DEGREE PROJECT, IN APPLIED MATHEMATICS AND INDUSTRIAL ECONOMICS , FIRST LEVEL
STOCKHOLM, *SWEDEN 2015*

Modelling Football as a Markov Process

ESTIMATING TRANSITION PROBABILITIES THROUGH REGRESSION ANALYSIS AND INVESTIGATING IT'S APPLICATION TO LIVE BETTING MARKETS

GABRIEL DAMOUR, PHILIP LANG

KTH ROYAL INSTITUTE OF TECHNOLOGY

SCI SCHOOL OF ENGINEERING SCIENCES

Modelling Football as a Markov Process

Estimating transition probabilities through regression analysis and investigating it's application to live betting markets

G A B R I E L D A M O U R
P H I L I P L A N G

Degree Project in Applied Mathematics and Industrial Economics (15 credits)
Degree Progr. in Industrial Engineering and Management (300 credits)
Royal Institute of Technology year 2015
Supervisors at KTH: Boualem Djehiche, Anna Jerbrant
Examiner: Boualem Djehiche

TRITA-MAT-K 2015:11
ISRN-KTH/MAT/K--15/11--SE

Royal Institute of Technology
School of Engineering Sciences

KTH SCI
SE-100 44 Stockholm, Sweden

URL: www.kth.se/sci

Abstract

This degree thesis aims for a modeling of football set pieces (i.e Throw Ins, Free Kicks, Goal Kicks and Corners) through the use of Markov theory. By using regression analysis on a various range of covariates we will try to estimate the transition probabilities of such a process from a state to another and investigate what factors might have an impact on these probabilities. Although not reaching a sufficiently high level of variance explanation, the model constructed shows strong significance and let us believe that an articulation of it could lead to a strong model for these set pieces. Furthermore we will proceed with an analysis addressing the application of such modeling within the pricing processes of betting companies, based on a case study of Metric Gaming. Undertaking an operational management perspective, we will assess which level of implementation of such modeling is the most efficient, and what consequences it will have in two sub-perspectives; the risk management and branding of the company.

Acknowledgments

We would here like to take the opportunity to thank those who have aided us through this project. Our supervising professor, Boualem Djehiche, for guiding us in the right directions and always being available for consultation. The people at Metric Gaming and Betpump for assisting us with interviews and valuable insights. Thank you.

Contents

1 Introduction **7**
- 1.1 Background . . . 7
- 1.2 Purpose and Aim . . . 7
- 1.3 Problem Definition . . . 8
  - 1.3.1 Definition and Assumptions: the Set Pieces . . . 8
- 1.4 Industrial Engineering Application . . . 8
  - 1.4.1 Purpose and Aim . . . 9
  - 1.4.2 The Vig . . . 9

2 Theoretical Framework **11**
- 2.1 Markov Theory . . . 11
  - 2.1.1 Chapman Kolmogorov . . . 12
  - 2.1.2 Absorbing states . . . 12
  - 2.1.3 Communication between States and Irreducibility . . . 12
- 2.2 Regression Analysis . . . 12
  - 2.2.1 The *Logit* Regression . . . 13
  - 2.2.2 Discarding Models and Improving upon them . . . 13
    - 2.2.2.1 *Likelihood-Ratio* test . . . 13
    - 2.2.2.2 The *Wald-test* . . . 14
    - 2.2.2.3 AIC-test . . . 14
    - 2.2.2.4 Goodness of Fit - R^2 . . . 14
  - 2.2.3 Errors . . . 15
    - 2.2.3.1 Multicollinearity . . . 15

3 Methodology **17**
- 3.1 Set-up of the Markovian Model . . . 17
  - 3.1.1 The Transition Matrix . . . 17
  - 3.1.2 The Impact of Game Related Factors . . . 18
  - 3.1.3 The Impact of Team Related Factors . . . 18
- 3.2 Data Collection . . . 19
  - 3.2.1 The States and Game Related Covariates . . . 19
  - 3.2.2 The Match Odds . . . 21
- 3.3 Estimating the Transition Probabilities . . . 22
  - 3.3.1 Counting Procedure . . . 22
  - 3.3.2 Logit Regression . . . 22
  - 3.3.3 Choosing the Model for Different Transitions . . . 22
  - 3.3.4 The Final Transition Matrix . . . 23
- 3.4 Computation of the *RF-1* Odds . . . 25
  - 3.4.1 Probability Distribution at End of the Proxy Period . . . 25
  - 3.4.2 Probability of Next *RF-1* Event . . . 25

4 Results **27**
4.1 Estimation of the Transition probabilities . . . 27
4.1.1 M1 Matrix . . . 27
4.1.2 The Wald-test . . . 27
4.1.3 Reducing the models AIC test . . . 28
4.2 Validation of the Regression Models . . . 30
4.2.1 The VIF-test . . . 30
4.2.2 Log-likelihood ratio . . . 31
4.2.3 Goodness of Fit . . . 34
4.3 The Resulting Odds . . . 35

5 Analysis **38**
5.1 Mathematical . . . 38
5.1.1 Overall Interpretation of the Transition Model Matrix . . . 38
5.1.1.1 The θ_{odds} Factor . . . 38
5.1.1.2 The θ_{time} and θ_{score} Factors . . . 39
5.1.2 Comparing Results with SuperLive Odds . . . 39
5.1.3 Propositions for Improvement . . . 39
5.1.3.1 The High Unexplained Variance - More Factors Needed? . 39

6 Industrial Application **42**
6.1 Metric Gaming . . . 42
6.2 Methodology . . . 43
6.3 Theoretical Framework . . . 44
6.3.1 Risk Management . . . 44
6.3.2 Business Process Re-Engineering . . . 44
6.3.3 B2B Branding - Innovation . . . 45
6.4 Results of Investigation . . . 45
6.4.1 Operational Risk Management . . . 45
6.4.1.1 Trading, liability management and pricing . . . 45
6.4.1.2 The High Fixed Cost . . . 47
6.4.2 Branding - RF-1 an innovative product . . . 48
6.5 Analysis - Implementation . . . 49
6.5.1 Three Levels of Implementation . . . 49
6.5.1.1 Full Automatisation . . . 49
6.5.1.2 Semi Automatisation . . . 49
6.5.1.3 Support Tool . . . 49
6.5.2 Risk Management . . . 50
6.5.2.1 Liability Management and Quality Improvement . . . 50
6.5.2.2 Profitability Risk - Reducing High Fixed Cost . . . 51
6.5.3 Branding . . . 51

7 Conclusions **53**
7.1 Our Modeling 53
7.2 Choosing The Level of Implementation 53

1 Introduction

1.1 Background

The, by far, most popular sport in the world is football. Practiced by 250 million and with a fan base exceeding 3.4 billion[1], more than half of earth's population are somehow engaged in the sport. In the USA quantitative analysis has been frequently applied to determine the dynamics of american sports. However, since football (or as Americans prefer it, soccer), has been trailing in popularity (USA), the research around the sport has suffered. In the last decennium however, the industry regarding quantitative analysis on football has become established [2]. The prime example of this is OptaSports (Opta™), a company solely devoted to collecting and analysing data for the benefit of football clubs, media and the betting industry [1].

Being great football fans ourselves, as well as students of mathematics, quantifying the dynamics of the game became an interest of ours. An idea which struck us a year ago was whether the probabilities regarding set pieces could be derived and explained using *Markov Theory*, modeling said set pieces as states in a Markov Chain.

Applying Markov Theory within sports is not a new phenomena, especially when analysing results. Modelings of game dynamics is more sparse, primarily focused on Baseball [3] and American football [4], both of whom are somewhat *turn based*[2]. So how does one approach football, a sport whose dynamics is based on "free flow" and never stops for more than a few seconds, when one is interested in modeling the game as a Markov Chain[3]?

1.2 Purpose and Aim

Why would one be interested in determining a way to explain the probabilities for certain events during a football game? First and foremost, using Markov Theory as we propose, further pushes the applicability of mathematics within sports science making it possible to, hopefully, use our results in other sports which presents similar dynamics.
Our study aim to research whether this way of modeling could explain why the outcomes of aforementioned set pieces vary as they do. From these results we will try to build a predictive model to be used for determining probabilities for events happening in football games.

[1] For a very hefty charge of course, these types of enterprises make fortunes selling data

[2] The game frequently starts and stops between plays leading to each play being considered as a "mini game" itself.

[3] Not accounting for irregularities such as substitutions, sending off's and injuries

1.3 Problem Definition

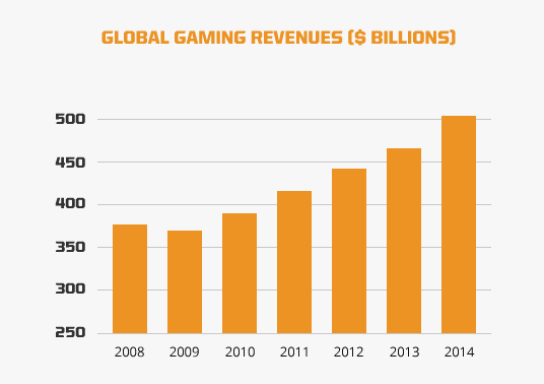
Can we model the outcomes of set pieces in football as a Markov chain? What factors are prone to alter the transition probabilities and how can we integrate these into the modeling?

1.3.1 Definition and Assumptions: the Set Pieces

The football set pieces we wish to model are; Throw Ins, Free Kicks, Goal Kicks and Corners, which all occur during a game of football. To make the model more plausible for Markov assumptions (see (1) in 2.1) we include passes as a fifth event in our model. Furthermore, we divide the passes, throw ins and free kicks in three field-location/action categories; Attack (A), Defence (D) and Central (C). Note that this classification takes into account which team that is in possession of the ball. This, we believe, is needed for our central assumption that the transition probabilities between the 11 aforementioned states are in fact Markov, ergo possess the so called *memoryless* property - what that constitutes will be covered later on (see 2.1). Indeed, we believe that the probability of the ball being played from a certain state A to another state B solely depends on the current state A.

1.4 Industrial Engineering Application

Today, the markets of betting and gambling are greater than ever. An increased globalisation, combined with the evolvement of the Internet and deregulations has lead to the global betting market surpassing a total revenue of $500 billion in 2014 [5]



We will, for our Industrial Engineering Application, operate under the assumption that we have managed to create a sufficient model for determining the probabilities of outcomes. We will perform a case study to see how our results may be implemented at the company where we work, (Betpump, a subsidiary of Metric Gaming). They offer very innovative markets which they sell to betting companies in form of a fully integrated product called the Superlive™-platform.

Part of the above mentioned platform is the *RF-1* market, short for *Rapid Fire*. It works in such a way that when it opens, the punters[4] have the possibility to bet on four different outcomes, namely the set pieces which we study.

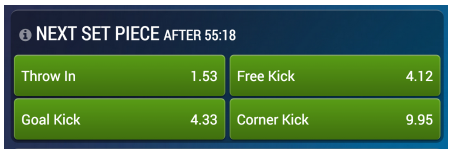


Figure 1: How it looks from an end-customers perspective

They have 40 seconds to place their bets after which a 10 second "cooling of period" takes place, an instrument which helps the company to hedge themselves from time discrepancies. This means that there might be people at the actual game, i.e. having information about the outcomes before Metric Gaming, who covers all games through high speed live feeds but still may experience information delays from 0 to a couple of seconds.

By analysing how to integrate a software for odds calculation (using our model) and what benefits it might bring, from two perspectives; *Risk Management* and *Branding*. We will thus answer the following research question:

How can our results be implemented at Metric Gaming's pricing process for the RF-1 market and to what extent, from an risk management perspective, and how will it affect the company brand?

1.4.1 Purpose and Aim

Our purpose with this module is to investigate how our modeling of football set pieces should be integrated in the pricing process of betting companies, which offers betting markets on said events. We aim to find an optimal implementation within an risk management perspective. Goals that we believe to be important to keep in mind whilst doing this; improve the product, improve the brand and minimise costs/increase profits.

1.4.2 The Vig

Furthermore, we need to present the basis regarding how odds work and how the money is made. All gambling companies operates with a *vig*; the tax the company imposes on its prices, a probability space, $\Omega = 1$, is multiplied by, e.g. 1,07 if the company wishes to have a tax of 7% [6]. If you have two outcomes that are equally probable, this is how you calculate the odds with the vig.

$$o_1 = \frac{1}{2} = o_2$$

[4]A punter is the end customer; those who bets on markets offered by gaming company.

Then the odds would be 2.00 and 2.00. But if you impose the tax,

$$(o_1 + o_2) = 1.07$$

now the probabilities for each of the outcomes is equal to 53.5%. When you invert these you get the odds for both to be 1.87. This is how they make money, and are hence looking for as large turnovers as possible. By, as mentioned in the interview, increasing player turnover to multiple times a game, the Metric Gaming creates great value for their clients.

2 Theoretical Framework

In this section we will present the different theories, stemming from different areas of science, which we have used to build our model and analyse our results.

2.1 Markov Theory

Fundamental to our thesis is the theory of *Markov Processes* and more specifically *Markov chain*. Such can be defined as a stochastic process which is a Markov chain if [7, p. 9],

$$\begin{aligned} P(X(t_{n+1}) = i_{n+1} | X(t_n) = i_n, X(t_{n-1}) = i_{n-1}, ..., X(t_0) = i_0) \\ = P(X(t_{n+1}) = i_{n+1} | X(t_n) = i_n) \quad \forall\, n \,\wedge\, i \end{aligned} \tag{1}$$

where i is any state within a given Markov chain,

$$i \in \Omega = \{1, ..., N\}$$

which is the probability space containing all possible states within a given Markov Process. The space could just as easily be countable but infinite however, relevant to our research is the case where we have a finite and countable set of states. This means that the only relevant information is in the *current* state, but the process does not regard "historic jumps" between states. The definition just proposed illustrates the core element of what constitutes Markov Processes namely the *Markov Property*, i.e. that the stochastic process be *memoryless*.

We are now ready to introduce the *transition probabilities* for a *time homogeneous*[5] Markov chain, i.e the probability of moving from one state to another (or staying in the same) at any given time using the above, (1), definition.

$$p_{ij} = P(X_n = j | X_{n-1} = i) \quad i, j \,\in \Omega \tag{2}$$

All the possible transition probabilities form a *transition matrix*, $\boldsymbol{P}$, defined as follows,

$$\boldsymbol{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \dots & p_{1N} \\ p_{21} & p_{22} & p_{23} & \dots & p_{2N} \\ p_{31} & p_{32} & p_{33} & \dots & p_{3N} \\ \vdots & \vdots & \vdots & \ddots & \\ p_{N1} & p_{N2} & p_{N3} & \dots & p_{NN} \end{bmatrix} \tag{3}$$

In the above matrix, the following holds,

[5] If the "transition operator", (2), does not change across transitions, a Markov chain is said to be *time homogeneous*

$$\sum_{i=1}^{N} p_{ri} = 1 \quad \forall\, r \in \Omega \tag{4}$$

As one can easily see, each row constitutes a full probability space, given that one currently is in a specific state, as represented by a corresponding row.

2.1.1 Chapman Kolmogorov

Below we have the Chapman-Kolmogorov Equations [7, p. 13], relations which one uses when wanting to derive a transition matrix after a certain, n, amount of jumps.

$$\begin{aligned} &a) \quad p_{ij}^{(m+n)} = \sum_{k \in \Omega} p_{ik}^{(m)} p_{kj}^{(n)} \\ &b) \quad \boldsymbol{P}^{(m+n)} = \boldsymbol{P}^m \boldsymbol{P}^n \\ &c) \quad \boldsymbol{P}^{(n)} = \boldsymbol{P}^n \\ &d) \quad \boldsymbol{p}^{(n)} = \boldsymbol{p}^{(0)} \boldsymbol{P}^{(n)} = \boldsymbol{p}^{(0)} \boldsymbol{P}^n \end{aligned} \tag{5}$$

2.1.2 Absorbing states

Absorbing states can be defined as a state to which the process can enter and then never leave. Therefore the following holds, where T_i is the number of time-steps required to reach the absorbing state, i,

$$P(T_i < \infty) = 1 \tag{6}$$

Then, of course, $p_{ii} = 1$, i.e. the probability of staying in the same state given that you have entered it is 1 [7, p. 16].

2.1.3 Communication between States and Irreducibility

States are said to be *communicating* with each other if it is possible to jump between states and back again (possibly with intermediary states in between) [7, p. 24]. If all states are communicating with all other states a chain is said to *irreducible*, i.e. that all states are *transient*.

2.2 Regression Analysis

We will, for estimating the aforementioned transition probabilities, use regression analysis since we believe that several factors have impacts on said probabilities. Regression analysis is, simply put, a statistic branch where one uses observations/data to estimate the relations between different parameters.

2.2.1 The *Logit* Regression

The *Logit* regression model is an appropriate model when dealing with probabilistic dependent variables. Otherwise, using an ordinary linear model might result in covariates being, $x_i\beta > 1 \vee < 0$. The Logit model is defined as follows [8, p. 12-14],

$$y_i = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} = p(x_i\beta), \qquad i = 1, \ldots, n \tag{7}$$

where y_i is an observation of the the probabilistic dependent random variable y, n is the number of observations, x the covariates and β the covariates' coefficients. The observational data upon which the regression is computed are dummy variables, d_i, which can undertake the value one for occurrence, zero if not. In our modeling for estimating the probability of jumping from a state to another, the data will consist of dummies, d_i, for the jump from a state to another state, s, such that

$$d_i = \begin{cases} 1, & \text{if the process jumps to state s} \\ 0, & \text{if the process jumps to another state} \end{cases}$$

The estimation of y is then computed by maximising the *log-likelihood* function,

$$ln(L) = \sum_{i=1}^{n} ln[(2d_i - 1)p(x_k\hat{\beta}) + 1 - d_i]$$

over the estimated coefficients $\hat{\beta}$. The interpretation of the influence of the estimated β_js is not the most intuitive. The impact of a positive β_j on the dependent variable y will indeed be positive but depends on the value of the other covariates. To visualise this we can rewrite the Logit specification as,

$$ln\left(\frac{y_i}{1 - y_i}\right) = x_i\beta$$

which shows the relation between the β_js and their logarithmic impact on the probability.

2.2.2 Discarding Models and Improving upon them

2.2.2.1 *Likelihood-Ratio* test For testing Logit regression models we can use the Likelihood-Ratio test [8, p. 13]. As in the F-test, one tests the likelihood of one or several coefficients of the covariates are equal to zero. Let $ln(L_*)$ be the log-likelihood function where the restrictions have been applied, then,

$$2ln(L) - 2ln(L_*)$$

is approximately a $\chi^2(r)$ distributed variable, where r is the number of restrictions.

2.2.2.2 The *Wald-test* The Wald test [8, p. 17-18] is much like the likelihood-ratio test and can be used for testing the significance of covariates in a model, i.e. testing a proposed null hypothesis[6]. Unlike the log-likelihood ratio test the Wald test does not require comparing with another regression model. By calculating the Z-statistic,

$$Z = \frac{\hat{\beta}}{SE}$$

where $\hat{\beta}$ is the estimated coefficient of the tested covariate and SE its respective estimated standard error. If we then square the Z-statistic, we obtain the so called Wald-statistic, W, which is a χ^2 distributed variable. We then get the probability of the null hypothesis being true by,

$$P(x > W) = 1 - P(x < W)$$

2.2.2.3 AIC-test The *AIC*-test, or *Akaike Information Criterion* [8, p. 21], computes the relative quality of a model to another with a different choice of covariates. Unlike the above mentioned tests, it takes into account the trade-off between model complexity and *goodness of fit* (GOF). The model to be chosen is the one which minimises the AIC value obtained by,

$$n\ ln(|\hat{e}|^2) + 2k \tag{8}$$

where k is the number of covariates and n the number of observations. For testing on Logit regression models, the value to minimise will be,

$$-2\ ln(L) + 2k \tag{9}$$

where L is the aforementioned log-likelihood formula.

2.2.2.4 Goodness of Fit - R^2 R^2, commonly referred to as the *coefficient of determination*, is a quantity which measures how well a set of observations fits a proposed model. The value is the fraction of explained variation and total variation. It is defined as follows [8, p. 8],

$$R^2 = 1 - \frac{Var(\hat{e})}{Var(y)} = \frac{Var(x\hat{\beta})}{Var(y)} \qquad R^2 \in [0:1] \tag{10}$$

where R^2 ranges from 0 to 1. It is actually possible to get a negative value of R^2, in that case the fit is worse than fitting the data to a horisontal line. If such an value would arise it indicates that one should contemplate entering a constant term into the model. However, when performing logistic regression analysis, the above mentioned coefficient of determination is not exactly applicable. Instead there are a number of

[6]Note that it can be used for other types of restriction which will not be needed in this paper

different approaches, which one should use though there is no consensus about. [9] One popular approach is the *McFadden Pseudo* R^2. It is calculated as follows,

$$R^2_{MCF} = 1 - \frac{ln(L_M)}{ln(L_0)} \tag{11}$$

2.2.3 Errors

When dealing with statistical modelling errors of different sorts often arise. This section presents how one can identify and, possibly, mitigate them.

2.2.3.1 Multicollinearity A problematic situation, named *Multicollinearity* [8, p. 15], is when two (or more) covariates are linearly dependent, i.e. they are highly correlated. Something that is important to point out is that multicollinearity is not a problem when wanting to make predictions. However, when doing a structural interpretation regarding how the covariates impacts the response variable it may pose a big problem. The individual P-values might be large even though the covariate is of intuitive importance. Also, the confidence intervals for the covariates might be too large, and perhaps even include 0. There are different ways to identify/test for multicollinearity;

- If the estimated standard deviations for some covariates are very large [8, p. 15], something that also can be due to too few observations.
- Large changes in estimated covariate coefficients when covariates are added or removed. [10]
- Testing the *Variance Inflation Factor* (VIF), a measure that quantifies how much multicollinearity that exists in an *OLS*-regression model. The measure shows how much of the variance of a covariate that is due to multicollinearity. E.g. for $i = 1$ It is calculated as follows,

  $$X_1 = \alpha_0 + \alpha X_2 + \cdots + \alpha X_k + e \tag{12}$$

  Then the VIF for $\hat{\beta}_1$ is defined as follows [11],

  $$VIF(\hat{\beta}_1) = \frac{1}{1 - R^2_1} \tag{13}$$

  where R_1 being the *coefficient of determination* for the previously regression (12). What VIF's that are deemed good or bad depends solely on the situation at hand, requirements of significance and the importance of the covariate one is testing. A $VIF(\hat{\beta}_i) > 10$ is commonly referred to as a limit for high multicollinearity. [11]

When multicollinearity has been identified there are a number of ways to remedy it, of course, depending on what have caused it;

- The first approach is to take a look at the data you have, and evaluate whether if it is sufficient. Obtaining more data to regress over will produce more accurate estimates covariates, and so forth lead to smaller standard errors.
- Test dropping a covariate, to see if one gets more significant results for the other covariates. This will, of course, lead to the model having less information, and perhaps biased covariate estimates for the remaining covariates which were correlated with the omitted.
- Look into, if there are any, dummy variables so that you have not fallen into the *Dummy Variable Trap* [12], as it is commonly named. If one enter a set of dummies which accounts for all possible outcomes amongst the observations, e.g. female and male when observations will possess either attribute, one will instead achieve perfect multicollinearity. To remedy this you either can drop the intercept or one of the dummies. The dropped dummy will then be the base case that you compare the others against.

3 Methodology

3.1 Set-up of the Markovian Model

Assuming the *memoryless*-ness of the states, set pieces and passes, we can set up a transition matrix representing the Markov process associated with these states. Our aim being to derive the probabilities for the process to enter one of the set piece-states (i.e Throw In, Free Kick, Goal kick or Corner).

3.1.1 The Transition Matrix

It is safe to assume that players have different incentives for entering a specific state depending of the field-location of the current state. For example it is understandable that a free kick in an offensive position is more likely to result in an offensive pass or a goal kick than a pass in a defensive position. Hence, for us to enhance the accuracy of our models, we divide some of the states[7] into three sub-states; Offensive (A), Defensive (D) and Safe/Central possession (C; representing a midfield possession).The transition matrix will thus be as follows,

$$\begin{bmatrix}
p_{p_cp_c}(\theta) & p_{p_cp_a}(\theta) & p_{p_cp_d}(\theta) & p_{p_ct_c}(\theta) & p_{p_ct_a}(\theta) & p_{p_ct_d}(\theta) & p_{p_cf_c}(\theta) & p_{p_cf_a}(\theta) & p_{p_cf_d}(\theta) & p_{p_cg}(\theta) & p_{p_cc}(\theta) \\
p_{p_ap_c}(\theta) & p_{p_ap_a}(\theta) & p_{p_ap_a}(\theta) & p_{p_at_c}(\theta) & p_{p_at_a}(\theta) & p_{p_at_d}(\theta) & p_{p_af_c}(\theta) & p_{p_af_a}(\theta) & p_{p_af_d}(\theta) & p_{p_ag}(\theta) & p_{p_ac}(\theta) \\
p_{p_dp_c}(\theta) & p_{p_dp_a}(\theta) & p_{p_dp_a}(\theta) & p_{p_dt_c}(\theta) & p_{p_dt_a}(\theta) & p_{p_dt_d}(\theta) & p_{p_df_c}(\theta) & p_{p_df_a}(\theta) & p_{p_df_d}(\theta) & p_{p_dg}(\theta) & p_{p_dc}(\theta) \\
p_{t_cp_c}(\theta) & p_{t_cp_a}(\theta) & p_{t_cp_a}(\theta) & p_{t_ct_c}(\theta) & p_{t_ct_a}(\theta) & p_{t_ct_d}(\theta) & p_{t_cf_c}(\theta) & p_{t_cf_a}(\theta) & p_{t_cf_d}(\theta) & p_{t_cg}(\theta) & p_{t_cc}(\theta) \\
p_{t_ap_c}(\theta) & p_{t_ap_a}(\theta) & p_{t_ap_a}(\theta) & p_{t_at_c}(\theta) & p_{t_at_a}(\theta) & p_{t_at_d}(\theta) & p_{t_af_c}(\theta) & p_{t_af_a}(\theta) & p_{t_af_d}(\theta) & p_{t_ag}(\theta) & p_{t_ac}(\theta) \\
p_{t_dp_c}(\theta) & p_{t_dp_a}(\theta) & p_{t_dp_a}(\theta) & p_{t_dt_c}(\theta) & p_{t_dt_a}(\theta) & p_{t_dt_d}(\theta) & p_{t_df_c}(\theta) & p_{t_df_a}(\theta) & p_{t_df_d}(\theta) & p_{t_dg}(\theta) & p_{t_dc}(\theta) \\
p_{f_cp_c}(\theta) & p_{f_cp_a}(\theta) & p_{f_cp_a}(\theta) & p_{f_ct_c}(\theta) & p_{f_ct_a}(\theta) & p_{f_ct_d}(\theta) & p_{f_cf_c}(\theta) & p_{f_cf_a}(\theta) & p_{f_cf_d}(\theta) & p_{f_cg}(\theta) & p_{f_cc}(\theta) \\
p_{f_ap_c}(\theta) & p_{f_ap_a}(\theta) & p_{f_ap_a}(\theta) & p_{f_at_c}(\theta) & p_{f_at_a}(\theta) & p_{f_at_d}(\theta) & p_{f_af_c}(\theta) & p_{f_af_a}(\theta) & p_{f_af_d}(\theta) & p_{f_ag}(\theta) & p_{f_ac}(\theta) \\
p_{f_dp_c}(\theta) & p_{f_dp_a}(\theta) & p_{f_dp_a}(\theta) & p_{f_dt_c}(\theta) & p_{f_dt_a}(\theta) & p_{f_dt_d}(\theta) & p_{f_df_c}(\theta) & p_{f_df_a}(\theta) & p_{f_df_d}(\theta) & p_{f_dg}(\theta) & p_{f_dc}(\theta) \\
p_{gp_c}(\theta) & p_{gp_a}(\theta) & p_{gp_a}(\theta) & p_{gt_c}(\theta) & p_{gt_a}(\theta) & p_{gt_d}(\theta) & p_{gf_c}(\theta) & p_{gf_a}(\theta) & p_{gf_d}(\theta) & p_{gg}(\theta) & p_{gc}(\theta) \\
p_{cp_c}(\theta) & p_{cp_a}(\theta) & p_{cp_a}(\theta) & p_{ct_c}(\theta) & p_{ct_a}(\theta) & p_{ct_d}(\theta) & p_{cf_c}(\theta) & p_{cf_a}(\theta) & p_{cf_d}(\theta) & p_{cg}(\theta) & p_{cc}(\theta)
\end{bmatrix}$$

Figure 2: Transition Matrix $\boldsymbol{P}$

where p_c is the pass in safe zone state, p_a pass in offensive zone state, p_d pass in defensive zone state and so forth[8]. Thus, the probability space will be,

$$\Omega = \{p_c, p_a, p_d, t_c, t_a, t_d, f_c, f_a, f_d, g, c\}$$

Here $\boldsymbol{\theta}$ is the parameter vector containing the factors that could induce modification of the transition probabilities, something we will discuss in the following section.

[7] All excluding Goal kicks and Corners, since they have definitive field-locations

[8] where t stand for Throw In, f for Free Kick, g for Goal Kick, c for Corner

3.1.2 The Impact of Game Related Factors

In this section we will put forward the factors that we believe may have an impact on the transition probabilities.

Proposition 1: *Time has an impact on some, if not all, transition probabilities*

This is a intuitive assumption that we have found to be confirmed when working with football set pieces. Indeed it is understandable that the mentality of the players (depending on whether they have the lead or not) change due to the decrease in remaining play-time. For example, it is safe to assume that a loosing team will try to get the ball to the goal area faster if time is scarce, affecting the *Defensive Possession* to *Offensive Possession* transitions positively. Furthermore transitions from any state to *Throw In* states are likely to be reduced since players will try to not create any time-loosing moments. One could go on and on, drawing a large number of logical conclusions on how the probabilities change depending on what the time of the match is. We will later interpret our regression results and try to explain different phenomena with logical inferences based on our knowledge of football. This will be the first factor, θ_{time}, of our parameter vector.

Proposition 2: *The score has an impact on some, if not all, transition probabilities*

This is believed to be true for every sport. Although it may not have an impact on all transition probabilities, the fact that frustration and increased defensive/offensive play occur will change the transition probabilities. Given the three score lines {3-2,1-1,0-4} one can fairly easily determine that the motives and incentives for the players to perform and taking decisions will differ; trailing by four goals is not as motivating as trailing by one.

Some concerns have arisen regarding a possible multicollinearity between the score and the time covariates. Indeed it seems rather evident that as time goes the propensity of goals increases. Hence, we have instead chosen to take the score difference as second factor, θ_{score}, in our parameter vector.

3.1.3 The Impact of Team Related Factors

The factors specific for each moment in the game are not the only one we believe have an impact on the transition probabilities. In this section we discus the possible team related factors that could induce a change of said probabilities.

Proposition 3: *Differences in team quality have an impact on some transition probabilities*

We believe that differences in team quality will affect the probabilities for the occurrence of the set pieces. If one team is significantly better than the other, they will retain

possession more, hence passing the ball around more, increasing the probability for such transitions.

It remains the problem of assessing the quality of the different teams. Since a lot of factors specific to each game and, to some extent, unrelated to the success rate of the team is to be taken into account, doubts in our ability to rate the teams in an objective manner arise. Even if we manage to create a complete classification/ranking of the teams (which would require an enormous amount of time) there is always the possibility of a variation of the team's *quality* depending on their opponents' characteristics. A good way to bypass this issue is to use the *pre-game* odds offered by open markets like the *Betfair Exchange* (See 3.2.2) which can be translated into probabilities.

To transform these odds into a unique variable we will resize the odds for team 1 respectively team 2 to win by changing the probability space from three to two events[9] in the following way,

$$c = \frac{a}{(a+b)} \qquad d = \frac{b}{(a+b)}$$

where a and b are the odds for team 1 respectively team 2 to win, c and d the reallocated odds for team 1 respectively team 2 to win. Then, by taking the absolute value of the difference of their inverse (which represent the probabilities), we can compute the value, θ_{odds},

$$\theta_{odds} = \left| \frac{1}{c} - \frac{1}{d} \right|$$

reflecting the *quality difference* of the teams thus giving us the third parameter of our $\boldsymbol{\theta}$ vector.

$$\boldsymbol{\theta} = \{\theta_{time}, \theta_{score}, \theta_{odds}\}$$

3.2 Data Collection

3.2.1 The States and Game Related Covariates

The data required for this research consist of the Throw Ins, Free Kick, Goal Kicks, Corners, and Goals that occurs during a game and some notion of possession. This with their respective time of occurrence, team possession and field location. The dataset we have procured ourselves, and its source need be confidential since this type of data is extremely valuable; something which has been approved with Prof. Boualem Djechiche,

[9] Excluding the draw, commonly referred to as *Draw-No-Bet*

our supervising professor. It contains one season of the Barclays Premier League (2011-12) and displays a wide range of game events all from set-pieces to passes and shots[10] and their respective data[11],

type	pos	date	Ateam	Bteam	period	time	scoreA	scoreB	redsA	redsB	gamezone	odds
pass	B	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	0	0	0	0	0	S	0.3953488
pass	B	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	0	0	0	0	0	S	0.3953488
pass	B	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	0	0	0	0	0	S	0.3953488
pass	B	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	0	0	0	0	0	S	0.3953488
pass	A	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	0	0	0	0	0	S	0.3953488
pass	B	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	0	0	0	0	0	S	0.3953488
pass	A	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	0	0	0	0	0	S	0.3953488
pass	B	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	0	0	0	0	0	S	0.3953488
pass.fk	A	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	0	0	0	0	0	S	0.3953488
pass	A	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	0	0	0	0	0	A	0.3953488
pass	B	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	1	0	0	0	0	S	0.3953488
pass	B	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	1	0	0	0	0	A	0.3953488
pass	B	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	1	0	0	0	0	D	0.3953488
pass	A	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	1	0	0	0	0	S	0.3953488
pass	A	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	1	0	0	0	0	S	0.3953488
pass	A	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	1	0	0	0	0	S	0.3953488
pass	A	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	1	0	0	0	0	S	0.3953488
pass	A	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	1	0	0	0	0	S	0.3953488
pass	B	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	1	0	0	0	0	S	0.3953488
pass.ti	A	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	1	0	0	0	0	S	0.3953488
pass	A	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	1	0	0	0	0	S	0.3953488
pass	A	2011-08-13	Blackburn Rovers	Wolverhampton Wanderers	1	1	0	0	0	0	S	0.3953488

Figure 3: Snap of our dataset

which we have reorganised to our purposes by:

- Deleting GameID and PlayerID
- Removing the events[12] that do not fall under our probability space.
- Assigning possession states (std) to all passes.
- Dividing them into Offensive (A), Defensive (D) an Safe (C) by combining the team-possession and the field position of the event, e.g. a pass by team A near their own goal will be a Defensive pass.

Figure 4: Illustrating how we have divided the field

- Discretise of the time to minutes and adding the θ_{odds} variable.

[10]Events: Passes, Shots, Throw Ins, Free Kicks, Goal Kicks, Corners, Offsides, Penalty, Red/Yellow Cards, Goal, Fouls

[11]Variables: Possession, Game Id, Teams, Period, Time, Player Involved, Score of both teams, number of red cards, Field Location

[12]Shots, Yellow/Red Cards, Fouls, Offsides, Breaks and Goals

3.2.2 The Match Odds

The *market based* assessment of the probabilities of respective teams to win is, in our opinion, the most objective way to estimate the *true* probabilities. Several ranking of premier league teams exists but that in itself is an argument for not using them. Moreover these rankings can impossibly take into account information about the line-up[13], the importance of the game or the team's specific strength/weakness relative to their opponent. Betting odds are much more dynamic and are updated on a minute basis depending on the game information that is publicly released. We have chosen to use the odds from the *Betfair Exchange* instead of the ones of a classic bookmaker which could be biased by the tweaking of odds they exerts for risk management purposes.

Indeed the *Betfair Exchange* person-to-person betting is considered to be a significantly more efficient market than the bookmaking one [13]. It offers a quite different betting environment to the one of conventional bookmakers. They act as actual markets (much like a stock exchange) for the *selling* (Book) and *buying* (Lay) of odds by making it possible for individuals to offers and place bets (see Figure 5, red side is for laying bets blue is for placing).

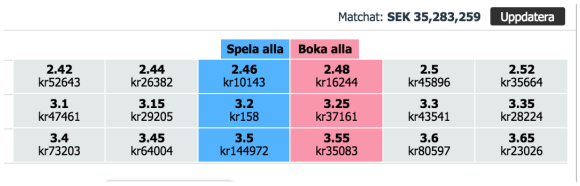


Figure 5: Snap of a Betfair Exchange Market

The *Betfair Exchange* provides any client holding a minimum of 100 *Betfair points*[14] with historical data consisting of all the odds for any betting market dating back to 2004. Being punters ourselves we have gained access to that information. Thus we have extracted the *Match Odds*, at the start time of the game, of all the games contained in our dataset.

For the other types of markets we could have used for more specific characterisation of the teams, we have some concerns about the liquidity of these relatively small markets [14]. Indeed, a lack of supply/demand on these market could considerably bias the odds. However, this is not a concern since match odds is the most popular market and have great liquidity [14]. This is especially true for the *Barclays Premiere League* [15] (as you can note in Figure 4; see the market depth under the odds).

[13] which is released one hour before every games

[14] A measure of client loyalty, see http://data.betfair.com/

3.3 Estimating the Transition Probabilities

In this section we put forward the different models we will compare for estimating the 121 different transition probabilities.

3.3.1 Counting Procedure

A counting procedure is a very straightforward method for estimating transition probabilities. Starting by dividing the data by the state that preceded them and then counting the different transitions we get the amount, n_{ij}, of transitions from a state, i, to another state j. Then by dividing it with the total jumps out of state, i, we get the transition probability,

$$P(X(t_{n+1}) = j | X(t_n) = i) = \frac{n_{ij}}{\sum\limits_{k \in \Omega} n_{ik}}$$

This method will be used for transitions where no correlation between their probability and the aforementioned factors can be established or when the sample size doesn't allow us to regress a probability.

3.3.2 Logit Regression

To asses the impact of the aforementioned factors we will regress the transition probabilities using a Logit regression. As previously mentioned, the Logit regression is well suited for our needs since it regresses a probabilistic response variable. The covariates to be used in this model will consist of the factor vector $\boldsymbol{\theta}$ (see 3.1.3). For the regression we will use the computer language R, which offers a default *glm*[15] function for Logit regression[16]. For this we will, like for the counting procedure, divide the events by state that preceded them so to have a dataset for every row of the transition matrix on which we will regress said probabilities. The model will thus be,

$$y = \frac{e^{\beta + x_1\theta_{time} + x_2\theta_{score} + x_3\theta_{odds}}}{1 + e^{\beta + x_1\theta_{time} + x_2\theta_{score} + x_3\theta_{odds}}} \tag{14}$$

where β is the intercept. In the next section we describe how we will compare the different models for choosing the one best suited to each transition.

3.3.3 Choosing the Model for Different Transitions

Due to the large amount of regression to be analysed and interpreted[16] we will be in need of a structured procedure for choosing the model for the prediction of different state-transitions. We will begin by investigating if there are transitions which are not

[15] Generalised Linear Models

[16] 111 transitions to be regressed with different possible models

possible or are, at least, highly unlikely to discard them and thus reduce the number of transitions to regress on. Furthermore we will undertake several tests for the choosing of suitable model for each transition. This will be done stepwise:

1. The first step will be to regress on all of the transition probabilities with our initial Logit model containing all of the covariates of the parameter vector $\boldsymbol{\theta}$. By performing Wald-tests(See 2.2.2.2) on every covariate we will identify the transitions where we cannot, within a 90% confidence interval, reject the null hypothesis, i.e. that the coefficient linked to a certain covariate is zero. We will then discard all models where there is not, at least, one coefficient whose p-value is lesser than 0.10 and apply the counting procedure model for the estimation of their probabilities.

2. After having scaled down the number of regressions to be analysed we will proceed with eventual reduction of the models by analysing the relevance of the covariates in the model. The AIC-test will be used for a preliminary choosing of the most suitable combination of covariates θ_i in each model.

3. The next step will be for us to assure ourselves that no multicollinearity exists between the chosen covariates. This will be done by performing VIF-tests on the remaining models to be used. This is a cautionary measure since we are fairly confident that the transformation of our variables, i.e. score to score difference will exonerate us from any time-score correlation. A possible source for concern is the possible correlation between difference in quality and score difference.

4. Furthermore a log-likelihood ratio test, with a null-restriction[17] will be performed for reconfirm the previous Wald-testing. We will here increase the confidence level to 95% to identify if any remaining model should be analysed in depth.

5. For the regressions that do not meet the above requirement we will once more perform a Wald-test on each covariate of said regressions and set a requirement of 95% confidence interval for rejecting the null hypothesis. The regression where the probability of the null hypothesis of the coefficients do not fall within this interval will be discarded and will be changed by a counting model.

6. Finally a Goodness of fit test will be used for us to interpret to what level our models describe the data at hand. This will later be used for assessing the quality of the model and how it might be improved.

3.3.4 The Final Transition Matrix

When we have chosen the models (Logit regression models with different covariates and counting procedure) we will compute a Matrix in R containing prediction functions for the transitions where we have chosen regression models for and fixed probabilities for

[17] Null hypothesis leaving only the intercept, i.e. testing for all the estimated covariate coefficient being equal to zero

the others. We will henceforth denote M1 the counting model, M2 the Logit regression model with $\boldsymbol{\theta}$ as covariates, M3 the one with θ_{time} and θ_{score} as covariates, M4 the one with θ_{time} and θ_{odds} as covariates, M5 the one with θ_{score} and θ_{odds} as covariates, M6 the one with θ_{time} as sole covariate, M7 the one with θ_{score} as covariate and M8 the one with θ_{odds} as covariate. In the same way we will denote P1, P2, P3, P4, P5, P6, P7 and P8 the transitions estimated by the M1, M2, M3, M4, M5, M6, M7 and respectively M8 model.
A problem arise: let one row of the transition matrix be composed of some factor dependent probabilities and some constant ones. If the P2-8 transition change the row-sum of probabilities would yield a different value than 1, which is not consistent with Markov theory. Thus we have to rephrase the M1-model for it to take into account this constraint and normalise the probability space by reallocating the difference $(1 - P_{row})$ onto the transition probabilities proportionally to their weight,

$$M1: \quad P(X(t_{n+1}) = j | X(t_n) = i) = \frac{n_{ij}}{\sum\limits_{k \in \Omega} n_{ik}} \left(1 + \frac{1 - P_{row_i}}{P_{row_i}}\right)$$

$$M2: \quad y = \frac{e^{\beta + x_1\theta_{time} + x_2\theta_{score} + x_3\theta_{odds}}}{1 + e^{\beta + x_1\theta_{time} + x_2\theta_{score} + x_3\theta_{odds}}} \left(1 + \frac{1 - P_{row_i}}{P_{row_i}}\right)$$

$$M3: \quad y = \frac{e^{\beta + x_1\theta_{time} + x_2\theta_{score}}}{1 + e^{\beta + x_1\theta_{time}}} \left(1 + \frac{1 - P_{row_i}}{P_{row_i}}\right)$$

$$M4: \quad y = \frac{e^{\beta + x_1\theta_{time} + x_2\theta_{odds}}}{1 + e^{\beta + x_1\theta_{score}}} \left(1 + \frac{1 - P_{row_i}}{P_{row_i}}\right)$$

$$M5: \quad y = \frac{e^{\beta + x_1\theta_{score} + x_2\theta_{odds}}}{1 + e^{\beta + x_1\theta_{odds}}} \left(1 + \frac{1 - P_{row_i}}{P_{row_i}}\right)$$

$$M6: \quad y = \frac{e^{\beta + x_1\theta_{time}}}{1 + e^{\beta + x_1\theta_{time}}} \left(1 + \frac{1 - P_{row_i}}{P_{row_i}}\right)$$

$$M7: \quad y = \frac{e^{\beta + x_1\theta_{score}}}{1 + e^{\beta + x_1\theta_{score}}} \left(1 + \frac{1 - P_{row_i}}{P_{row_i}}\right)$$

$$M8: \quad y = \frac{e^{\beta + x_1\theta_{odds}}}{1 + e^{\beta + x_1\theta_{odds}}} \left(1 + \frac{1 - P_{row_i}}{P_{row_i}}\right)$$

where $P_{row_i} = \sum\limits_{k \in \Omega} p_{ik}$ is the row-sum of all probabilities[18] on the row i. Now that we have normalised the probability space our transition matrix $\boldsymbol{P}$ is ready.

[18] With simple counting probabilities ex ante modifications

3.4 Computation of the *RF-1* Odds

For computing the *RF-1* Odds we will use the Markov model we have set up by calculating the probabilities of one of the four events; Throw In, Free Kick, Goal Kick or Corner; first happening.

3.4.1 Probability Distribution at End of the Proxy Period

Using the Chapman Kolmogorov equation (see equation 5.d) we can derive the expected distribution for the states after the 50 seconds proxy (see 1.3.1). We will for this need to estimate the mean number of transitions, μ, that occur during 50 seconds. This value will be specific to every game and has to be calculated in game although start μ will be a global mean taken from historical data.

To initialise this computation we will need information on the current state of the process which we put in vector form. This vector is the *start vector*, $\boldsymbol{u}$, of our process,

$$\boldsymbol{u} = (u_{p_c}, u_{p_a}, u_{p_d}, u_{t_c}, u_{t_a}, u_{t_d}, u_{f_c}, u_{f_a}, u_{f_d}, u_g, u_c)$$

where u_i yield the value 1 if the process is in state i and zero if not. The start vector is then multiplied by the transition matrix, P, μ times,

$$\boldsymbol{u}\boldsymbol{P}^{\mu} = \boldsymbol{u'} \tag{15}$$

thus giving us the expected distribution $\boldsymbol{u'}$ which will be used as start vector in the following step.

3.4.2 Probability of Next *RF-1* Event

The next step is for us to, from the expected distribution $\boldsymbol{u'}$, get the probabilities of the process first entering on of the set-piece events. For this we set all of these states as absorbing ones (see 2.1.2) to get the matrix $\boldsymbol{A}$,

$$\begin{bmatrix}
p_{p_cp_c}(\theta) & p_{p_cp_a}(\theta) & p_{p_cp_d}(\theta) & p_{p_ct_c}(\theta) & p_{p_ct_a}(\theta) & p_{p_ct_d}(\theta) & p_{p_cf_c}(\theta) & p_{p_cf_a}(\theta) & p_{p_cf_d}(\theta) & p_{p_cg}(\theta) & p_{p_cc}(\theta) \\
p_{p_ap_c}(\theta) & p_{p_ap_a}(\theta) & p_{p_ap_a}(\theta) & p_{p_at_c}(\theta) & p_{p_at_a}(\theta) & p_{p_at_d}(\theta) & p_{p_af_c}(\theta) & p_{p_af_a}(\theta) & p_{p_af_d}(\theta) & p_{p_ag}(\theta) & p_{p_ac}(\theta) \\
p_{p_dp_c}(\theta) & p_{p_dp_a}(\theta) & p_{p_dp_a}(\theta) & p_{p_dt_c}(\theta) & p_{p_dt_a}(\theta) & p_{p_dt_d}(\theta) & p_{p_df_c}(\theta) & p_{p_df_a}(\theta) & p_{p_df_d}(\theta) & p_{p_dg}(\theta) & p_{p_dc}(\theta) \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

Figure 6: Transition Matrix, $\boldsymbol{A}$, with Absorbing States

Much like in (15) we will here use the Chapman Kolmogorov equation (see equation 5.d) and multiply the expected distribution $\boldsymbol{u'}$ with the matrix $\boldsymbol{A}$ n times,

$$\boldsymbol{u}'\boldsymbol{A}^n = \boldsymbol{\pi} \tag{16}$$

in such a way that the probability of the process being in one of the pass states,

$$P_p = \sum_{i \in \{s,a,d\}} \pi_{p_i}$$

is deemed negligible. We will here set 0.01 as the upper bound probability. The probabilities of every event is then obtained by adding every location divided states and reallocating the remaining probability for passes onto the other ones in the following way,

$$\begin{aligned}
P_t &= \pi_{t_i} + \frac{\sum_{i \in \{s,a,d\}} \pi_{t_i}}{\sum_{i \in \Omega} \pi_i} P_p, \\
P_f &= \pi_{f_i} + \frac{\sum_{i \in \{s,a,d\}} \pi_{f_i}}{\sum_{i \in \Omega} \pi_i} P_p, \\
P_g &= \pi_g + \frac{\pi_g}{\sum_{i \in \Omega} \pi_i} P_p, \\
P_c &= \pi_c + \frac{\pi_c}{\sum_{i \in \Omega} \pi_i} P_p,
\end{aligned}$$

where P_i, $i \in \{t, f, g, c\}$, is the probability of the process first entering state i. Finally we simply invert these probabilities to obtain the odds.

4 Results

In this section we present the results obtained by following our methodology. For the estimations of the transition probabilities and the computation of the RF-1 odds we have used the computer language R a script language used specifically for statistical analysis.

4.1 Estimation of the Transition probabilities

4.1.1 M1 Matrix

We will begin by presenting the transition matrix with P1 probabilities, i.e. transition probabilities obtained by counting procedure, for identifying the transitions which yield a zero probability.

$$\begin{bmatrix}
0.7323 & 0.1656 & 0.0176 & 0.0291 & 0.0130 & 0.0041 & 0.0196 & 0.0064 & 0.0006 & 0.0075 & 0.0029 \\
0.3088 & 0.4659 & 0.1152 & 0.0184 & 0.0143 & 0.0109 & 0.0157 & 0.0072 & 0.0023 & 0.0258 & 0.0119 \\
0.2579 & 0.2063 & 0.2440 & 0.0203 & 0.0179 & 0.0428 & 0.0242 & 0.0027 & 0.0027 & 0.0811 & 0.0786 \\
0.7379 & 0.1827 & 0.0132 & 0.0183 & 0.0105 & 0.0039 & 0.0202 & 0.0085 & 0 & 0.0027 & 0.0015 \\
0.2841 & 0.5255 & 0.1021 & 0.0154 & 0.0126 & 0.0147 & 0.0147 & 0.0084 & 0.0042 & 0.0140 & 0.0042 \\
0.1118 & 0.3291 & 0.4284 & 0.0136 & 0.0136 & 0.0240 & 0.0261 & 0.0021 & 0.0073 & 0.0188 & 0.0219 \\
0.6070 & 0.2279 & 0.0564 & 0.0244 & 0.0120 & 0.0099 & 0.0259 & 0.0049 & 0.0019 & 0.0209 & 0.0069 \\
0.2541 & 0.3398 & 0.1684 & 0.0135 & 0.0180 & 0.0270 & 0.0225 & 0.0060 & 0 & 0.0962 & 0.0300 \\
0.2538 & 0.1769 & 0.1461 & 0.0153 & 0.0153 & 0.0461 & 0.0153 & 0 & 0 & 0.1307 & 0.1461 \\
0.6359 & 0.2495 & 0.0074 & 0.0358 & 0.0189 & 0.0019 & 0.0303 & 0.0134 & 0.0004 & 0.0044 & 0.0009 \\
0.2137 & 0.1882 & 0.2832 & 0.0229 & 0.0245 & 0.0254 & 0.0432 & 0.0025 & 0.0016 & 0.1153 & 0.0585
\end{bmatrix}$$

Figure 7: Transition Matrix $\boldsymbol{P}$

Here we see that $p_{f_d f_a}$, $p_{f_d f_d}$, $p_{f_a f_d}$ and $p_{t_s f_d}$ are equal to zero which reduces the number of transition to be estimated.

4.1.2 The Wald-test

When regressing the remaining transition probabilities with M2 we have discarded the ones whose regression do not yield, at least, one coefficient[19] with a z-value less than 10%. Indeed, if no coefficient has a z-value under 10% we cannot reject the null hypothesis of the corresponding coefficient being zero, i.e. the covariate having no significance for the model. We display the result in the following matrix which we will henceforth refer to as a *transition model matrix* $\boldsymbol{M}$,

[19] Intercept excluded

$$\begin{bmatrix} U & U & U & U & U & M_1 & U & M_1 & M_1 & U & U \\ U & U & U & M_1 & M_1 & M_1 & M_1 & U & M_1 & U & M_1 \\ M_1 & M_1 & U & M_1 & M_1 & M_1 & M_1 & U & M_1 & M_1 & M_1 \\ U & U & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & 0 & U & M_1 \\ M_1 & M_1 & U & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 \\ M_1 & M_1 & U & M_1 & M_1 & M_1 & M_1 & U & M_1 & M_1 & M_1 \\ U & M_1 & U & U & M_1 & M_1 & M_1 & M_1 & M_1 & U & M_1 \\ M_1 & U & U & M_1 & M_1 & M_1 & M_1 & M_1 & 0 & M_1 & M_1 \\ M_1 & M_1 & M_1 & M_1 & M_1 & U & M_1 & 0 & 0 & M_1 & M_1 \\ U & U & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & U \\ M_1 & M_1 & M_1 & M_1 & U & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 \end{bmatrix}$$

Figure 8: Transition Model Matrix $\boldsymbol{M}_1$

where $M1$ stands for the counting model presented in 3.3.4., which is to be used for respective transition probabilities, and U stands for unknown model which can be any of the M2-8 models. The next step will be to determine which one of these models is most suited for the transition at hand.

4.1.3 Reducing the models AIC test

To do this we have regressed all of the 32 remaining transition on each of the M2-8 models and compared their AIC values (see 2.2.2.3). The AIC value indicates the relative quality between models and an indication of which covariates to use in the modelling of said probabilities. The model with the lowest value holds the best quality relative to the others. The following table shows the different AIC-values for the transitions which are marked U in Figure 8 for the 7 different possible combination of covariates.

Trans.	M2	M3	M4	M5	M6	M7	M8
$p_{p_s p_s}$	**67749**	67769	67811	67821	67837	67844	67839
$p_{p_s p_a}$	52409	**52407**	52426	52456	52424	52454	52455
$p_{p_s p_d}$	**10370**	10375	10378	10378	10384	10384	10379
$p_{p_s t_s}$	15379	15406	15378	**15377**	15407	15404	15377
$p_{p_s t_a}$	8128.1	8139.4	8135	**8127**	8148.2	8138.7	8133.2
$p_{p_s f_s}$	11278	**11279**	11297	11286	11298	11286	11296
$p_{p_s g}$	5192.1	5190.1	5194.2	5190.7	5192.3	**5188.7**	5192.3

Trans.	M2	M3	M4	M5	M6	M7	M8
p_{p_sc}	2360	2363.7	2358.1	2359.9	2362	2363.9	**2358**
$p_{p_ap_s}$	32533	32538	**32534**	32555	32539	32559	32553
$p_{p_ap_a}$	3681	36382	**36379**	36382	36380	36383	36381
$p_{p_ap_d}$	18816	**18813**	188117	18823	18814	18820	18822
$p_{p_af_a}$	2243.4	2241.9	2242.1	2248	**2240.7**	2246.6	2246.1
p_{p_ag}	6307.1	6304.9	6306.6	6314	**6304.3**	6311.7	6312
$p_{p_dp_d}$	7697.9	7704.2	7696.3	7699	**7602.5**	7705.4	7697
$p_{p_df_a}$	259.52	265.23	258.87	**258.24**	264.11	264.04	259.01
$p_{t_sp_s}$	2959.8	2964.2	2957.9	2958.3	2962.3	2962.8	**2956.3**
$p_{t_sp_a}$	2449.1	2452	2447.2	2447.1	2450	2450	**2445.2**
p_{t_sg}	96.577	96.288	95.668	100.69	**95.594**	100.62	98.713
$p_{t_ap_d}$	944.03	947.83	942.13	942.7	946.03	946.63	**940.7**
$p_{t_dp_d}$	1311	1311.9	1309.1	1309	1309.9	1309.9	**1307.1**
$p_{t_df_a}$	28.06	31.679	28.693	26.061	32.234	29.682	**27.501**
$p_{f_sp_s}$	2680.9	2682.8	**2679.7**	2686.7	2681.9	2689.1	2684.8
$p_{f_sp_d}$	865.99	875.14	864.32	864.3	873.24	873.35	**862.44**
$p_{f_st_s}$	461.34	459.44	466.37	461.25	464.36	**459.31**	464.49
p_{f_sg}	404.98	403.03	404.66	413.73	**402.76**	411.79	411.73
$p_{f_ap_a}$	850.96	856.58	849	849.51	854.76	855.04	**847.84**
$p_{f_ap_d}$	601.16	608.44	599.17	599.31	606.45	606.55	**597.31**
$p_{f_dt_d}$	51.043	52.785	50.894	50.91	52.544	51.591	**49.506**
p_{gp_s}	2633.9	2636.6	2632	2632.2	2634.8	2634.8	**2630.6**
p_{gp_a}	2247.1	2261.5	2245.1	2245.2	2259.5	2259.5	**2243.2**
p_{gc}	34.372	36.419	34.023	33.103	35.595	34.917	**32.068**
p_{ct_a}	274.74	273.47	272.82	276.95	**271.54**	275.62	275.42

Table 1: AIC values of the Different Models

For the choice of covariates for respective transition probability model we summarise the results obtained in Table 1 in the following matrix by changing the transition probabilities by the model to use,

$$\begin{bmatrix} M_2 & M_3 & M_2 & M_5 & M_5 & M_1 & M_3 & M_1 & M_1 & M_7 & M_8 \\ M_4 & M_4 & M_3 & M_1 & M_1 & M_1 & M_1 & M_6 & M_1 & M_6 & M_1 \\ M_1 & M_1 & M_6 & M_1 & M_1 & M_1 & M_1 & M_5 & M_1 & M_1 & M_1 \\ M_8 & M_8 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & 0 & M_6 & M_1 \\ M_1 & M_1 & M_8 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 \\ M_1 & M_1 & M_8 & M_1 & M_1 & M_1 & M_1 & M_8 & M_1 & M_1 & M_1 \\ M_4 & M_1 & M_8 & M_7 & M_1 & M_1 & M_1 & M_1 & M_1 & M_6 & M_1 \\ M_1 & M_8 & M_8 & M_1 & M_1 & M_1 & M_1 & M_1 & 0 & M_1 & M_1 \\ M_1 & M_1 & M_1 & M_1 & M_1 & M_8 & M_1 & 0 & 0 & M_1 & M_1 \\ M_8 & M_8 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & M_8 \\ M_1 & M_1 & M_1 & M_1 & M_6 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 \end{bmatrix}$$

Figure 9: Transition Model Matrix $\boldsymbol{M}_2$

where M_i represents the selected model as described in 3.3.4.

4.2 Validation of the Regression Models

We will in this section proceed to several tests for the 32 transitions for which we have chosen to regress the probabilities.

4.2.1 The VIF-test

We begin by testing for the multicollinearity of the covariates of transition whose probability is modelled with several covariates, i.e. M2-5.

Trans.	$\theta time$	$\theta score$	$\theta odds$
$p_{p_s p_s}$	1.274010	1.279693	1.005480
$p_{p_s p_a}$	1.271385	1.271385	-
$p_{p_s p_d}$	1.244975	1.251058	1.006157
$p_{p_s t_s}$	-	1.005497	1.005497
$p_{p_s t_a}$	-	1.004574	1.004574
$p_{p_s f_s}$	1.239215	1.239215	-
$p_{p_a p_s}$	1.000090	-	1.000090
$p_{p_a p_a}$	1.000078	-	1.000078
$p_{p_a p_d}$	1.235212	1.235212	-
$p_{p_d f_a}$	-	1.006478	1.006478
$p_{f_s p_s}$	1.000149	-	1.000149

Table 2: VIF-test Results

As we can see in Table 2 all multivariate regressions yield very low ($<<$ 10, see 2.3.3.1)) VIF-values, thus it is safe to say that no multicollinearity exists between the different covariates.

4.2.2 Log-likelihood ratio

The next step is to test the log-likelihood of the models compare to the regression with the *null*-model regressing solely on the intercept.

Trans.	p-value	Trans.	p-value
$p_{p_s p_s}$	2.280811e-24	$p_{p_s p_a}$	1.351248e-11
$p_{p_s p_d}$	1.017026e-04	$p_{p_s t_s}$	1.023651e-07
$p_{p_s t_a}$	8.217796e-06	$p_{p_s f_s}$	1.133169e-05
$p_{p_s g}$	**5.736316*e*-02**	$p_{p_s c}$	1.503139e-02
$p_{p_a p_s}$	1.008443e-07	$p_{p_a p_a}$	5.180665e-04
$p_{p_a p_d}$	7.926483e-03	$p_{p_a f_a}$	1.393922e-02
$p_{p_a g}$	6.395749e-03	$p_{p_d p_d}$	**8.59601*e*-02**
$p_{p_d f_a}$	7.163972e-03	$p_{t_s p_s}$	1.067479e-02

Trans.	p-value	Trans.	p-value
$p_{t_s p_a}$	2.754313e-02	$p_{t_s g}$	2.422576e-02
$p_{t_a p_d}$	1.483587e-02	$p_{t_d p_d}$	4.727346e-02
$p_{t_d f_a}$	2.288513e-02	$p_{f_s p_s}$	1.283991e-03
$p_{f_s p_d}$	8.894682e-04	$p_{f_s t_s}$	2.290607e-02
$p_{f_s g}$	2.620068e-03	$p_{f_a p_a}$	3.347832e-03
$p_{f_a p_d}$	1.867370e-03	$p_{f_d t_d}$	**7.72366*e*-02**
$p_{g p_s}$	9.398821e-03	$p_{g p_a}$	3.886137e-05
p_{gc}	**5.85898*e*-02**	p_{ct_a}	3.243490e-02

Table 3: Log-Likelihood Ratios

As we can see in Table 3 the probability of the restrictions being statistically probable are sufficiently low (ranging from 2.280811e-24 to 4.727346e-02) for rejecting the null restrictions. Some exception are identified; the transitions p_s to g, p_d to p_d, g to c and f_d to t_d which yield probabilities over 5%. This calls for a closer look at these regressions.

Cov.	Estimate	SE	z-value	P<|z|
Intercept	-4.80502	0.06070	-79.163	<2e-16
scordiff	-0.09175	0.04935	-1.859	0.063

Table 4: Summary of Regression for $p_{p_s g}$

Cov.	Estimate	SE	z-value	P<|z|
Intercept	-1.216357	0.057552	-21.135	<2e-16
time	0.001795	0.001048	1.713	0.0868

Table 5: Summary of Regression for $p_{p_d p_d}$

Cov.	Estimate	SE	z-value	P<|z|
Intercept	-3.59867	0.57719	-6.235	4.52e-10
odd	0.13822	0.07058	1.958	0.0502

Table 6: Summary of Regression for $p_{f_d t_d}$

Cov.	Estimate	SE	z-value	P<|z|
Intercept	-8.17899	1.25363	-6.524	6.83e-11
odd	0.17708	0.08489	2.086	0.037

Table 7: Summary of Regression for p_{gc}

As we can see from Table 4 and 5 the regression on $p_{p_s g}$ respectively $p_{p_d p_d}$ yield probabilities of the null hypothesis,

$$H_0 : x_1 = 0$$

where x_1 is the coefficient of their covariate, being true which are significantly over the 5% boundary we have set ourselves. Thus we cannot reject it and must discard these in our estimation of the transition matrix $\boldsymbol{P}$. Hence we reformulate the transition matrix to hold the models for each transition according to the following transition model matrix $\boldsymbol{M}2$,

$$\begin{bmatrix}
M_2 & M_3 & M_2 & M_5 & M_5 & M_1 & M_3 & M_1 & M_1 & M_7 & M_8 \\
M_4 & M_4 & M_3 & M_1 & M_1 & M_1 & M_1 & M_6 & M_1 & M_1 & M_1 \\
M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & M_5 & M_1 & M_1 & M_1 \\
M_8 & M_8 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & 0 & M_6 & M_1 \\
M_1 & M_1 & M_8 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 \\
M_1 & M_1 & M_8 & M_1 & M_1 & M_1 & M_1 & M_8 & M_1 & M_1 & M_1 \\
M_4 & M_1 & M_8 & M_7 & M_1 & M_1 & M_1 & M_1 & M_1 & M_6 & M_1 \\
M_1 & M_8 & M_8 & M_1 & M_1 & M_1 & M_1 & M_1 & 0 & M_1 & M_1 \\
M_1 & M_1 & M_1 & M_1 & M_1 & M_8 & M_1 & 0 & 0 & M_1 & M_1 \\
M_8 & M_8 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1 & M_8 \\
M_1 & M_1 & M_1 & M_1 & M_6 & M_1 & M_1 & M_1 & M_1 & M_1 & M_1
\end{bmatrix}$$

Figure 10: Transition Model Matrix $\boldsymbol{M}_3$

where we have, for the transitions $p_{p_s g}$ and $p_{p_d p_d}$, shifted to a counting procedure model M1.

4.2.3 Goodness of Fit

For testing the goodness of fit of our models we use the McFadden R^2 statistic which can give an indication of the appropriateness of a regression model.

Trans.	R^2	Trans.	R^2
$p_{p_s p_s}$	0.0016675474	$p_{p_s p_a}$	0.0009543083
$p_{p_s p_d}$	0.0020292877	$p_{p_s t_s}$	0.0020897454
$p_{p_s t_a}$	0.0028753819	$p_{p_s f_s}$	0.0020163760
$p_{p_s g}$	0.0006990536	$p_{p_s c}$	0.0025055302
$p_{p_a p_s}$	0.0009895394	$p_{p_a p_a}$	0.0004158173
$p_{p_a p_d}$	0.0005141472	$p_{p_a f_a}$	0.0027355285
$p_{p_a g}$	0.0012294189	$p_{p_d p_d}$	0.0004536837
$p_{p_d f_a}$	0.0376823756	$p_{t_s p_s}$	0.0022031504
$p_{t_s p_a}$	0.0019854061	$p_{t_s g}$	0.0525320611
$p_{t_a p_d}$	0.0062970828	$p_{t_d p_d}$	0.0030111840
$p_{t_d f_a}$	0.1805259530	$p_{f_s p_s}$	0.0049555219
$p_{f_s p_d}$	0.0127024120	$p_{f_s t_s}$	0.0113303477
$p_{f_s g}$	0.0222465734	$p_{f_a p_a}$	0.0100974210
$p_{f_a p_d}$	0.0160460221	$p_{f_d t_d}$	0.0642039606
$p_{g p_s}$	0.0025615592	$p_{g p_a}$	0.0075022916
$p_{g c}$	0.1130307777	$p_{c t_a}$	0.0170486239

Table 8: Pseudo R^2 values

As we can see in Table 8 the values are very low with a maximum of 18% of variation being explained. Our initial reaction was that the model should be discarded without further investigation. But, when we think about the covariates incorporated in our model and the millions of factors which cannot be accounted for, it is quite obvious that much of the variance cannot be explained. One must assess what levels of R^2 that is significant and feasible relative to the proposed model. We read about a dissertation where the aim of the study was to see if religiosity could be used to predict predict physical health[17]. They used frequency of religious attendance to indicate religiosity. They got an $R^2 = 0.04$ and still the model was deemed significant. In conclusion one need to assess what parts that the model should explain, and in doing so we feel that the values we have got is

significant. Then why does our values differ so much; from only a tenth of a percent to 18%? That is partially explained by the large differences in sample size. The transitions containing the passes all contain more than 5000 observations; whilst throw in-s in a defensive position is less than a thousand.

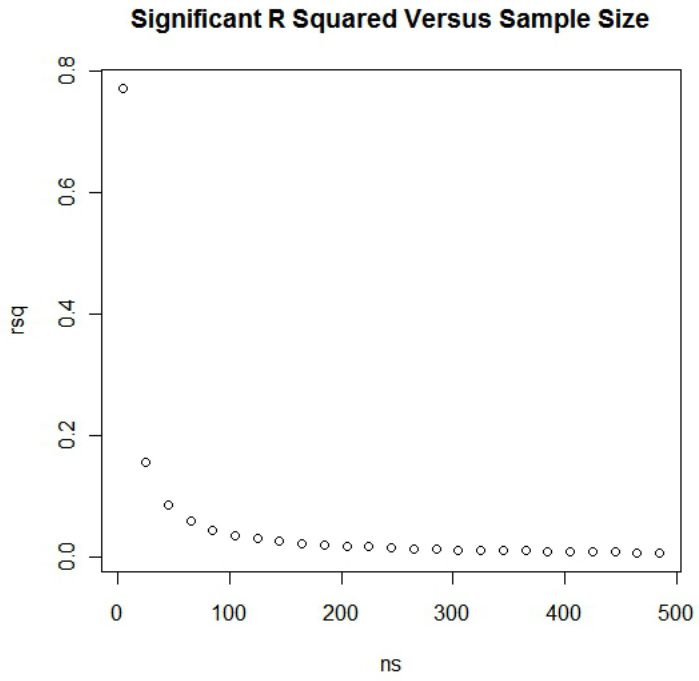


Figure 11: A graph showing what R^2 is significant as sample size increases.

Commenting on the above plot, this is something which is applied for OLS-regression. As there currently exists no consensus in the scientific society regarding R^2 and its validity for logistic regression, we offer this plot as something to think about rather than using it validating our results.

4.3 The Resulting Odds

Now that we have our final transition model matrix we can compute the transition matrix $\boldsymbol{P}$ with the prediction models according to $\boldsymbol{M}_3$. By following the methodology described in 3.4 we can compute the RF-1 odds on different fictive test games.

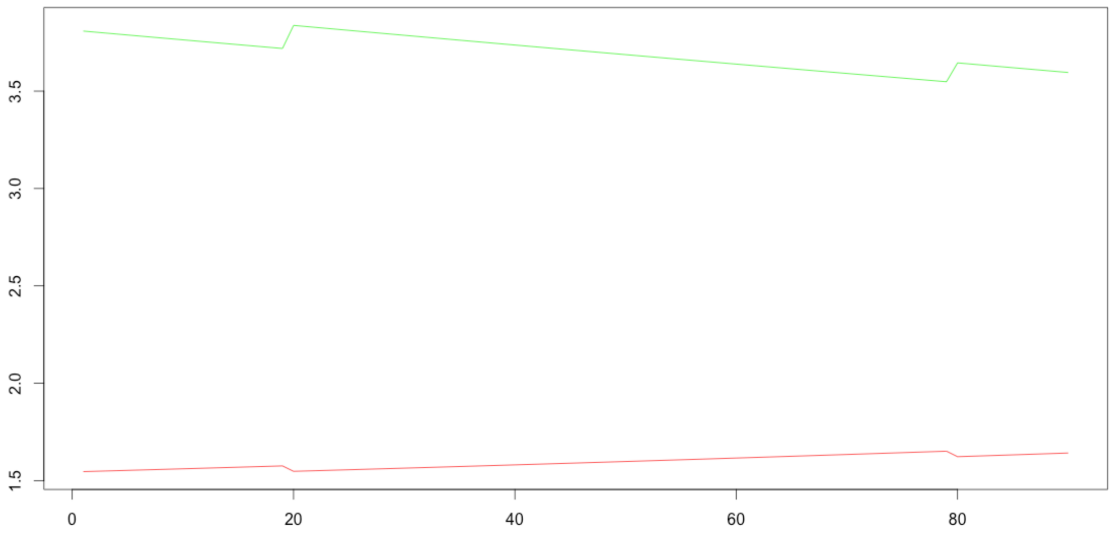


Figure 12: Throw In and Free Kick Odds variation on time and score

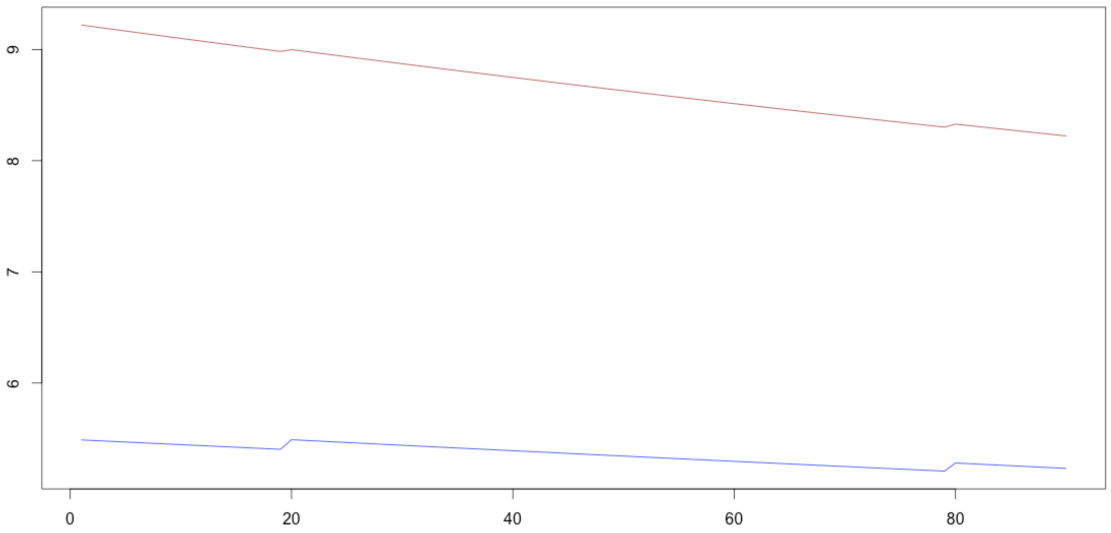


Figure 13: Goal kick and Corner Odds variation on time and score

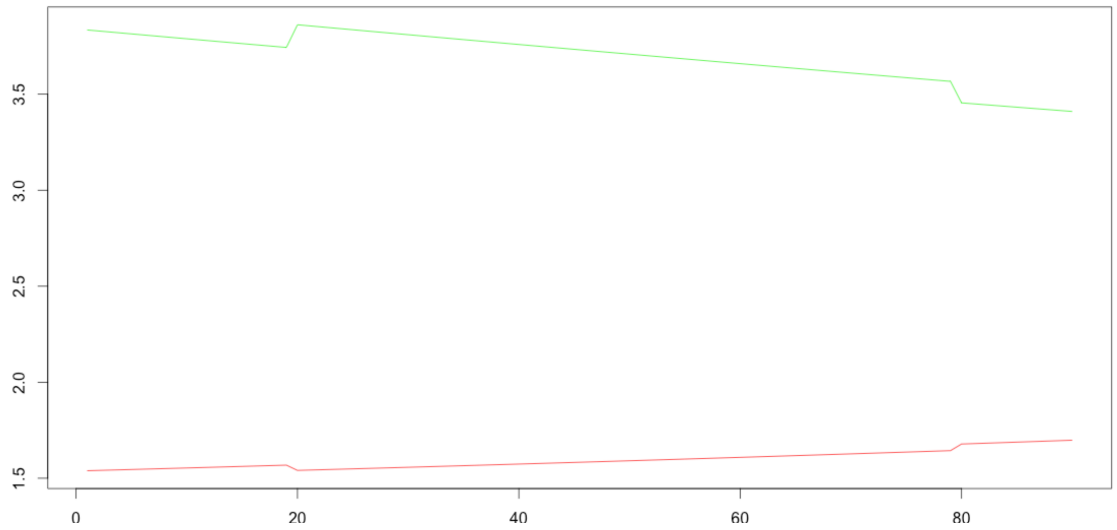


Figure 14: Throw In and Free Kick Odds variation on time and score

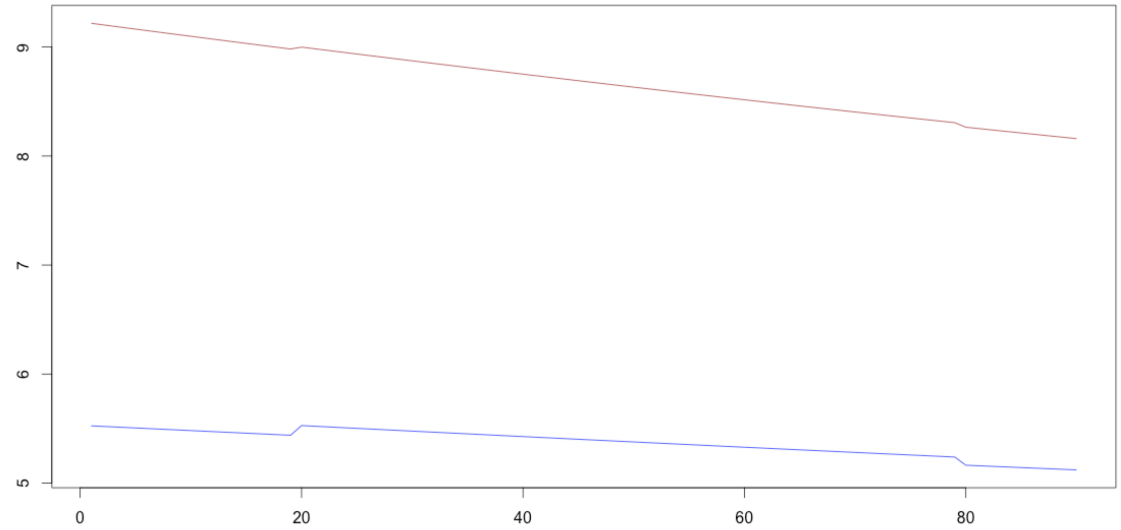


Figure 15: Goal kick and Corner Odds variation on time and score

Figure 11 and 12 shows the variation of the odds (green: Free Kick, red: Throw In, brown: Corner and blue: Goal Kick) during a game opposing two teams with a difference in probability of winning, θ_{odds}, of 0.7 and with a score difference going from 0 to 1 at minute 20 and from 1 to 2 at minute 80. Whereas Figure 13 and 14 shows the variation of the odds during a game opposing two teams with a difference in probability of winning, θ_{odds}, of 0.1 and with a score difference going from 0 to 1 at minute 20 and from 1 to 0 at minute 80. In both examples we have used a average, μ (see 3.4.1), of 10 transition per 50 seconds and a number, n (see 3.4.2), of 25 transitions.

5 Analysis

5.1 Mathematical

5.1.1 Overall Interpretation of the Transition Model Matrix

Although we wont proceed with an in-depth structural interpretation of each regression, a couple of general conclusions can be drawn from the $\boldsymbol{M}_3$ matrix on how the factors used in estimating our transitions have an impact on football games. We will in this sections confront these observation to intuitive logical explanation of the impact induced by the time, score difference and the quality difference.

5.1.1.1 The θ_{odds} Factor First and foremost we can conclude that the difference in quality of the teams is the dominant factor which is present, in a significant manner in 20 of the regression models. It affects almost all of the transitions to a defensive states. A intuitive explanation of this could the team with lowest quality fails to play up the ball to offensive positions and is forced to remain in a defensive zone. Furthermore the the dominant team might be more capable to intercept the ball when the opponent has an offensive possession thus increasing the number of transitions from offensive passes to defensive ones.

Trans.	$\theta odds$-Estimate
$p_{p_s p_d}$	-0.01992
$p_{t_a p_d}$	-0.05956
$p_{t_d p_d}$	0.03276
$p_{f_s p_d}$	-0.09389
$p_{f_a p_d}$	-0.09242
$p_{f_d t_d}$	0.13822

Table 9: Some Coefficients Estimates of θ_{odds}

Observing the values displayed in Table 9 we see that this intuition is not warranted. Although we cannot interpret to what degree the factor impact the probability without comparing their estimated value with the others we can conclude that the quality difference has a overall negative impact on the transitions to defensive states except for the $f_d t_d$ and $t_a p_d$ transition which might fall under the above explanation. We can thus conclude that the difference quality is more prone to reduce the transition probability to defensive states which might come from the fact that the dominant team stay in the safe/offensive states, e.g. the $f_d t_d$ transition yield a positive coefficient of 0.0104203, and that the lesser team aim to do counter attacks when they get the ball.

5.1.1.2 The θ_{time} and θ_{score} Factors

From figures 12 to 15 we can clearly observe the impact of the θ_{time} and θ_{score} factors on the RF-1 odds. While the probability of Throw In happening first is decreasing with time the Goal Kick, Corner and Free Kick ones are clearly increasing. Furthermore we can note that an increase in score difference increases the probability of Throw In happening and deceases the remaining ones.

5.1.2 Comparing Results with SuperLive Odds

When observing the resulting odds obtained with our model we are surprised by the low odds, or high probability, of the Throw In state happening first. Indeed the MetricGaming traders often price the Throw In within a range 1.7 to 2. It seems that the impact of time on the probability for Throw In happening is lesser than on the other set-pieces. Looking over our model matrix we can see that all but one[20] transition to a Throw In state have probabilities that depend on time. By analysing the modelling of these transitions we found that some of these, when modelled with time as sole covariate yielded rather good p-values. Although they did not meet the 95% confidence interval requirement we believe that an extended data set could reduce these p-values and maybe give time a bigger role in the absorption matrix thus making the predicted probability of Throw In happening not taking into account the *real* impact of time.

5.1.3 Propositions for Improvement

5.1.3.1 The High Unexplained Variance - More Factors Needed?

As presented in the results' section, we have obtained some quite low values of R^2 for our regressions. Hence it is safe to say that our proposed covariates fail in explaining the lion part of the variance of the data. Our conclusion is that this is due to a large number of factors which are very hard to quantify and incorporate into a model like ours.

- First and foremost one must assess how decisions are taken at any given point of the game, or to put it in the "lingo" of our project, any given state of the match. When a player receives a pass, he immediately thinks about what to do with the ball. He might be a player with great technique and flair, therefore he might be more inclined to drive the ball up field by himself. He might also be a master of the pass, with a great perception of the game; therefore he might opt to pass the ball to get into a dangerous position. He could also be affected by something that has happened earlier in the game; a missed opportunity or a dispute with an opponent, something which might have an impact on the decision he makes. We believe that a solution to this could be to implement the ratings system of players in the, very popular, video game; FIFA 15. EA Sports, who produces the series, do extensive quantitative analysis to derive these ratings which contains more than 30 different parameters such as flair, vision, speed, balance etc [18].

[20] The c to t_s is modelled by M_6 but holds a very low probability

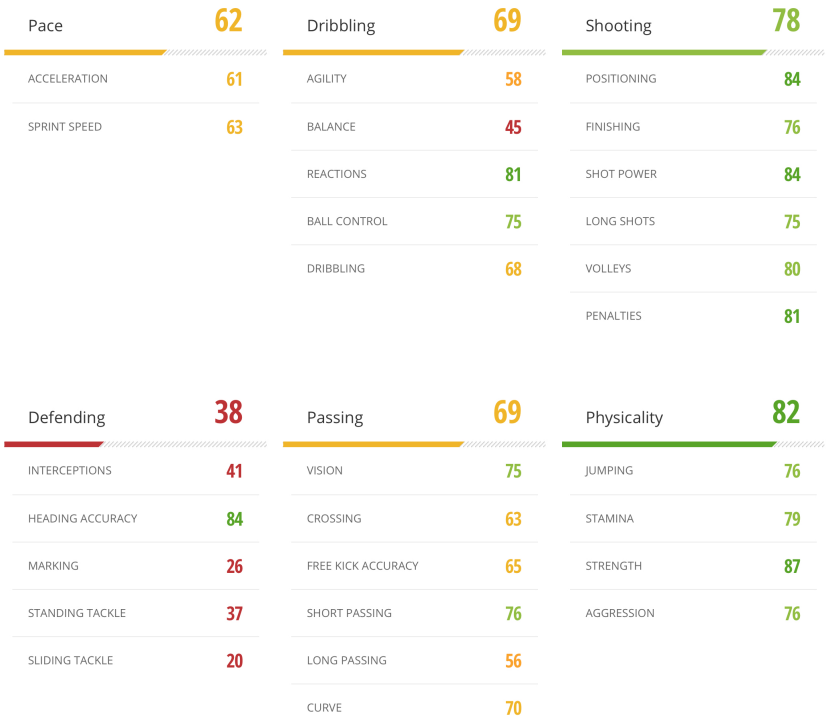


Figure 16: The ratings of Olivier Giroud by EA

- What more affects this decision making? Where the others are located on the pitch is the first thing that come to mind, but building a system/method which makes it possible for us to incorporate this factor into our model is beyond our scope both in terms of time and the necessary skill/experience. Acquiring data of that detail would probably cost millions of dollars.

- Another thought we have had is how the current game intensity affects the probability, and if it were something we could create covariates out of, we believe would be amongst the best possible predictors. Our proposition for this is to be use the distance covered by all players over the last minute(s) and see how this affects the probabilities. We believe that a low sum of distances would correlate with an increase of the probabilities, hence lowering the odds, for free kicks and throw ins while a high covered distance would correlate with more corners and goal kicks. Implementing this in a model which is to be used in real time, determining odds, would not work as it is impossible to access such data live; it is only distributed after games.

- Depending on tactics and so forth on player material, club culture and so on, it is clear that different teams play in different ways. We believe that this also could go a long way in bettering our model. We have, however, not come up with a solid proposition to how these could be incorporated into the model. We have discussed

studying the possibility of create team specific distributions containing their passes with how long/short they are, perhaps dividing length into 3/5/7 categories. We have also pondered the possibility of incorporating more odds-related covariates to describe playing style eg., who gets the next- corner/yellow card. We do however suspect significant multicollinearity with our already implemented odds covariate.

We have also thought about whether we can develop our currently employed odds variable further. As of now it only regards the relative, assumed, strength between two teams. Let the odds of each team winning a game be equivalent. It does not tell you whether it is two equally good teams that are facing each other or if they are just equally bad. By making some kind of historical analysis of pre-game odds for each team, we would be able to assign teams an absolute coefficient of strength which we do believe could improve upon our model.

We have also found a study which may of interest if one wants to pursue a more complete model. [19]

6 Industrial Application

In this section we will present the industrial application of our mathematical thesis which focuses on how the implementation of a pricing tool (using our modeling), should be conducted for reaching an optimal production process in terms of operational management within betting companies. We will base our analysis on *Metric Gaming* (MG), the company where we work and which produces RF-1 market (see 1.4.1). Several perspectives are to be taken into account in this discussion. We have identified three possible levels of implementation; *full-automatisation*, *semi-automatisation* and as *supporting tool*, all of which have different consequences on different parts of the company. We have chosen to focus our research on three perspectives, namely the risk management and business operations parts of the pricing process as well as the branding and marketing positions of the company. We begin with a brief introduction of the company.

6.1 Metric Gaming

Metric Gaming (MG) is a limited liability company based out of Las Vegas, Nevada with subsidiaries in Stockholm, Malta and Isle of Man. They provide customers with both the product being their SuperLive™ platform, where they offer unique betting markets as the *RF-1*, and by delivering trading and settling services. They distinguish themselves by their markets by taking the term *live betting* to a new meaning. Markets open and settle within a minute, hence giving end customers the opportunity to bet a large amount of times during a match.
MG is a business-to-business (B2B) company, who's prospective clients are the gambling companies with an online betting service eg., BWin, Betsson etc. The clients choose what level of integration they like with the platform and are so forth able to offer MG's unique markets as a part of their own product range. One of MG's unique selling points is, according to the founder in an interview with *SBC News*;

"This innovation will drive both customer acquisition and loyalty, particularly with the new generation of sports bettor that has grown up in today's smartphone and social media-driven society, where up-to-the-second information and results are now demanded and expected. SuperLive will also naturally increase player turnover because wagers are graded in real time, allowing winnings to be rolled over instantly – sometimes hundreds of times per match."[20]

To explain further, the possibility of having your end customer turn over his wagering capital many times a game complemented with gaining access to a product that is highly synergiseable with the evolution of the smart phone culture, is something that should be highly attractive among MG's prospective clients.

6.2 Methodology

Our methodology used to be able to answer our research question can be divided into two parts. We will primarily perform qualitative interviews, then complement them through researching earlier work and news articles in order to gain sufficient knowledge regarding our areas of interest within the sector. We will interview the CEO, Vice President of Trading and a trader at MG's subsidiary here in Stockholm to get their views on the mentioned (see introduction) three different perspectives. Studying the structure-of-interview framework provided to us by Anneli Linde, we define our interview methodology to be on the right hand side of the scale; we use a somewhat predefined structure of questions, all with open answers. For each perspective we ask questions revolving different sub-areas of the subject, all in order to obtain relevant insight. All will be consulted on the risk management perspective while the CEO will answer the questions regarding the remaining two perspectives. The question basis were prepared by us both while P. Lang conducted the interviews. Admittedly, we do already know a lot about our research areas ourselves since we are employed by the company. Therefore we feel it being of capital importance that we conduct these interviews with others so that we do not succumb to *subjective bias*[21], something which may occur when working so close with one case study.

Regarding risk management, we are specifically interested in what practices are employed in the market today and whether we can define them. We would also like to know what risks a bookmaker faces when pricing and thence draw conclusions to which of these may affect our proposed implementation or vice versa, thereafter analyse the implications to choose an optimal level of implementation from a risk management perspective.

For the perspective of cost reducing, we want to assert that we have the right understanding regarding how the production of the market, the RF-1, is conducted. We would, furthermore, like to get an idea for what MG's long term aim is in order cut costs in their production chain and what challenges lie in front of them in order to accomplish this.

We will finally argue that such a modeling could impact on the company's branding. Since the market they offer is innovative we will investigate the betting industry's position towards innovations and how our model could change their view on MG's product.

When we have shed light on the questions above we will analyse and evaluate which level of implementation that would be optimal, weighing the pros and cons of each perspective for each level against each other. While doing so, keeping in mind the goals (1.4.1) of the company; improve the product, improve the brand and minimise costs/increase profitability.

6.3 Theoretical Framework

6.3.1 Risk Management

Risk management as defined by the Financial Times's lexicon is the process of identifying and managing the uncertainties of business activity's outcomes. Several types of risk can be identified; strategic failures, operational failures, financial failures, market disruptions, environmental disasters, and regulatory violations, but for our purposes we will focus on the operational failures that might arise within the pricing process of betting companies.

Operational risk is defined as *"the risk of loss resulting from inadequate or failed processes or systems, human factors or external events"*[22]. We will not perform a quantitative estimation of these risks, but rather a qualitative one for which several methods exists in *operational risk measurement*.

Little literature addressing the way to manage risk for gambling companies exist. This is probably since betting company each have constructed their own operational risk management praxis based on different strategies and are very protective of them. *Ladbrokes*, a major actor within the betting industry [23], present in their business [24] review the different types of operational risk they expect to encounter.

- Trading, liability management and pricing: The losses induced by a failure in the pricing process i.e odds not being *"determine accurately the odds in relation to any particular event and/or any failure of its risk management processes."*
- High fixed cost base: When considering a possible drop in revenue, Ladbrokes explain that the fact that they operate with significant fixed costs; i.e. *"employee, rental and content costs associated with its betting shop estate"*, could greatly impact their profitability.
- Loss of key locations: Here they mean to address the risk of centralisation of their operations. Changes in legislature in Gibraltar for example, where the online betting and gaming operations are based, could lead to a need of costly re-localisation.

We will in our analysis proceed from the two first points of this review, since they are the type of risk that might be encountered when changing the production process of the company. We address this in the following section where we present the results of our investigation.

6.3.2 Business Process Re-Engineering

Business process re-engineering was introduced by *Michel Hammer* and *James Champy* and describes the *"fundamental rethinking and radical redesign of business processes to achieve dramatic improvements in critical, contemporary measures of performance, such as cost, quality, service, and speed [25]"*. As we will see later, the proposed implementation could induce improvements both on cost and quality depending on what level of integration is chosen.

6.3.3 B2B Branding - Innovation

The definition of branding, as proposed by *Philip Kotler* in his book *B2B Brand Management* consists of three dimensions; brand distinction, brand communication, brand evaluation and brand specialities. These are the perspective which a company should take into account when using brand to create value in B2B businesses i.e create *brand equity*. Here brand equity is defined as the,

- Perceived quality
- Name awareness
- Brand associations
- Brand loyalty

of the company's brand. Based on this definition we will argue that a change in the pricing process of MG's RF-1 product will have an impact on the perceived quality of the brand [26].

6.4 Results of Investigation

To perform this analysis we have performed a research of the current situation at Metric Gaming, with an overview of the betting industry as a whole, within each of the aforementioned perspective.

6.4.1 Operational Risk Management

Metric Gaming is a service provider, hence not taking any own financial risk on the bets they price. Therefore our research will start by showing how the gambling companies manage operational risk today, as they are those who might be interested in the risk management services included in Metric Gaming's product.

6.4.1.1 Trading, liability management and pricing We begin by presenting the result of our investigation concerning the risk associated to pricing process of bookmakers where we seek to answer the following questions.

- *What type of risk does a bookmaker expose itself to?*
- *How is risk currently managed in the betting industry?*

We have identified two main types of risk linked to the pricing process that bookmakers are exposed to. Both of them are financial risks; the *value on bet* and the *exposure* risks.

1. Value on bet risk: We here make use of an example to illustrate this type of risk. Let a bet on *Real Madrid* to win against *Barcelona* hold an odds of 5, this would translate (without taking into account the vig; see 1.3.1) into a probability

of $\frac{1}{5} = 0.2$ of Real Madrid winning. Assume now that this is not the case and that this event (Real to win) in fact holds a probability of 0.3 then there is a value of 50%. If a punter manages to identify such bets and betting on those a significantly large amount of times then, by the law of large numbers, he is ensured a 10% profit without taking any risk, similarly to an *arbitrage* situation.

2. Exposure risk: Let there be a market with two possible outcomes; e.g. team A to win (A) and team B to win (B), hold odds of 1.5 respectively 2.3 (when inverting those, the sum of the probabilities would give 1.10, i.e. a vig of 10%; see 1.4.2). The ideal situation for a bookmaker would be to have $\frac{2.3}{1.5} = 1.53$ times the amount of money on bet A than bet B so that whatever the outcome of the match the bookmaker earns the 10% vig. If this is not the case the bookmaker take a risk position where one of the outcomes would lead to losses for him.

Through research concerning the different types of bookmakers and our interviews, we have found out that, amongst gambling companies, there are three strategies employed in the industry.

1. *Risk Surveillance*
   Most large gambling companies have their own risk management system, tailored to their preferences and organisation. They do research to come up with starting prices, when the markets are live the algorithms in their risk systems adjust the prices. How this process is conducted exactly we do not know, but we believe that there are traders/risk managers monitoring multiple games at once ready to go into "manual mode" when needed [27]. From our own experience this system is not that dynamic, they do not alter prices just because they are getting some exposure on one side of a market. They usually impose higher taxes (vig) on their prices than the next category. Examples being; Bet365, Bwin etc.

2. *Risk Minimising*
   Here the company takes an approach to risk that is muck easier to define. By being extremely reactive and carefully monitoring incoming bets they manage to keep sharp lines at all times. This means that the money placed on outcomes and the corresponding prices match up so that the company never succumbs to exposure. This is analogous to a supply/demand situation where the company try to obtain and control equilibrium. By doing this they offer prices with very low imposed tax and their business model instead being predicated on large turnovers. Pinnacle Sports is world leading in this category.

3. *Outsourced Risk*
   Lastly, we have the open exchanges, aforementioned Betfair being one of those. They do not take any risk as they merely offer a place for punters to meet and place bets against each other. It works very similarly to a stock exchange. As this sector does not have any risk to lever, it is not relevant to our question at hand. Rather interesting though, and something we discussed in our methodology, is that

these can be considered to the "truest" odds out there [13]. They, if the market is moderately liquid, represent the market's view on the true worth of the outcomes, a situation which can be closely linked to economics theory. We believe that these prices are heavily used by the two former categories as an indicator of where the market is at.

From the interviews we have conducted with football traders at MG it is our understanding that the company employs a risk surveillance strategy where it is up to the traders to asses if no further exposure can be taken or if the situation is so uncertain that any odds would lead to risk. In very uncertain situations, where the odds do not reflect the real probabilities of the outcomes (leading to value on bet risk) it is common practice for the traders to freeze certain outcomes if not the whole market. Situations like this are, among others:

- Corners: Where the trader often freezes the Goal Kick and Corner outcomes since the likelihood of these happening become very uncertain and/or the trader cannot change the odds, for them to reflect the new probabilities, quickly enough.
- Dangerous Free Kicks: Same as for Corners
- Injuries: It is a rule of thumb to freeze the Throw In outcome when injury breaks happen since players often stops the game by sending out the ball to Throw In without it appearing on the live televised feed (which often transmits replays of the injury).
- Technical issues: In case of major technical errors, e.g. if the trader cannot access the live televised feed it is common to freeze all markets.

This freezing of market outcomes is mainly (not for technical errors or injuries) because of the limitations of the human trading which can impossibly react (reassess the probabilities and hence change the odds) instantly to the aforementioned events. For the risk of exposure, the trader tend to lower the odds of popular outcomes and raise the other ones so to attract clients on them and try to get balanced markets.

6.4.1.2 The High Fixed Cost Furthermore, the implementation we aim to investigate is expected to induce major changes in the high fixed cost base of the company which is manly composed of employee wages, research and development and rent costs. To investigate the impact it might have we start off by answering the following question;

- *What is the organisational structure surrounding this area of production?*
- *The company's future operations strategy, what is it and what challenges lie ahead?*

The production part takes place as follows. For every game that the company covers there is a *trader* and a *scorer* working on it. It is the trader's responsibility to monitor

the market and bets that are coming in, set the prices and also suspending[21] markets when needed. The scorer is the one who meticulously monitors the game, putting in the times for-/types of- outcomes, aided by external services to ensure accuracy in form of a precise *live feed*[22]. Such a feed is never a 100% but more in the regions of 95% accuracy. That the scoring is done live ensures that the company can settle and pay out the bets in real time. All the trader/scorer teams is backed by a central support unit which can aid them in various issues they cannot manage in real time. This can be controlling unclear outcomes or reversing scored bets when errors in scoring are made.

When interviewing about the future of operations we find that re-engineering of the pricing process through automatisation is an integral part in bringing the company forward. However it is hard to know to what extent as there is a lot of issues to be addressed. Looking at the scoring part of the supply chain, it could be outsourced to other specialised scoring services. However, then they need to look at whether a drop in accuracy, $\sim 4\%$, and the price for the service is proportional to the decreased wage costs. Looking at the pricing process automatisation, is also the direction to work towards where some kind of semi automatisation is on the horison.

Indeed, MG is a relatively new company who has been in an expansionary phase where a stagnation in revenue and client base could induce an adverse impact on its profitability if no cost reductions are engaged.

6.4.2 Branding - RF-1 an innovative product

We will in this section investigate how the betting industry react to innovative products and how MG's Super Live platform, and especially the RF-1, is met by the gambling companies. We believe that implementation of our model could have an impact on the general opinion regarding this innovative product and improve MG's brand equity.

- *How innovative is the betting industry as a whole?*
- *How is Metric Gaming's innovative product met by the betting industry?*

Innovation and development is an area important to every industry. The gambling sector is not an exception but it is a industry which is stagnant in multiple aspects as technology, product development. There has been a false sense of security in the market, with an *"If it ain't broke, don't fix it"*-attitude [28]. A panel with executives from the biggest actors in the industry was conducted at a gambling conference in the end of 2013 on the subject of innovation.

"I agree we are waiting for the next big invention, but it will come definitely out of the mobile space. It will come from the mobile way of thinking." - George Zenzefilis, CEO, Intralot

[21] When dangerous situations occur, e.g. a penalty or dangerous free kick
[22] Much like the one used for collecting the data we use in this paper

MG are at the frontier of innovation in the industry, capitalising on technological advances such as smart phones. The SuperLive™-platform is an innovation which has potential to change the way of thinking in the market, and it is released in an interesting transition within the sector, going from conservative to progressive. It is great to be among the market leaders in innovation but it can also be an obstacle. Since MG's potential customers are gambling companies, and, as aforementioned, the sector is innovatively stagnant, being market leaders in this aspect induces difficulties in terms of integration, from both a technological perspective and that it is hard to persuade conservative actors that this is the future. It is understandable that they are reluctant to invest in a product that presents uncertainty and has not been established on the market, this especially when thinking that they could always integrate this product after market breakthrough without expecting great client loss. Hence, MG is very capable to reap the benefits of their innovative position but may have to wait on the industry to catch up in order to truly unlocking their potential.

6.5 Analysis - Implementation

In this section we will propose three different ways for implementation of our results and see what implications they might have using our perspectives of analysis. As mentioned in the introduction we here operate under the assumption that we have been able to create a predictive model with an accuracy which is satisfactory.

6.5.1 Three Levels of Implementation

In this section we present the tree possible levels of implementation of our model within the pricing process of the RF-1 market and the analyse the .

6.5.1.1 Full Automatisation A full automatisation of the pricing process would, in light of our mathematical results and with proposed development of the model, be fully possible in terms of available data, technology etc. All of the input needed for the initialisation of probabilities calculation (state of the process at market opening, transition propensity etc.) can be derived from the support unit described in 5.1.2. In this level of implementation the trader and scorer would be completely replaced by the software.

6.5.1.2 Semi Automatisation This level of implementation allows for a trader to work on several games at a time where his role would shift to a controller. Here the software would run the games and the controller would supervise the process, intervening if he feels that the game presents abnormal characteristics which would lead to miscalculations of the probabilities by the software.

6.5.1.3 Support Tool Finally, the software could be used as a simple support tool assisting the trader with the probabilities/odds predicted by our model. Here the pricing

will remain under the trader's responsibility and it will be up to him/her to decide whether to base his pricing on the mathematically derived odds, hedging strategies or simply on experience-based instinct.

6.5.2 Risk Management

6.5.2.1 Liability Management and Quality Improvement When implementing our model as a software for a fully automatised pricing process it is of capital importance to address the risk linked with presenting probability based odds. They are a good way for the company to eliminate the risk of bets presenting value, i.e. if the odds do not represent the probability of the event happening one could bet on such outcomes (with higher odds that probability based ones) a large number of times and by the law of large numbers ensure a positive return[23] on investment (one could see this as an arbitrage situation). For example our model would be very effective in eliminating the need for freezing outcomes of the market since it will instantly take into account the current state of the process and re-evaluate the odds to take into account drastic changes in probabilities (see 5.1.1). This would lead to a smoother way of betting (betting is possible at all time) and with a odds fluctuation that could, in our opinion, enhance the betting experience (be able to at specific time take the opportunity of betting on temporarily higher odds).

While this is important for the clients, i.e. the bookmakers, they often focus their risk management so to never be expose to risk of great losses. A way this to be taken into account in a fully automatised pricing process the software would require an extension so to change the prices in function of the betting propensity of each outcome. This type of odds tweaking should require a deep mathematical and economical analysis to be integrated within the software. Indeed, due to the very short time until settlement of the bets, one would need to take into account a lot of factors for the algorithm to asses how much the odds should be change in order to rapidly reduce exposure on one or several outcomes. It would require the construction of a supply and demand model to analyse at which level of added positive/negative value an outcome receives increased/decreased bet propensity. This model would then be used for quantifying the required odds change to hedge a specific amount of exposure for a specific outcome. We expect this type of modelling to be extremely complex and both time- and money consuming for MG, especially for a betting market including four different outcomes.

In the semi automatised model the freezing of the markets would still be eliminated although the trader would be able to freeze markets manually if necessary. Thus the risk management would fall under the trader's responsibility where he could freeze the markets in order to ensure that too much exposure is not attained. We expect the trade off, between risk and betting experience induced by this type of solution would be positive since the frustration of not being able to place a bet would be more rare.

[23] Which should be impossible when taking the vig into account, see 1.4.2

Finally, when using the software as a simple support tool the situation would be the same as today, i.e. risk management resting solely upon freezing of the markets both for the odds-value risk and the exposure risk.

6.5.2.2 Profitability Risk - Reducing High Fixed Cost While an implementation of our model as a support tool would not lead to any significant cost reductions in the pricing process, it would have a major impact if some kind of automatisation is chosen. This level of integration would only require the cost of setting up the model since risk management issues would remain under the responsibility of the trader. There is also the possibility, regardless of implementation strategy, to integrate the aforementioned live feeds (see 5.1.2) and thus eliminate the scorer completely. One must then analyse the trade-off in data accuracy, how much extra work it will be to re-settle bets in retrospect and what impact such processes might have on the brand (and the view on the RF-1) versus the decrease in wage costs. Let us henceforth operate under the assumption that it is worth to perform such an integration.

In the scenario of a fully automatised production, drastic changes would be possible; the most significant one being the cost reduction induced by cut backs in terms of manpower. Looking over the production process and assuming that trader hold slightly higher wages than scorer, we estimate that the trader and scorer wages stand for $\sim 40\%$ respectively $\sim 25\%$ of the production costs. Both of them which would be replaced by the software. Although this seems very attractive, concerns arise concerning the set up cost of such a system and quality of the production. The aforementioned need for a sophisticated risk management plug-in is expected to require significant investment.

For the semi automatised implementation much of the problems that arise in the previous scenario could be avoided. Here we expect a possibility of eliminating the scoring part of the production ($\sim 25\%$ cost) and reducing the number of traders by 75%, if not more, by allowing traders to supervise several (say four) games at the same time. The fact that they would be in charge of making sure that the software does not produce odds that are unreasonable and that exposure levels are under a certain level, eliminates the need for constructing the aforementioned risk management plug-in. It would be a smaller investment to develop the model than in the preceding case. Finally, such a shift in the trader's role would require educational seminars etc., in order to introduce them to the new tasks they will have to accomplish.

6.5.3 Branding

As in the preceding section, a support tool would not change much in terms of brand equity. With an semi-/full automatisation there is a number of brand equity consequences to investigate. An implementation would make an innovative product more solid. In our introduction we featured an excerpt from an interview with the company's founder and director. He stated that this product would increase betting propensity to perhaps

hundred of times per match. That means that probabilistically correct odds would, by the law of large numbers, minimise volatility and stabilise return for MG's clients. Therefore, an automated way of pricing could be a great way for the company to market themselves, of course depending on the model being solid enough. It could help the product, which today can be seen as a *niche* innovation, taking the trajectory towards being an established product. This does however require an appreciation from the clients, that the pricing our model entails is a strategy they are willing to bet on. As discussed in the section regarding risk management 5.1.1, most of MG's prospective clients would place in the first category. It is therefore hard for us to assess whether the market group as a whole could implement such a pricing model given that they all have unique ways of pricing and managing risk. However, these processes are not interacting with the process of the RF-1, it is therefore hard to see that they would object to our way of pricing on an integration level. The objection would perhaps be based upon a strategical difference of opinion regarding pricing, a discrepancy which MG can mitigate trough good marketing.

7 Conclusions

7.1 Our Modeling

Through this paper we have been able to derive an explanation for how some of the transition probabilities, used for our modelling of football as a Markov process, are affected by the aforementioned factors; time, score difference and difference in team quality. We have seen that modelling it as a Markov Chain is indeed possible and its use could be incorporated in a automated software for application in the betting industry. This said, the building of regression models explaining the probabilities variance to a greater degree than we have, would require a more complete in-depth analysis of football games. Our choice of covariates was based on intuition and although it seems like it was warranted we believe that a many more factors would be needed for better predictions. We have proposed a few things which we believe can improve upon the model; but all of which is beyond the scope and time frame of this project. Analysing our project we can conclude that building a model which can explain the main body of variance seems to be an extremely complex task where much more resources and experience are needed. We do however hope that our results may be something to build upon and we would encourage anyone interested to further our research.

7.2 Choosing The Level of Implementation

In light of the results of our research concerning the operations management consequences of implementing our model (in form of a software for pricing of the RF-1 market) on the mentioned levels, we believe that the semi-automatisation (see 5.2.1.2) is the most suitable model of implementation. It presents high cost reduction possibilities and presents the most effective risk management solution. Furthermore we believe that an integration of mathematically derived odds will be a good argument for convincing the rather conservative betting industry of the potential of innovative markets like the one proposed by MG. One need to know though, that at end of the line, the end customers are the ones who will have the final verdict regarding whether mathematically derived odds is something appealing. That will of course, by extension, affect MG's pricing strategy. Reflecting on our aim we can conclude that integrating a model like ours into existing production at a company like Metric is more complex than we first realised, particularly if one is to automate production completely. We do however hope that we have shed light on some things which are of value to the company and that it will aid them in forthcoming work.

References

[1] FIFA. *The FIFA Big Count*. 2007.

[2] Howard Hamilton. Failure — the motivator for expanded use of soccer analytics. 2010.

[3] Bruce Bukiet, Elliotte Rusty Harold, and José Luis Palacios. A markov chain approach to baseball. *Operations Research*, 45(1):14–23, 1997.

[4] Ben Zauzmer. Modeling nfl overtime as a markov chain. 2014.

[5] Metric Gaming. Global sports betting market. 2015.

[6] BetUS Team. What is the "vig" when betting? 2015.

[7] Jan Grandell and Jan Enger. *Markovprocesser och Köteori*. 2014.

[8] Harald Lang. *Elements of Regression Analysis*. 2014.

[9] Ph.D Paul Allison. *What's the Best R-Squared for Logistic Regression?* Statistical Horizons, 2013.

[10] R. William. *Standard errors for regression coefficients; Multicollinearity*. University of Notre Dame.

[11] John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.

[12] R.G. Pierse. *Lecture 5: Omitted Variables, Dummy Variables and Multicollinearity*. Win Solve.

[13] Michael A Smith, David Paton, and Leighton Vaughan Williams. Market efficiency in person-to-person betting. *Economica*, 73(292):673–689, 2006.

[14] Betfair trading where has all the money gone. 2011.

[15] Steve Howe. Trading on betfair: Making a profit on the premier league winner market. 2014.

[16] Torsten Hothorn and Brian S Everitt. *A handbook of statistical analyses using R*. CRC Press, 2014.

[17] Karen Grace-Martin. *Small R-Squared*. The Analysis factor, 2012.

[18] EA Sports Ultimate Team. Fut database. Player num:178509, Olivier Giroud, 2015.

[19] Soccer Performance. Factors affecting soccer performance.

[20] Ted Menmuir. *Interview with Martin De Knijff by SBC News*. SBC news, 2014.

[21] Bent Flyvbjerg. Five misunderstandings about case-study research. *Qualitative inquiry*, 12(2):219–245, 2006.

[22] Anna S Chernobai, Svetlozar T Rachev, and Frank J Fabozzi. *Operational risk: a guide to Basel II capital requirements, models, and analysis*, volume 180. John Wiley & Sons, 2008.

[23] Artik. *The Big Three; William Hill, Ladbrokes and Coral*. Reliable Bookies.

[24] Ladbrokes. *Business Review, Risk*. 2012.

[25] Michael Hammer and James Champy. Reengineering the corporation. 1993. *Haper-Collins, New York*, 1993.

[26] Philip Kotler and Waldemar Pfoertsch. *B2B brand management*. Springer Science & Business Media, 2006.

[27] Wayne Dews. *bookmaker-risk-management-how-it-has-changed*. BetAsia, 2012.

[28] William Chambers. *World Executive Gaming Summit*. Gamasutra, 2012.

TRITA -MAT-K 2015:11
ISRN -KTH/MAT/K--15/11--SE