_4 R_s$$

1(a)

1(b)

The mean, λ of this generalized linear model (GLM) represents the expected number of goals scored after a substitute is introduced. According to, (Lee, 1997) the Poisson distribution mean has to be positive, this is an assumption on which our model is built. We express the mean as a linear combination of minutes played by the substitutes, the overall performance of the substitute (goals scored and yellow cards obtained) over the minutes played and the effect of home advantage.
The probability of obtaining a value x in a Poisson distribution is defined in Equation 2(a), where n represents the number of observations and λ is the mean of the Poisson distribution function. The probability that a home substitute will optimize their performance during the match, λ is represented in Equation 2(a)

$$\Pr[X = x] = \frac{\lambda^x e^{-\lambda}}{x!} \quad for\ x = 1, 2, 3, \dots, n \qquad \textbf{2(a)}$$

Similarly, the probability that an away substitute will optimize their performance during the match, λ is represented in Equation 2(b)

$$\Pr[X = x] = \frac{\mu^x e^{-\mu}}{x!} \quad for\ x = 1, 2, 3, \dots, n \qquad \textbf{2(b)}$$

4. Data Analysis

The data was obtained at the primary source from the KPL offices, the referees' reports and other match official documents. The data collected represents a sample of 4039 observations of substitution patterns in the KPL. The variables of interest used are the aggregated goal differential before and after each substitution, minute of substitution, goals scored by substitutes, yellow and red cards obtained by the substituted player and the home advantage. The data is collected from the matches played in the KPL from the 2012-2015 seasons. All players are considered to have a role in attack and also defense of the score in this paper. This contradicts the idea that substitutions made by the managers hold a particular position in the formation employed in the game. In addition, according to (Bradley, Peñas, & Rey, 2014)it would be difficult to classify whether or not a substitution is attacking or defending and also not knowing enough about the players on the team. Lastly, the "home-effect" is included in the model as a test variable to determine if it impacts the substitutions made by the managers.

4.1 Statistical properties of the data.

The tests in the research are carried out on Stata© 14 to determine the efficiency of the model and to estimate the parameters of the model.

First, we obtained the summary statistics of the data collected for each season. Table 1(a) shows the mean, median and standard deviation of the variables contained in the data obtained for this research.

Table 1 Summary statistics for KPL 2012-1015 season

YEAR	2012	2013	2014	2015
Mean	64.20	65.36	64.31	67.17
Median	65	66	65	68
Mode	60	46	46	47
Std. Dev	15.45	14.57	15.47	14.93
N	1010	1009	1008	1009

In the first season, the distribution of the average time of substitution is skewed around the 64th minute, around the half point of the 2nd half. The other seasons' substitutions also average almost

he same minute with 65th, 64th and 67th minute for the next three seasons respectively. The mode ninute for introduction of substitutes is at the 46th minute for all seasons studied bar the first. This s rational considering most managers get a chance to make tactical changes during the half time ɔreak. The data is left skewed showing most substitutions are made in the second period of the natch. The substitutions made in the first half are mostly as a result of injuries not the willingness ɔf the managers or coaches to alter the team. Any substitution made after the 90th minute 'additional time) is rounded off as the 90th minute.

Table 2: Summary statistics for the 2012-2015 KPL season substitutions

	Mean	Median	Mode	Std. Dev	N
Min Played	65.38404	66	46	15.05228	4036
Goal Diff	0.63107	0	0	0.940982	4036
G. Scored	0.039643	0	0	0.205027	4036
Y. Cards	0.053846	0	0	0.280603	4036
R. Cards	0.001734	0	0	0.04161	4036

Table 2 provides a breakdown of the summary statistics of 4036 observations over the 2012-2015 KPL season. As seen in table 1, the mean time of a substitution made averages between the 64th and 65th minute constant with the overall mean at 65th minute. Most substitutions are done immediately after the break. Most managers get to discuss their tactics to the players offering a platform to make tactical changes and substitutions. This however was not the case in 2012 where most substitutions were done in the 60th minute. No clear explanation is offered in the paper why this is the case.

On average, a substitute increased his team goal differential by 0.63 goals. However, the substitutes themselves scored an average of 0.039 goals per match once introduced while accumulating 0.053 yellow cards and 0.0017 red cards per match after introduction.

Making substitutions is a common feature amongst the KPL managers. Over the periods selected for this research, 83.65% of the matches played had managers making all 3 substitutions. 14.25% of the matches had two substitutes and 2.1% used one substitute in the match. No match played in the period considered in the research had a game without either team making a substitution.

4.2 Independent Poisson Regression model

A standard χ^2 test is carried out to determine the significance of the variables in explaining the expected number of goals scored by substitutes. Two important assumptions are considered while carrying out the hypothesis tests that help us determine the factors that influence the optimization of a substitute in the KPL. First, it is assumed that goals scored by substitutes are independent. It is also assumed that goals scored in the match follow a Poisson distribution. Table 3 shows the results obtained from Portmanteau's Q-statistic test to determine the significance of the variables in the model.

Table 3 Q-Statistic test results

	Q-statistic	Prob > Chi2
T	704.3132	0.000
G.S	112.8108	0.000
Y	704.3132	0.000
R	30.9505	0.8471

***T**- Minutes played by sub* ***G.S**- Goals scored by sub* ***Y**- Yellow cards* ***R**-Red cards*

The test was carried out at 95% confidence interval ($\alpha = 0.05$). The hypothesis tests if the model's variables are significant in explaining the difference in scores in a match. Minutes played by the substitute, goals scored and yellow cards obtained by the substitute are significant. However, red cards obtained does not influence the scores in the match. This may be explained as a lack of enough data on substitutions obtaining red cards.

4.3 Poisson regression analysis

Assuming the goals scored in a match follow a Poisson distribution, we calculated the parameters for variables affecting expected goals scored by substitutes and determined the optimal time for substitution. We also include the effect of the substitution if the team is playing at home or away from home. We regress the goal differential after a substitution made by the following factors:-

- Minutes played by the substitutes. (T_s)
- Goals scored by substitutes. (G_s)
- Yellow cards obtained by the substitutes. (Y_s)
- Red cards obtained by the substitutes. (R_s)
- Home advantage. (D_1H), dummy variable, HOME=1 and AWAY=0.

We want the mean, λ of this distribution to reflect the number of minutes played by a home substitute, the ability of the substitute (as a factor of goals scored and yellow cards obtained) and the home advantage, if it applies. We do this by expressing the logarithm of each mean as a linear combination of the factors. This neatly builds in the requirement that the mean of the Poisson has to be positive. Our equation for the mean of the home team is:-

$$\lambda = \beta_1 T_s + \beta_2 G_s + \beta_3 Y_s + \beta_4 R_s + D_1 H$$

The away team's mean of the distribution similarly represents the number of minutes played by a home substitute and the ability of the substitute. However, the home advantage dummy is zero, hence this mean been used as a control variable for the home advantage of the home team. The equation for the mean of the away team is:-

$$\mu = \beta_1 T_s + \beta_2 G_s + \beta_3 Y_s + \beta_4 R_s$$

We studied 4039 observations collected over 526 matches played across the 2012-2015 KPL seasons. Equation 1(a) shows the independent Poisson regression model used in this research. Table 4 shows the coefficients obtained from running the regression on the observations collected in the four KPL seasons till 2015.

Table 4 Poisson regression coefficients on the 4039 substitutions in KPL (2012-2015)

Diff	Coeff.	Std. error	Z	P> l z l
T	0.0342	0.0018	-5.350	0.0000
G	0.6336	0.2371	4.170	0.0000
Y	-0.0798	0.1072	-2.410	0.0161

$\mathbf{D_1H}$	-0.0511	0.0534	-2.330	0.0200

Therefore, the overall model can be represented by Equation 3 below. The dummy variable representing the home advantage takes a value of 1 if it is the home team and 0 if it is an away team substitute.

$$Diff = 0.0342T_s + 0.6336G_s - 0.0798Y_s - 0.0511 * D_1H \quad 3$$

To determine the optimal time a substitute should be introduced, we carried out a categorical outcome Poisson regression model on each period of interest during the match. We divided the 90 minutes in a match into four periods: the first 45 minutes, 46th - 60th minute, 61st - 75th minute and 76th – 90th minute. In each period we check for the effect of the variables on the goal difference of the match. The coefficients to the explaining variables are shown in Appendix 1.

5. Results

After fitting the model, and estimating the parameters using the maximum likelihood estimator, we can represent the overall model by Equation 3 below. The dummy variable representing the home advantage takes a value of 1 if it is the home team and 0 if it is an away team substitute.

$$Diff = 0.0342T_s + 0.6336G_s - 0.0798Y_s - 0.0511 * D_1H \quad (3)$$

The constant had to be dropped in order to incorporate the dummy variable. An inclusion of the constant causes a multicollinearity problem of the constant and the home advantage variable. The minutes played parameter is 0.0342, 0.6336 for the goals scored parameter, -0.0798 for the yellow cards received parameter and the home advantage parameter with 0.0511. This shows that for a 'typical' away team the average goal difference after substitution will be 1.02 while the home side will average a goal difference of 0.97 goals after a substitution is made. [6]

The 'home advantage' effect in the KPL for the four seasons sampled prove to be a disadvantage after substitution. This may be explained by the fact that most teams in the league do not have a home stadium. The home advantage parameter had a p-value of 0.026, however, when the data was analyzed annually, the factor had no significance in determining the optimal value of a substitution.

5.1 Optimal time of substitution

For each of the four periods discussed in section 4.3 above we carried out a Poisson regression to obtain the factor variable's coefficients. The results are shown in Appendix 1. We grouped the first half substitutions as the first period. One possible explanation for not grouping the substitutes into 15 minute periods is due to the low number of substitutions made before the first 30minutes.

In the first period (minute 1-45), minutes played by the substitute, the goals scored and home advantage parameters are statistically significant in increasing the game's goal difference. However, there is a slight significance of yellow cards reducing the score obtained by a substitute player during this period. Using the rationale in the section above to calculate the average goal difference after a substitution, we determine that, the home team on average increase their goal difference by 1.09 in comparison to the away team that on average increases their goal differential

[6] See Appendix 2

›y 1.47 after a substitution. This shows an advantage of 0.38 goal to the away side if both the nanagers decide to make the substitutions before the 45th minute.

From the 46th to 60th minute, only the minutes played and the goals scored are statistically ;ignificant in explaining the difference in scores between the two teams. The yellow card ›arameter is slightly statistically significant. 'Home advantage' is not a significant factor in :xplaining the difference in scores between the home team and away team. If both the home team ınd away team make a substitution each during this period, the home team will on average score).48 less goals than their opponents with an expected goal differential of 1.48 during this period.

The third period represents the modal range of substitutions in the KPL as seen in section 4.1 ıbove. The variable coefficients are all statistically significant in this period. Substitutions made luring this period represents 34.39% of the total number of substitutions made during the 2012-2015 KPL seasons. The home team averages a goal difference of 1.59 while the away team has an :xpected goal difference of 1.61 after a substitution made during this period. This shows a slight lifference in the effect of 'home disadvantage'. A possible explanation to this is the intensity of ›lay during this period where both teams are giving their best in order to defend their lead, reduce :he difference in scores or to comeback from a losing position.

Substitutions made during the last period (minute 75-90), represents 28.72% of the substitutes nade. The minutes played by the substitute is not statistically significant in explaining the goal lifference in the match. This is an expected trend given that of the 1160 substitutions in this period, 593 substitutes played 10 minutes or less. That represents over 50% of the observations. Furthermore, 125 substitutes were introduced in the 89th minute or later. The goals scored and ıome advantage parameters are statistically significant with the yellow card parameter slightly ;tatistically significant in explaining the difference in scores. This period also represented the ıighest goal differential in all the four periods examined in this section. An average of 2.3 goal lifferential for the home team and 1.98 for the away team shows that this is the period that most ;ubstitutes make an impact after introduction. For each substitution made during this period :hough, the away team averaged 0.15 more goals than the home team. This difference is not so ;ignificant but shows, away teams in the league appear to fair better after substitutions than the ıome team.

5. Conclusion

In this paper, we analyze the optimization of a substitute. A substitute is optimized only if, during his time on the pitch the goal difference of his team does not decrease. To do so, we implement the multivariate independent Poisson regression. Thus this model analyzes the goals scored by the substitute, the yellow cards obtained, minutes played and the home advantage effect. We analyze the substitutions made from the Kenya Premier League in the 2012 to 2015 season.

The results show that the away team has a greater expected goal difference compared to the home team. This disqualifies the 'home advantage' as a disadvantage. Specifically, the best range of time a home team can make a substitution is from the 60^{th} to 75^{th} minute. The away team can make the substitution either in the first or final 15 minutes of the second half of the match in order to optimize their performance. The value of this research is that the estimated model optimizes the performance of a substitute and to show the impact of a home substitute being introduced in the match.

7. Suggestions for further research

This study was based on substitutions made in the Kenyan Premier League and the impact they have on matches played. One of the key suggestions for further research is the replication of this study in other sporting disciplines. This suggestion relies heavily on the availability of data on substitutes and the distribution of scores in the sport.

The approach to modeling the optimization of substitutions is a little simplistic. First, we have taken no account of the fact that a substitute may either be a defensive or attacking. We also assume that all substitutions are equal which is unrealistic assumption. In addition, the model used only has four variables which was due to lack of data. Future research in the topic can be made more robust if data is available for other factors affecting the scores in the match such as the position played by the substitute and assists by substitutes.

The assumptions made on the model enable us to determine the optimal time range the substitute should be introduced. There is a gap to be filled in this area of research by obtaining the optimal position a substitute should play and also the optimal value a particular substitutes can add to the match.

Finally, this study can also be extrapolated to determine if the optimal use of substitutions has an impact in the overall league standings after a season. This shows the need for a team to optimize

ıeir substitution throughout the season and the effects it may have in determining the teams' :ague position.

. Limitations to the study

'he research carried out in this paper turned out to be a success. However, we faced various hallenges in analyzing the data and obtaining the results. First, the data provided does not feature ıe assists made by substitutions. This is an omitted variable that can make the model more ccurate. The data provided also did not show the positions of the players. This made it difficult ɔ categorize the substitutions as either an attacking or a defensive tactic.

'he other limitation is the use of goal difference in scores after as substitution as a proxy for the ıptimization of substitutes. A proxy only allows partial results to be obtained hence the optimal 'alue of the substitutions will only be close to optimum but not the optimum.

'or the model, teams that tend to score a lot of goals against weak opponents are overrated by a nodel that looks only at scores and teams that settle into a defensive shell once they have got the ead may be underrated. This is due to the fact that this model focuses on the goals scored by ubstitutes as the only attacking variable a substitute brings after coming on.

). References

:orral, J. D., Barros, C. P., & Rodríguez, J. P. (2007). The Determinants of Soccer Players Substitutions; A Survival Analysis of the Spanish League. *Journal of Sports Economics*.

3radley, P. S., Peñas, C. L., & Rey, E. (2014). Evaluation of the Match Performances of Substitution Players in Elite Soccer. *International journal of sports physiology and performance*.

:arlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson Models. *The Statistician, 52*(Part 3), 381-393.

3ilva , R. M., & Swartz, T. B. (2016). Analysis of Substitution Times in Soccer.

3ilva , R. M., & Swartz, T. B. (2016). Analysis of Substitution Times in Soccer. *Journal of Quantitative Analysis in Sports*.

\ziza, B. (2010, June 6). *The Future of Soccer is in its Analytics.* Retrieved from Smart Data Collective: http://smartdatacollective.com/Home/27719

3radley, P. S., Peñas, C. L., & Rey, E. (2014). Evaluation of the Match Performances of Substitution Players in Elite Soccer. *International Journal of Sports physiology and Performance*, 415-424.

:orral, J. D., & Barros, e. a. (2007). The Determinants of Soccer Players Substitutions; A Survival Analysis of the Spanish League. *Journal of Sports Economics*.

3reen, W. (2003). *Econometric Analysis.* New Jersey: Prentice Hall.

3ulhane, T. F. (2015). Sports injuries: Causes, Symptoms, Treatment and Prevention. *International Journal of Physical Education,Sports and Health*, 107-109.

IFAB. (2017, January). *Fédération Internationale de Football Association (FIFA).* Retrieved from The International Football Association Board: Fifa.com

Lee, A. J. (1997). Modelling scores in the Premier League: Is Manchester United really the best? doi:10.1080/09332480.1997.10554791

Maher , M. J. (1982). Modelling association football scores. *Statis- tica Neerlandic, 36*, 109-118.

Mählmann, , P. (1992). The Role of Sport in the Process of Modernisation: The Kenyan Case. *Journal of Eastern African Research and Development*, 120-31.

Mmbaya, K. L. (2013). FINANCIAL PERFORMANCE OF FOOTBALL CLUBS IN KENYA.

Myers, B. R. (2012). A Proposed Decision Rule for the Timing of Soccer Substitutions. **Journal of Quantitative Analysis in Sports*.

Omondi, E. W. (2008). STRATEGIC PLANNING OF FOOTBALL CLUBS IN THE KENYAN PREMIER LEAGUE.

Purnomo, H. D. (2014). Soccer game optimization for continuous and discrete problems. *Juarnal Metris*.

0. APPENDICES

Appendix 1: Regression analysis of the 'four' periods

Minute Range					
		Coefficient	**Std. Error**	**z**	**P>|z|**
0 – 45	**TIME**	0.0218	0.0059	3.7000	0.0000
	GS	0.8031	0.1545	5.2000	0.0000
	Y	-0.4396	0.2698	-1.6300	0.1030
	DH	-0.2962	0.1278	-2.3200	0.0210
46-60	**TIME**	0.0127	0.0071	1.7800	0.0750
	GS	0.3870	0.1059	3.6500	0.0000
	Y	0.0213	0.1407	0.1500	0.8800
	DH	-0.0324	0.0666	-0.4900	0.6270
61-75	**TIME**	-0.0020	0.0095	-0.2200	0.8300
	GS	0.6875	0.1233	5.5800	0.0000
	Y	-0.1884	0.1729	-1.0900	0.276
	DH	-0.0139	0.0735	-0.1900	0.85
76 – 90	**TIME**	0.0074	0.0095	0.7800	0.4370
	GS	1.1722	0.1755	6.6800	0.0000
	Y	-0.4253	0.2481	-1.7100	0.0860
	DH	-0.0694	0.0852	-0.8100	0.4160

Appendix 2: Log difference of the scores in the match

The home side goal difference

$$Diff = 0.0342T_s + 0.6336G_s - 0.0798Y_s - 0.0511 * D_1H$$

$$e^{(0.0342+0.06336-0.0798-0.0511)} = 0.97$$

The away side goal difference

$$Diff = 0.0342T_s + 0.6336G_s - 0.0798Y_s$$