PARIS SPORTIFS AU FOOTBALL : L'INTERET DES EXPECTED GOALS

Paul Steffen, Léo Gerville-Réache, Nicolas Bisoffi

PARIS SPORTIFS AU FOOTBALL : L'INTERET DES EXPECTED GOALS

Paul Steffen[1,2], Léo Gerville-Réache[2], Nicolas Bisoffi[1]

[1] *Betclic 51 Quai Lawton, 33300 Bordeaux, p.steffen@betclicgroup.com*
n.bisoffi@betclicgroup.com
[2] *Univ. Bordeaux, CNRS, IMS, UMR 5218, 33405 Talence, leo.gerville-reache@u-bordeaux .fr*

Résumé. Pour l'établissement des cotes du résultat d'un match de football, l'estimation des probabilités des issues possibles 1/N/2 est fondamentale. Les nombreux modèles statistiques qui sont utilisés se servent du résultat réel. Dans cette communication nous proposons d'utiliser les « issues espérées » construites à partir des buts espérés (expected goals, xG). En utilisant le modèle polytomique ordonné, nous comparons, sur plusieurs années de Ligue 1, les résultats obtenus selon l'utilisation des issues réelles ou des issues espérées.

Mots-clés. Football, buts espérés, issues espérées, modèle polytomique ordonné.

Abstract: To determine soccer results odds, estimate outcomes probabilities 1/N/2 is fundamental. Many statistical models used utilize the real result. In this paper, we propose to use "expected outcomes" constructed from the expected goals (xG). With the ordered polytomics model, we compare, on several years of Ligue 1, the results obtained using real or expected outcomes.

Keywords. soccer, expected goals, expected outcomes, ordered polytomic model.

1 Introduction

L'établissement de cotes pour un ensemble de paris sportifs passe, entre autres, par l'établissement de probabilités d'un ensemble d'événements caractéristiques. Pour l'issue d'un match de football, le score à la mi-temps est un événement. Le score final est aussi un événement (dépendant du score à la mi-temps). Sur les sites de paris sportifs, on peut également parier sur les buteurs, l'équipe qui ouvrira le score, etc... Pour établir une estimation fiable de la probabilité d'un événement, il faut construire un modèle probabiliste et statistique aussi pertinent que possible. Dans le cadre d'un projet Région Nouvelle-Aquitaine / Betclic / Université de Bordeaux, une thèse CIFRE sur l'établissement des cotes est en cours. Parmi les pistes de recherche, nous nous sommes questionné sur la pertinence d'une statistique relativement récente : l'expected goal (voir par exemple Rathke A. (2017)). Si l'on s'intéresse à la rencontre Bordeaux-PSG de décembre 2018, premier match nul pour le PSG depuis le début de la saison, le score final fut 2 à 2 alors que les expected goals étaient de 0,65 pour Bordeaux et 1,54 pour le PSG. Sous la réserve de la pertinence de ces expected goals, le déroulement de match aurait dû plus probablement produire la victoire au PSG. Mais il est bien connu que le résultat d'un match de football est très incertain. L'arrivée d'un but fait davantage penser au résultat d'un hasard sauvage qu'à celui d'un hasard bénin. Pour autant des régularités existent et des prévisions sont possibles.

Dans cette communication, après avoir présenté succinctement le Modèle Polytomique Ordonnée (MPO) utilisé dans cette recherche, nous présentons la méthode de construction des expected goals, la base de données de ligue 1 retenue, les principaux résultats obtenus et enfin les pistes d'évolution de cette recherche.

2 Le modèle polytomique ordonné

Le MPO est une généralisation de la régression logistique pour laquelle la variable à expliquer est multinomiale ordonnée. Pour notre étude, l'issue du match 1/N/2 est cette variable. En vue d'une modélisation, la précision suivante est nécessaire : 1 < N < 2. Le MPO exprime les probabilités des issues (Y) conditionnellement à un ensemble de covariables ($\boldsymbol{X}$) de dimension p :

$$P(Y \leq c|\boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(\alpha_c - \beta_1 x_1 - \cdots - \beta_p x_p)}{1 + \exp(\alpha_c - \beta_1 x_1 - \cdots - \beta_p x_p)}.$$

A travers une telle modélisation, seule la constante diffère suivant les différents niveaux de Y. L'utilisation de ce modèle a déjà fait l'objet d'une communication en 2014 sur les paris à handicap au rugby, voir Champagne L., Gerville-Réache L., Tiarks S. (2014).

3 Les expected goals

L'expected goal est une estimation du nombre de buts qu'une équipe aurait dû marquer lors d'un match. Il peut être basé sur les tirs tentés lors du match, ou sur les actions n'aboutissant pas à une frappe : les passes, les interceptions, les dribbles ou les tacles par exemple. Pour notre étude, nous nous intéresserons uniquement aux expected goals basés sur des tirs.

A chaque tir, est associée une probabilité de but, basée sur la distance, l'angle de tir et la partie du corps avec laquelle le tir a été effectué. Ces probabilités d'événement sont ensuite sommées afin d'obtenir les courbes d'expected goal de chaque équipe au cours du match. Cette mesure permet d'évaluer les performances des joueurs et des équipes. Dans un sport comme le football, où l'occurrence de buts est faible, le score final ne fournit pas toujours une image fidèle du match et de la performance des équipes. En mesurant statistiquement la qualité des opportunités de buts créées et concédées, il semble possible d'avoir une analyse plus fine de la rencontre.

Pour cette étude, les expected goal associés à chaque match sont ceux proposés par le site understat.com, qui utilise comme base de données, plus de 10 variables sur plus de 100 000 tirs afin d'estimer les expected goal via un réseau de neurones.

Ci-dessous, la représentation des tirs et le graphe de l'évolution de l'expected goal de chaque équipe, lors du match de la 15ème journée de Ligue 1, Bordeaux-PSG :

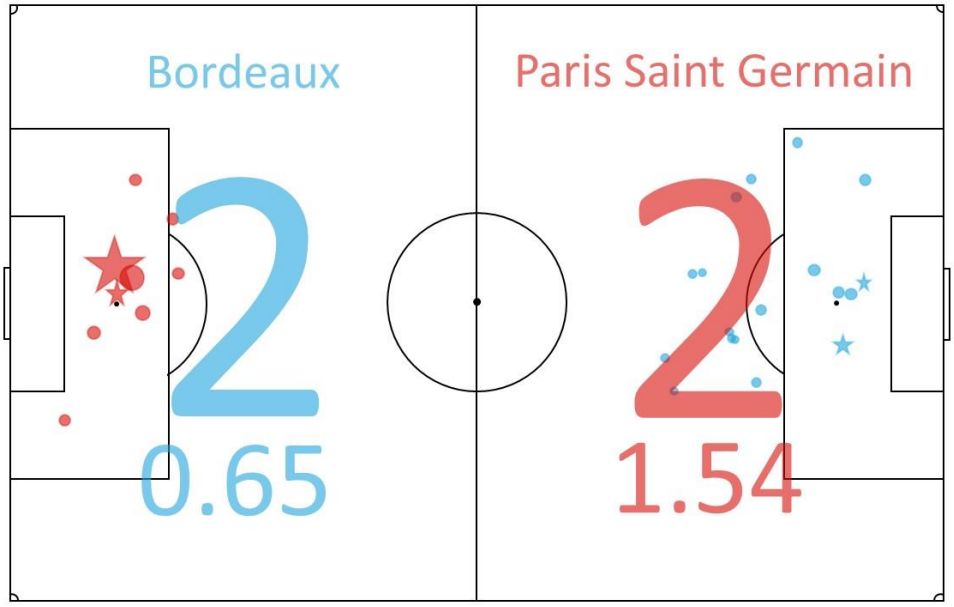

Fig1. Position des tirs (la surface des cercles traduisent les probabilités de buts et les étoiles, les buts)

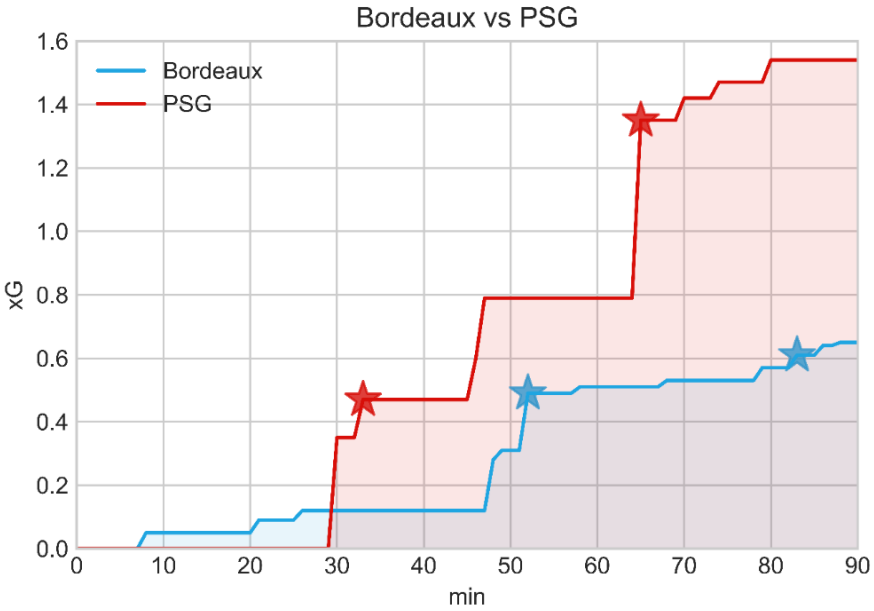

Fig 2. Expected goal des deux équipes en fonction du temps

Bien que le PSG ait moins tenté sa chance lors de la rencontre, les tirs du club de la capitale étaient statistiquement plus dangereux. Ainsi, avec seulement 9 tirs, dont 2 cadrés, le PSG a obtenu un expected goal de 1.54. A contrario, Bordeaux, ayant réalisé 18 frappes, dont 4 cadrées, a obtenu un expected goal de 0.65 à la fin de la rencontre.
On peut noter que, malgré la différence des comportements offensifs et la dangerosité des frappes des deux équipes, le match nul au terme de cette rencontre ne permet pas d'aboutir à une analyse aussi fine.

4 Prévision du 1/N/2 : l'exemple de la ligue 1

Afin d'estimer les différents coefficients du modèle, il est naturel et classique d'utiliser un ensemble de rencontres pour lesquelles la variable à expliquer est le résultat réel de la rencontre (1/N/2). Mais connaissant les expected goals à la fin de chaque match, on peut imaginer utiliser cette information pour définir un « expected outcome » de chaque match.

La base de données

A l'aide des expected goals d'understat.com, et des valeurs Elo de ClubElo un dataset des matchs de Ligue 1 depuis le début de la saison 2014/2015 jusqu'aujourd'hui a été construit. La valeur Elo, qui est un classement par point, mesure la qualité d'une équipe, d'après les résultats obtenus lors des précédentes rencontres et la qualité des opposants, en pondérant ces résultats par l'ancienneté des rencontres.

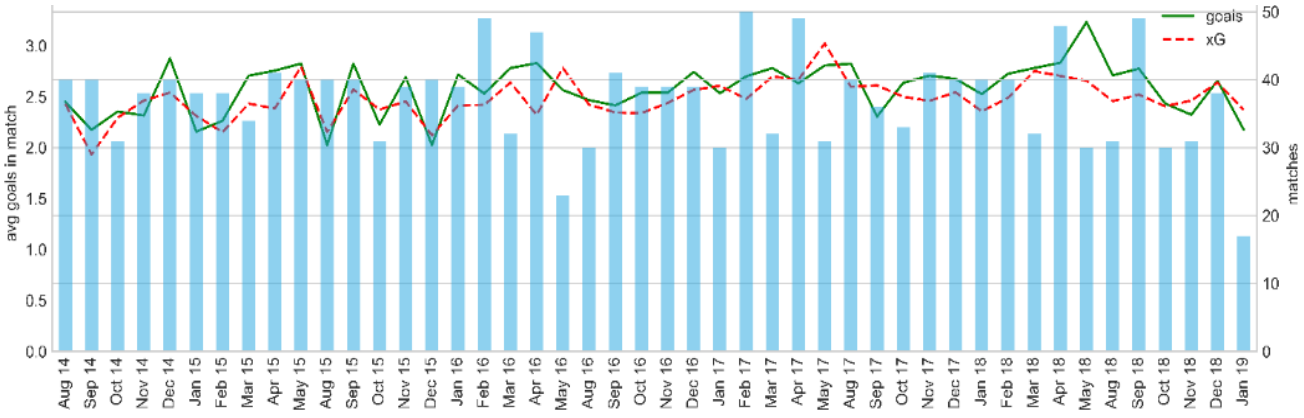

Fig 3. Moyenne des expected goal, du nombre de buts et nombre de matchs (par mois)

Lors des 1 716 rencontres ayant eu lieu en Ligue 1, depuis août 2014, on compte en moyenne 1,47 buts pour l'équipe à domicile et 1,11 buts pour l'équipe à l'extérieur. D'après l'expected goal, les équipes à domicile devraient marquer en moyenne 1,41 buts par match contre 1,06 pour les équipes à l'extérieur. Depuis cette même date, on compte 45% de victoires à domicile, 26% de matchs nuls et 29% de victoires à l'extérieur. En arrondissant à l'entier l'expected goal, on compte 43% de victoires à domicile espérées, 34% de matchs nuls espérés et 23% de victoires à l'extérieur espérées.

Malgré une différence de proportions non négligeable concernant les matchs nuls et les défaites constatées et celles espérées, le nombre de buts constatés et ceux espérés sont relativement proches à la vue des moyennes par match.

Afin de visualiser les positions des différentes équipes relativement aux buts marqués à domicile d'une part et à l'extérieur d'autre part, voici (Fig.4) les moyennes de buts marqués par les différentes équipes de Ligue 1 sur l'ensemble de la base de données. Sans surprise, on retrouve le PSG en haut à gauche du graphique. A l'opposé, on trouve Lens, Troyes et Nancy. Enfin, on peut remarquer que certaines équipes, comme Nîmes, ont une moyenne de buts marqués à l'extérieur supérieure à celle des buts marqués à domicile.

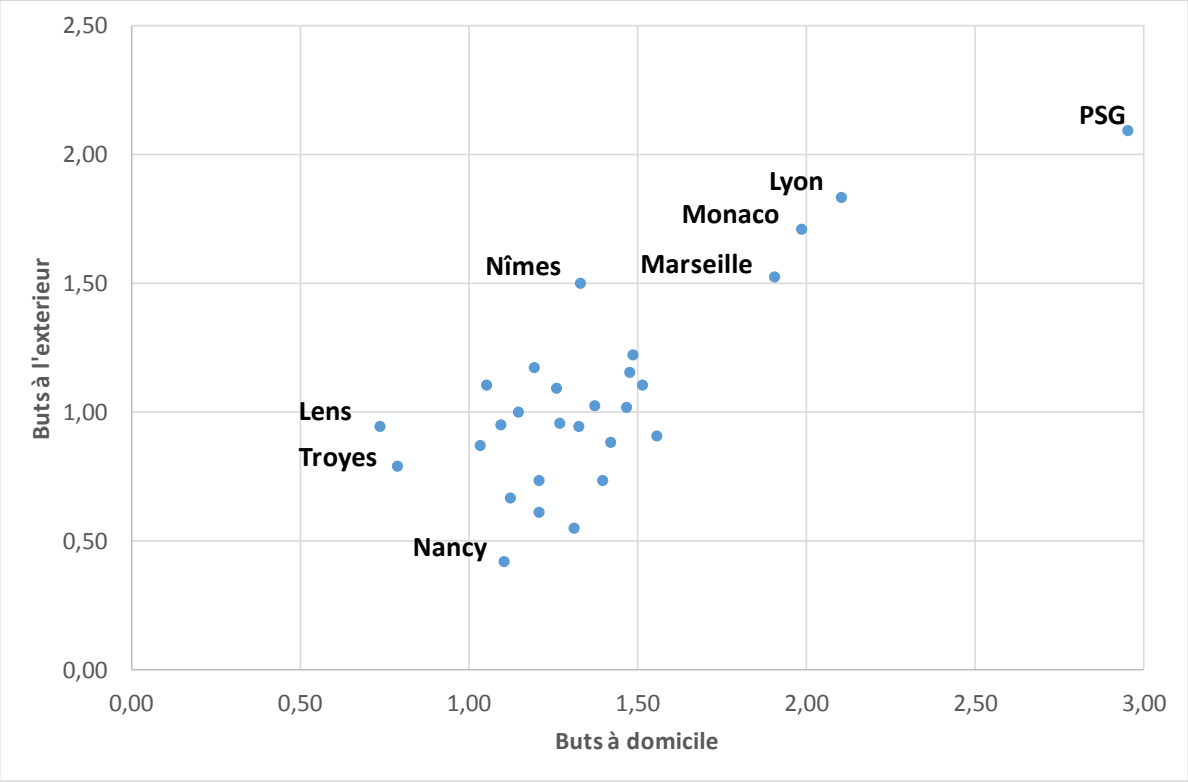

Fig 4. Moyennes des buts à domicile et à l'extérieur

Les issues espérées

Soit xG_1 l'expected goal de l'équipe à domicile et xG_2, l'expected goal de l'équipe à l'extérieur, c'est en fonction de la différence $xG_1 - xG_2$ que l'on définit l'espérance de l'issue du match *Y*. Le principe est alors le suivant :

$$\begin{cases} Y = 1 \ \ si \ \ xG_1 > xG_2 + \theta, \\ Y = 2 \ \ si \ \ xG_1 + \theta < xG_2, \\ Y = N \ \ sinon. \end{cases}$$

Où θ est une constante telle que, dans la base d'estimation, le pourcentage d'expected outcome nulles soit égale au pourcentage de matchs réellement nuls. Pour notre étude on trouve $\hat{\theta} = 0{,}36$. Précisément, si $|xG_1 - xG_2| \leq 0{,}36$, l'issue espérée sera *N*.

Le modèle concret (équipe *i* à domicile et *j* à l'extérieur) s'exprime de la manière suivante :

$$P(Y_{i,j} \leq c | \boldsymbol{X_{ij}} = \boldsymbol{x_{ij}}) = \frac{\exp(\alpha_c - \beta_i - \gamma_j - \delta \times DifElo_{ij})}{1 + \exp(\alpha_c - \beta_i - \gamma_j - \delta \times DifElo_{ij})}.$$

Avec $DifElo_{ij}$, l'écart de classement Elo entre l'équipe à domicile *i* et l'équipe à l'extérieur *j*. L'estimation par maximum de vraisemblance des paramètres a été réalisée avec le logiciel R.

Résultats

Sur les 1716 matchs présents la base, un échantillon tiré au sort de 1000 rencontres a servi de base d'estimation et les 716 autres ont servi de base de test. Afin de comparer la qualité des estimations des probabilités des différents résultats (1/N/2), la racine carrée de la distance du khi-deux par rapport aux résultats réels a été utilisée :

$$D = \sqrt{\sum_{i=1,N,2} \sum_{j=1,N,2} \left(n_{ij} - n_i p_{ij}\right)^2}.$$

Les résultats obtenus sur la base test pour les deux modélisations sont résumés dans le tableau ci-dessous :

		Probabilités estimées à partir des résultats réels			**Probabilités estimées à partir des expected results**		
Résultats réels	**Nombre de matchs**	**1**	**N**	**2**	**1**	**N**	**2**
1	311	**0,510**	0,247	0,243	**0,567**	0,249	0,184
N	192	0,448	**0,261**	0,290	0,500	**0,274**	0,227
2	213	0,347	0,270	**0,383**	0,354	0,286	**0,360**
Distance "khi-deux"		**302,4**			**293,1**		

Ici cette distance est de 3,1% inférieure pour la modélisation basée sur les issues espérées. Ce résultat est encourageant quant à son utilisation et donc l'utilisation des expected goals pour modéliser les probabilités (et les cotes associées) des différentes issues d'un match. Sur la droite, l'ACP montre les corrélations entre les estimations selon les deux modélisations (R : réelle et xG : espérée) pour la base test. Ces corrélations sont de 0,77 entre RN et xGN, de 0,83 entre R2 et xG et enfin de 0,86 entre R1 et xG1. Sont également positionnées les trois issues réelles. On peut en conclure que les probabilités des deux modélisations diffèrent mais que les deux modélisations restent cohérentes. Afin, il est intéressant de noter que les probabilités du match nul sont corrélées positivement au résultat « 2 » (la probabilité de victoire à l'extérieur).

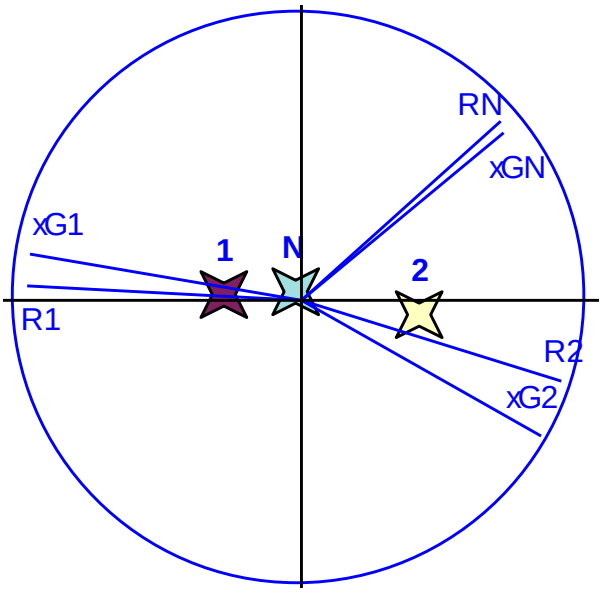

Fig 5. ACP sur les prévisions des deux modélisations

5 Discussion

Bien que les résultats précédents mettent en évidence un intérêt certain pour l'expected goal dans le but d'établir les probabilités d'issues de rencontre lors d'un match de football, cette mesure n'est pas sans faille. En effet, l'expected goal construit par understat.com, tel que nous l'utilisons à l'heure actuelle, ne semble pas prendre en considération toutes les variables que nous souhaitons utiliser pour construire cette mesure. Par exemple, un ajustement de la probabilité selon le joueur ayant tiré semble nécessaire : « un tir de Messi a plus de chances de finir au fond des filets, qu'une frappe de Mitroglou, *ceteris paribus* ». Dans un souci d'amélioration de cette variable et d'autonomisation du processus de prévision, nous devons réfléchir à la construction même de l'expected goal.

Dans cette étude, nous avons pris parti pour un modèle relativement simple, ne s'intéressant qu'à l'issue du match. Le comparatif plutôt encourageant de l'utilisation de l'expected goal, via l'expected outcome, par rapport à l'issue réelle d'un match, doit donc être relativisé. Premièrement, notre objectif est l'obtention de probabilités d'un bien plus grand nombre d'évènements : score final, score à la mi-temps, buteur, équipe qui ouvre le score etc… Deuxièmement, l'écart de qualité entre les deux modèles (issue espérée vs issue réelle) est faible et ne permet donc pas à l'heure actuelle de garantir sa réelle pertinence. Ainsi, se sont bien les courbes d'expected goal, qu'il faut introduire dans la réflexion.

D'une manière plus générale, les deux modèles construits ne sont pas de qualité suffisante. C'est au niveau des variables explicatives qu'il faut également réfléchir. Deux pistes semblent d'ores et déjà incontournables. La première a déjà fait l'objet d'une étude sur le rugby (voir Champagne L., Gerville-Réache L., Tiarks S. (2014)) ; il s'agit de l'intégration d'avis d'experts. La deuxième est la composition et le schéma tactique de l'équipe donnés par la feuille de match.

Bibliographie

[1] Champagne L., Gerville-Réache L., Tiarks S. (2014). Construction de cote : l'exemple du pari à handicap au Top 14, Journées de statistique, Rennes, 6p.

[2] Champagne L., Gerville-Réache L. (2015) *Autour des procédures de classement : exemples du tennis, du tennis de table et du golf*, Journal de la société française de statistique. Vol. 156(2), 5-24.

[3] Foulley JL. (2012). *Tentative d'évaluation et de classement des 16 équipes de l'Euro 2012*. w3.jouy.inra.fr/unites/miaj/public/.../applibugs.12_12_20.jlfoulley.pdf.

[4] Langville A. & Meyer CD. (2012). *Who's one ? : The Science of Rating and Ranking*. Princeton University Press.

[5] Massey K. (1997). *Statistical Models Applied to the Rating of Sports Teams.* Bluefield College, 1997 - 74.207.231.132.

[6] Rathke A. (2017) An examination of expected goals and shot efficiency in soccer, preceding 6th ISPAS. International Society of Performance Analysis of Sport, Ireland, pages 514-529.

[7] Borøy-Johnsen S. (2017) Beating the bookmakers using artificial neural networks to profit from football betting, these.

[8] Ensum J., Pollard R., Taylor S. (2004) Applications of logistic regression to shots at goal in association football: Calculation of shot probabilities, quantification of factors and player/team, *Journal of Sports Sciences*, 22 (6), 504

[9] Pollard R., Ensum J., Taylor S. (2004). "Estimating the probability of a shot resulting in a goal: The effects of distance, angle and space". *International Journal of Soccer and Science*. 2 (1): 50–55