UCLouvain

LOUVAIN
School of Management

Louvain School of Management

Le classement Elo et les "expected goals" comme indicateurs de performance lors de la Coupe du Monde 2018 de football.

Auteur : GERARD Thibault
Promoteur : SAERENS Marco
Année académique 2018-2019

UNIVERSITE CATHOLIQUE DE LOUVAIN
LOUVAIN SCHOOL OF MANAGEMENT

Le classement Elo et les "expected goals" comme indicateurs de performance lors de la Coupe du Monde 2018 de football.

Promoteurs : Marco Saerens

Mémoire présenté en vue de l'obtention du Master en ingénieur de gestion à finalité spécialisée par :
Thibault Gérard

Septembre 2019

remerciements

Je tiens tout d'abord à remercier toutes les personnes qui m'ont aidé lors de la rédaction de ce mémoire.

Je voudrais dans un premier temps remercier mon promoteur de mémoire Monsieur Saerens, professeur à l'UCLouvain, pour sa disponibilité ainsi que ses judicieux conseils, qui ont contribué à pousser et alimenter ma réflexion.

Je remercie particulièrement Monsieur Soetewey, Monsieur Degimbe et Monsieur Chen pour leur aide dans la compréhension du langage R et dans l'écriture de mon code.

Je souhaite également remercier Mme Denis pour son aide précieuse à la relecture et à la correction de mon mémoire.

Enfin, je remercie Lysa ainsi que ma famille et mes amis qui ont toujours été là pour moi. Leur soutien moral et leurs encouragements ont été d'une grande aide.

Résumé

Dans ce mémoire, nous allons nous pencher sur l'évaluation d'équipes de football et plus particulièrement celle des 32 nations ayant participé à la Coupe du Monde 2018. Afin d'évaluer leur niveau, nous nous sommes basés sur deux éléments de mesure, d'une part les buts marqués et encaissés par les équipes sous forme de résultats ou scores, et de l'autre les valeurs d'*expected goals* pour chaque équipe lors de ces rencontres. Les *expected goals* sont une mesure de qualité d'occasion de but, définie selon plusieurs paramètres tels que la distance du tir par rapport au but, la qualité de la passe décisive, etc.

Nous avons par la suite inclus les données de ces rencontres dans un classement Elo optimisant ses paramètres de façon à obtenir la meilleure capacité prédictive pour ces classements. Une meilleure capacité prédictive pour un classement signifie également une meilleure évaluation du niveau des équipes selon l'indicateur de performance.

Les classements les plus prédictifs pour chaque indicateur de niveau ont ensuite été comparé entre eux, la différence des deux classements déterminant les équipes ayant obtenu des résultats supérieurs, ou inférieurs, à leur production dans le jeu en termes d'occasions.
Ensuite, nous avons essayé d'établir des liens de corrélation entre ces surperformances et sous-performances avec des statistiques liées aux sélections présentes à la Coupe du Monde 2018, que ce soit sur les joueurs ou les sélectionneurs. Cependant les statistiques testées n'étaient pas assez fortement corrélées pour expliquer suffisamment les différences de performance.

Finalement, la capacité prédictive des *expected goals* a été testée par rapport à celle des buts réels en comparant les modèles les plus prédictifs des deux indicateurs sur les scores et les résultats futurs de la Coupe du Monde 2018. Les *expected goals* se sont montrés plus prédictifs que les buts réels à la fois sur les scores et les résultats. Cependant, cette différence de prédiction n'étant pas significative, nous ne pouvons pas affirmer une supériorité en terme de capacité prédicitve des *expected goals* sur les buts réels.

En parallèle, deux classements des joueurs ont été créés afin de déterminer le meilleur joueur lors de la compétition. Ces classements se basent sur les notes des joueurs provenant du site spécialisé *Whoscored.com* lors des rencontres ainsi que leur influence sur les gains ou pertes de points de leur équipe dans les classements Elo.

table des matières

Liste des tableaux i
Table des figures ii

I Introduction 1

II Partie théorique 4
1 Méthodes de classement 4
1.1 Système d'évaluation Elo . . . 4
1.1.1 Historique et état de l'art . . . 4
1.1.2 Description . . . 7
1.1.3 Propriétés . . . 13
1.2 Modèles concurrents . . . 16
1.2.1 Méthode des chaînes de Markov . . . 16
1.2.2 Méthode des moindres carrés de Massey . . . 17
1.2.3 Modèle de Bradley-Terry . . . 18
1.3 Choix du modèle . . . 20
2 Expected goals 22
2.1 Historique et état de l'art . . . 22
2.2 Modèle de Michael Caley . . . 25
3 Méthodologie 32
3.1 Elo basé sur les buts réels . . . 33
3.2 Elo basé sur les *expected goals* . . . 38
3.3 Analyse des différences de performance . . . 41
3.4 Capacité prédictive des *expected goals* . . . 42
3.5 Classement des joueurs . . . 43

III Partie empirique 45
1 Jeu de données 45
2 Application des modèles Elo 50
2.1 Evaluation des équipes selon leurs buts réels . . . 50
2.2 Evaluation des équipes selon leurs *expected goals* . . . 53
3 Analyse des performances 56
4 Prédiction des scores et résultats futurs 62

5 Classement des joueurs 66
6 Conclusion, limites et travail supplémentaire 70
6.1 Conclusion 70
6.2 Limites et travail supplémentaire 71

IV Annexes 73
A Equations des différents types de tir selon le modèle de Caley 73
B Optimisation des différents modèles Elo 74
C Code R 75

V Bibliographie 79

liste des tableaux

Table 1	Distribution de Poisson des "buts" lors de Suisse-Italie	9
Table 2	Variables incluses dans les différents modèles d'*expected goals*	24
Table 3	Distribution de Poisson des *expected goals* lors de Brésil-Belgique	39
Table 4	Valeurs de S selon la formule de différence de buts	40
Table 5	Equipes qualifiées à la Coupe du Monde 2018	46
Table 6	Classement de la phase de groupe	47
Table 7	Rencontres des huitièmes de finale	48
Table 8	Rencontres des quarts de finale	48
Table 9	Rencontres des demi-finales	49
Table 10	Finale et rencontre pour la 3$^{\text{ème}}$ place	49
Table 11	Valeurs de MSE des classements Elo optimisés basés sur les buts réels	50
Table 12	Classement des équipes suivant Elo 2	52
Table 13	Valeurs de MSE des classements Elo optimisés basés sur les *expected goals*	53
Table 14	Classement des équipes suivant Elo 6 et Elo 6 bis	55
Table 15	Performance des équipes	57
Table 16	Corrélation entre performance et différentes statistiques d'équipe	59
Table 17	Corrélation entre Elo 2, Elo 6 et différentes statistiques d'équipe	60
Table 18	Valeurs de MSE des différents classements Elo optimisés sur leur prédiction des scores réels	62
Table 19	Comparaison des prédictions de l'Elo 2 et l'Elo 6 sur les résultats de la phase finale	65
Table 20	Classement des joueurs selon leur note moyenne	66
Table 21	Classement des joueurs selon leur influence sur les classements Elo 2, Elo 6 de leur équipe	67
Table 22	Classement final des joueurs selon leur note moyenne et leur influence sur les classements Elo 2, Elo 6 de leur équipe	68
Table 24	Optimisation et valeur de MSE des classements Elo basés sur les buts réels et les *expected goals*	75
Table 25	Optimisation et valeur de MSE de différents classements Elo sur leur prédiction des scores réels	75

table des figures

FIGURE 1 Courbe sigmoïde 10
FIGURE 2 Valeurs de K dans le classement FIFA féminin. 12
FIGURE 3 Taux de conversion d'un tir selon son emplacement 26
FIGURE 4 Comparaison selon l'importance de l'occasion de but 30
FIGURE 5 Distribution des scores de football professionnel anglais de 1988 à 2014 33

Première partie
Introduction

Le sport a depuis de nombreuses années pris une place importante dans la société. Que ce soit en tant que vecteur social, moyen d'intégration ou pour mettre en lumière une région ou un pays. Cependant, le sport est régi par une donnée importante, le résultat. C'est la victoire qui rapproche les gens, comme on a pu le voir lors de la Coupe du Monde 2018 de football dans les pays dont la sélection a effectué un bon parcours (FE Online, 2018). C'est elle qui permet de mettre en avant des villes et pays comme Toronto et le Canada à la suite de leur victoire en Finales NBA 2019 (Cecco, 2019). C'est les bonnes performances du musulman Mo Salah et les victoires de son équipe qui ont aidé à réduire les discriminations vis-à-vis de certaines minorités, avec la baisse d'actes et de comportements islamophobes dans la ville de Liverpool depuis son arrivée (Alrababah et al., 2019).
Les légendes du sport que sont Merckx en cyclisme, Pelé en football ou Jordan en basketball sont avant tout reconnues pour leur capacité à gagner sans laisser la moindre miette à leur adversaire (Duseaux, 2017). C'est ainsi que ces noms sont restés si longtemps dans la mémoire collective, leur faculté à gagner que ce soit individuellement ou dans des équipes restées elles aussi dans la légende.

Néanmoins, la seule victoire est-elle un bon indicateur du niveau réel d'un sportif ou d'une équipe ? Les seuls résultats seraient-ils de bons indicateurs de la valeur d'un sportif ou d'une équipe à un moment donné ? Quelle part dans ces résultats tient de la valeur du sportif ou de l'équipe plutôt que de l'aléatoire du sport ?

Le football, sport sur lequel nous allons nous pencher durant ce mémoire, au travers de la Coupe du Monde 2018, a cette particularité d'être ce que les américains appellent un « low scoring game » (Smith, 2018), soit un sport dans lequel peu de buts sont marqués. Il est donc plus probable que la chance y ait une plus grande importance dans les résultats d'une équipe, que par exemple en basketball. Un tir s'écrasant sur le montant ou rentrant via celui-ci en football risque en effet de peser plus fortement sur le résultat final qu'un panier inscrit ou non en basketball. En 1968, Reep et Benjamin écrivaient déjà sur la chance et son importance dans le football. Ils en conclurent même que ce sport était dominé par la chance. Sur le terrain cette fois, certaines épopées comme celle de la Grèce au championnat d'Europe 2004 ou de l'équipe de Leicester, championne d'Angleterre en 2016, resteront à jamais des énigmes totales pour un grand nombre d'observateurs (Sprigings,

2016). La différence entre le jeu proposé et les résultats obtenus par ces équipes défiaient alors toute logique.

Cependant la majeure partie des systèmes d'évaluations d'équipes sont dictés par les seuls résultats, et leur influence va bien souvent au-delà d'un simple classement. L'économie du football, que ce soit les fédérations, les joueurs ou les entreprises de pari sportif, dépend en partie de ce type d'évaluation. Premièrement, les tirages au sort des compétitions internationales sont basés sur l'évaluation des équipes par le classement FIFA, lui-même basé sur l'historique des résultats des équipes nationales (FIFA.com, 2017). Ces tirages au sort ont, de plus, une influence directe sur les probabilités de victoire des équipes dans une compétition comme l'ont démontré Leitner et al. (2010) dans leur étude sur le Championnat d'Europe de football 2008. Ensuite, certains pays comme le Royaume-Uni utilisent ce classement comme paramètre décisionnel dans l'accord de permis de travail pour les footballeurs étrangers souhaitant évoluer dans un des championnats britanniques.
Il est également très important pour une industrie comme celle du pari sportif de pouvoir mettre une valeur chiffrée sur une équipe et ses performances. Les sommes en jeu sont importantes et une évaluation erronée pourrait donc leur coûter très cher.

Afin de réduire l'importance de la chance dans l'évaluation des équipes de football, un certain nombre d'analystes, sous l'impulsion de l'analyse de données dans les sports américains, appliquèrent un nouvel indicateur de performance aux rencontres de football, les *expected goals* (Stanton, 2017). Cet indicateur mesurant les opportunités de buts serait, par sa propension à réduire l'influence de la chance et du hasard dans la conversion d'une occasion en but, selon ses utilisateurs, un meilleur moyen de prédiction que les buts marqués. Cela signifierait donc que cet indicateur de performance pourrait déterminer la meilleure équipe sur une rencontre sans tenir compte des buts marqués par celle-ci, ni même de ses résultats.

Dans le but de déterminer la réelle plus-value des *expected goals* ainsi que leur capacité prédictive, nous allons devoir classer, à l'aide de la méthode la plus adaptée, les équipes selon leur faculté à obtenir de bons résultats d'une part et leur faculté à se créer des opportunités de buts de l'autre. La méthode de classement choisie pour ce mémoire et appliquée sur les matchs de la Coupe du Monde 2018 est le système d'évaluation Elo. L'Elo est ici choisi en tant que système de classement pour sa capacité prédictive supérieure, comme a pu le démontrer Lasek (2013). Il a comparé dans son étude diverses méthodes de classement parmi lesquelles différents classements Elo, les chaines de Markov, la méthode des moindres carrés de Massey et a constaté une capacité de prédiction supérieure pour les classements Elo. Cela

signifie donc que l'Elo possède une meilleure évaluation du niveau et de la valeur des équipes qu'il classe.

Le premier et unique croisement à ce jour entre les *expected goals* et le modèle Elo vient du blog de Worville (2016) où il comparait les performances des équipes de MLS (Ligue Américaine de football) à l'aide de deux modèles Elo différents. Le premier était basé sur les résultats (victoire, égalité, défaite) de chaque équipe et le second sur des *expected wins*. Ces *expected wins* sont identiquement calculées dans le modèle, cependant ces victoires ne sont pas basées sur l'équipe ayant marqué le plus de buts comme dans le premier modèle, mais celle ayant un avantage d'au moins 0,8 *expected goals* à l'issue du match.
Dans ce mémoire nous utiliserons donc également le classement Elo, afin de comparer les buts réels et les *expected goals*. Cependant, nous développerons un plus grand nombre de modèles se basant sur différentes variables (résultats, scores). Les *expected goals* seront également traités différemment, avec une distribution mathématique de ceux-ci afin de déterminer des probabilités de résultats ou de scores, plutôt qu'un choix arbitraire de différence d'*expected goals* résultant sur une victoire ou non, comme a pu le faire Worville.

Au cours de ce mémoire, nous nous intéresserons tout d'abord à cette première question : "Quels sont les modèles d'Elo se basant sur les buts réels ainsi que sur les *expected goals* les plus performants ?".
Ensuite, une fois les modèles d'Elo les plus performants connus, nous pourrons tenter de répondre aux deux questions suivantes : "Quelles sont les différences entre les évaluations des deux types de modèle (buts réels et *expected goals*) et comment peut-on les expliquer ?" et "Un modèle se basant sur les *expected goals* peut-il prédire efficacement les scores et résultats futurs d'une compétition ?".

Le terme d'*expected goals*, aussi abrégé par xG, ne sera pas traduit dans ce mémoire et restera donc dans sa langue d'origine, l'anglais. Les raisons sont, d'une part, une traduction française qui n'est que peu utilisée et pas encore popularisée, et de l'autre, cela permettra d'éviter une quelconque confusion avec le score ou résultat attendu compris dans la formule du système d'évaluation Elo.

Deuxième partie
Partie théorique

1 méthodes de classement

1.1 *Système d'évaluation Elo*

1.1.1 *Historique et état de l'art*

Le classement Elo est aujourd'hui une des méthodes de classement les plus répandues dans les compétitions sportives comme électroniques. Ce système d'évaluation est basé sur la comparaison, de joueurs ou de choix, par paires. David (1963) décrivait dans son livre ce principe de comparaison par paires, dans lequel chaque compétiteur se voyait attribuer une valeur de force ou de mérite permettant ensuite de les classer les uns par rapport aux autres.
Créé par le professeur Árpád Élő en 1959 dans le but de classer les joueurs d'échecs, le classement Elo est notamment utilisé par la Fédération Internationale d'Echecs (FIDE) depuis 1970. Elo a décrit son modèle en détail quelques années plus tard dans son livre « The rating of chessplayers » publié en 1978. Il y a développé les différentes variables de son modèle ainsi que la méthode mise à jour des notes attribuées à chaque joueur après une rencontre.

Dans un papier de 1995, Mark E. Glickman, professeur à l'Université de Boston, analysait les différents systèmes de classements appliqués aux échecs et plus particulièrement le modèle Elo. Il a détaillé les différents paramètres du modèle en pointant du doigt quelques défauts comme la gestion de joueurs inactifs ou celle des joueurs débutants. Il a donc décidé de créer son propre modèle, adapté spécifiquement aux échecs, le Glicko-system en 1995 ainsi qu'une version améliorée de celui-ci en 2012 avec le Glicko-2 system. Ces deux systèmes d'évaluations ont comme particularité de prendre en compte un paramètre "RD" (*rating deviation*) qui est l'écart-type du classement des joueurs, afin de répondre au mieux aux problèmes soulevés dans son analyse du modèle Elo.

Le modèle Elo a d'abord été utilisé dans le football pour sa capacité prédictive. En effet, Leitner et al. ont dans deux papiers différents datés de 2010, appliqué le système d'évaluation Elo afin de déterminer les vainqueurs de l'Euro 2008, ainsi

que de la Coupe du Monde 2010. Ils ont comparé notamment les probabilités de victoires selon le modèle Elo par rapport à celles de plusieurs sites de pari sportif, représentés dans leur étude par un consensus formé de différents *bookmakers*. Les différences entre les deux méthodes mettaient en avant la difficulté du calendrier des équipes prise en compte par les *bookmakers*. Ainsi une équipe tirée au sort dans un "groupe de la mort" voyait ses probabilités de victoires selon le consensus sévèrement réduites par rapport à celles du modèle Elo.

Par la suite, Hvattum et Arntzen (2010) décidèrent de comparer cette capacité prédictive du modèle Elo par rapport à des référents naïfs ainsi que d'autres modèles prédictifs tel que ceux développés par Goddard (2005). Les deux modèles de Goddard comparés ici utilisent une régression ordonnée probit afin de déterminer les résultats des matchs selon l'historique des buts pour l'un, des résultats pour l'autre. Tous les modèles étaient ensuite comparés sous le prisme du pari sportif, notamment les gains et pertes ainsi que le retour sur investissement. Ils en conclurent que la simple utilisation de la différence de valeur entre deux équipes dans le modèle Elo était un très bon indice de prédiction de résultats.

Lasek (2013) compara à son tour dans son étude la capacité prédictive de différents modèles Elo adaptés au football, notamment les modèles du site spécialisé *Eloratings.net*, celui de la FIFA pour classer les équipes féminines de football international et un modèle Elo vainqueur d'une compétition sur le site *Kaggle.com*. Il en ressortit que le classement FIFA ainsi que le classement *Eloratings.net* étaient les meilleurs performeurs par rapport à d'autres modèles tels que ceux de Markov, Massey et Colley.

Dans leur livre regroupant les principaux systèmes d'évaluation et de classements, Langville et Meyer (2012) synthétisèrent le système d'évaluation Elo, preuve de son importance dans la littérature scientifique. Ils passèrent donc en revue les différents paramètres et variables du modèle Elo de base, avec quelques analogies aux applications dans le football et le football américain. On peut également voir la description de l'avantage à domicile et son incorporation dans le modèle en tant que paramètre, augmentant la valeur de l'équipe jouant à domicile dans la prédiction du résultat. Une autre addition au modèle de base développée dans ce livre est le remplacement du résultat par la différence de buts ou de points, afin de récompenser plus conséquemment les larges victoires.

Le système Elo s'est également développé dans un grand nombre de sports en dehors du football et des échecs, bien que n'étant pas des classements officiels. Ces

systèmes de notations et classements sont toutefois considérés comme des références pour les suiveurs assidus. Le site spécialisé dans les sondages et classements autour entre autres de la politique et du sport, *Fivethirtyeight.com*, en a d'ailleurs fait son système d'évaluation de base. Ce site développa donc un système d'évaluation Elo pour le basketball, le football américain ainsi que la Formule 1. En 2015, *Fivethirtyeight.com* proposa une version améliorée de l'Elo pour classer les équipes NBA (National Basketball Association) avec son modèle « CARM-Elo » en référence au joueur Carmelo Anthony. Premièrement, ce modèle modifie les valeurs initiales d'équipes par des notes « CARM-Elo » basées sur les performances offensives et défensives des équipes. Le modèle avantage également les équipes jouant à domicile avec un avantage supplémentaire pour celles évoluant dans des villes situées en haute altitude. A l'inverse, il désavantage les équipes jouant en « back-to-back », c'est-à-dire le second match d'un enchainement de deux matchs en deux soirs, ainsi que celles ayant voyagé longtemps avant un match. Cependant ce modèle n'a jamais été appliqué à d'autres sports étant donné les spécificités des variables.

Dans leur papier publié en 2016, Sullivan et Cronin se sont penchés sur le paramétrage des différentes constantes et variables de l'Elo dans le football. En plus d'optimiser ses paramètres tels que le niveau de compétition ou celui relatif à l'avantage à domicile, ils inclurent également au modèle un avantage pour chaque équipe l'emportant par deux buts d'écarts ou plus. A cela ils ont ajouté un nouveau facteur représentant la forme des équipes, les récents résultats influençant donc positivement ou négativement les probabilités de victoire d'une équipe ayant enchainé les victoires ou les défaites. Ces modifications ont permis d'améliorer la capacité prédictive de leur modèle basé sur plusieurs saisons de première division anglaise.

Dans son étude de l'Elo appliqué à l'Euro de football féminin 2017, Chen (2018) s'est intéressé à l'optimisation des deux paramètres de ses modèles que sont le niveau de compétition et l'avantage accordé à l'équipe à domicile. Il s'est inspiré de la méthodologie utilisée par Brier (1950) dans son étude sur les prédictions météorologiques en utilisant la méthode des moindres erreurs au carré. A l'aide de cette méthode il est parvenu à trouver le modèle Elo avec la meilleur capacité prédictive, modèle basé sur les buts marqués par chaque équipe lors de chaque rencontre et dont les valeurs initiales des différentes équipes au début de la compétition étaient toutes identiques.

1.1.2 *Description*

Le modèle Elo est un système d'évaluation de joueurs ou d'équipes s'affrontant en matchs un contre un. Il se base sur la performance des joueurs/équipes qui est une variable aléatoire normalement distribuée X dont la "moyenne" μ équivalente à la force ou la valeur intrinsèque de chaque joueur/équipe ne peut changer que lentement avec le temps. Il faudra une certaine période de temps pour que les résultats d'un joueur/équipe influencent donc sa "moyenne" μ. Ces différents résultats influenceront selon leur rendement par rapport à cette "moyenne" positivement ou négativement celle-ci.

Le système Elo a en effet comme principe de base de récompenser plus fortement les victoires que l'on pourrait qualifier d'inattendues. Une victoire d'un joueur ou d'une équipe classée très haut dans le classement sur un joueur ou une équipe beaucoup plus faible ne rapportera que peu de points à celui-ci, et à l'inverse, en fera perdre très peu à l'opposant le plus faible. D'un autre côté, la victoire d'un « outsider » face à un adversaire considéré comme significativement plus fort lui rapportera plus de points et en fera donc perdre beaucoup au favori de la rencontre. Le modèle s'ajuste après chaque rencontre selon un ajustement linéaire simple et proportionnel à l'écart du joueur/d'une équipe par rapport à sa "moyenne". Pour un résultat S, son ancienne note $r(old)$ est mise à jour pour devenir sa nouvelle note $r(new)$ définie comme suit :

$$r(new) = r(old) + K(S - \mu) \tag{1}$$

K est un paramètre qui représente le niveau de compétition et S est le résultat de la rencontre.

Si l'on considère maintenant chaque compétiteur comme une équipe on peut déterminer les notations de deux équipes i et j après une rencontre lors de laquelle elles s'affronteraient. Leurs valeurs respectives $r_i(old)$ et $r_j(old)$ avant la rencontre seraient mises à jour et deviendraient $r_i(new)$ et $r_j(new)$ en utilisant des formules similaires à (1).

1.1.2.1 S - Résultat

La valeur de S représente le résultat de chaque match et peut se définir comme suit dans la formule (1) :

$$S = \begin{cases} 1 & \text{si victoire} \\ 0,5 & \text{si égalité} \\ 0 & \text{si défaite} \end{cases} \tag{2}$$

Elle peut également prendre une autre forme que celle du résultat tel qu'on le connait : victoire, match nul ou défaite. Dans des sports ou compétitions où il existe une marge de victoire on peut par exemple remplacer le résultat de la rencontre par le score de celle-ci. Cela permet ainsi d'avoir une différenciation entre une victoire sur un score serré d'une autre assez large. Le score de l'équipe i contre l'équipe j, S_{ij}, peut donc prendre la forme suivante dans la formule (1) :

$$S_{ij} = \frac{G_{ij} + 1}{G_{ij} + G_{ji} + 2} \tag{3}$$

Avec G_{ij} le nombre de buts marqués par l'équipe i et G_{ji} le nombre de buts encaissés par celle-ci. Les critères selon lesquels S_{ij} est toujours compris entre 0 et 1 et la somme de S_{ij} et de S_{ji} est égale à 1 sont aussi respectés.

Lorsque l'on souhaite tester une statistique liée au nombre de buts ou de points inscrits, n'étant pas un nombre entier mais un nombre décimal, on peut alors distribuer ce nombre décimal selon la loi de Poisson comme l'on fait Dyte et Clark (2000) dans leur étude sur la Coupe du Monde de football. Cela permet d'obtenir une probabilité pour chaque score comprenant uniquement des nombres entiers de buts ou de points. Prenons l'exemple d'un match de football entre la Suisse et l'Italie. La Suisse inscrit lors de ce match le nombre fictif de 2,1 "buts" et l'Italie a elle inscrit 0,8 "buts". La distribution de ces "buts" en nombres entiers est la suivante :

	Nombre de buts									
	0	1	2	3	4	5	6	7	8	9
Suisse	0,122	0,257	0,270	0,189	0,099	0,042	0,015	0,004	0,001	0,000
Italie	0,449	0,359	0,144	0,038	0,008	0,001	0,000	0,000	0,000	0,000

En croisant ces probabilités de nombres entiers de buts, on obtient une probabilité pour chaque score et ensuite pour chaque résultat.

Italie \ Suisse	0	1	2	3	4	5	6	7	8	9
0	0,055	0,116	0,121	0,085	0,045	0,019	0,007	0,002	0,001	0,000
1	0,044	0,092	0,097	0,068	0,036	0,015	0,005	0,002	0,000	0,000
2	0,018	0,037	0,039	0,027	0,014	0,006	0,002	0,001	0,000	0,000
3	0,005	0,010	0,010	0,007	0,004	0,002	0,001	0,000	0,000	0,000
4	0,001	0,002	0,002	0,001	0,001	0,000	0,000	0,000	0,000	0,000
5	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
6	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
7	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
8	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
9	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000

TABLE 1 – Distribution de Poisson des "buts" lors de Suisse-Italie

Les probabilités de score sont directement disponibles dans le tableau ci-dessus. Pour les résultats, il faut additionner les scores en faveur ou défaveur de chaque équipe. Ainsi, la probabilité d'une victoire italienne est la somme des probabilités du coin inférieur gauche, celle d'un match nul est la somme des probabilités de la diagonale lorsque le nombre de buts est identique pour les deux équipes, et la probabilité d'une victoire suisse est la somme des probabilités de scores du coin supérieur droit.
Cela nous donne les probabilités de résultats suivantes :

- Victoire de la Suisse = 67%
- Egalité = 20%
- Victoire de l'Italie = 13%

Si l'on applique cela à la formule (2) on obtient pour cet exemple un S pour la Suisse égal à 0,77 et pour l'Italie de 0,23. La somme des S étant aussi égale à 1.

1.1.2.2 μ - "Moyenne"

μ représente donc le rapport de force entre les valeurs de chaque joueur/équipe avant une rencontre. Il contient aussi la valeur prédictive du modèle Elo en établissant un favori lors de chaque match. Lors d'un match entre deux compétiteurs i et j, μ deviendrait quant à lui μ_{ij}, ou le pourcentage de points que i devrait gagner contre j. μ_{ij} suivant l'hypothèse d'une fonction logistique dont la différence de notation d_{ij} est égale à :

$$d_{ij} = r_i(old) - r_j(old) \tag{4}$$

La fonction logistique est définie comme $f(x) = 1/(1 + e^{-x})$. Bien que la majorité des formules Elo utilisées dans le sport comme les échecs et le football emploient un exposant de base 10 plutôt que e, leurs graphes suivent la même courbe :

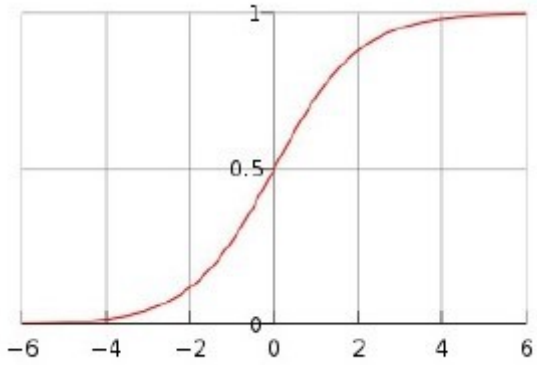


FIGURE 1 – Courbe sigmoïde

La valeur μ_{ij} pour l'Elo est donc définie aux échecs par la FIDE (Fédération internationale des échecs) ainsi qu'au football par la FIFA (Fédération internationale de football association) comme suit :

$$\mu_{ij} = L(d_{ij}/400) = \frac{1}{1 + 10^{-d_{ij}/400}} \tag{5}$$

Les formules mettant à jour les notes des joueurs et équipes sont donc telles que :

$$r_i(new) = r_i(old) + K(S_{ij} - \mu_{ij}) \tag{6}$$

$$r_j(new) = r_j(old) + K(S_{ji} - \mu_{ji}) \tag{7}$$

Avec S_{ij} tel que (2) et μ_{ij} tel que (5). La valeur de ξ est de 400 dans l'équation (5) telle que définie par la FIDE et la FIFA mais peut également prendre d'autres valeurs. Pour toute valeur de $\xi > 0$ on peut en déduire ceci :

$$\mu_{ij} = L(d_{ij}/\xi) = \frac{1}{1 + 10^{-d_{ij}/\xi}} = \frac{10^{r_i/\xi}}{10^{r_i/\xi} + 10^{r_j/\xi}} \tag{8}$$

Cela signifie donc que pour chaque avantage de valeur de note du compétiteur i sur le compétiteur j égal à ξ, celle-ci aura une probabilité dix fois plus grande de battre son adversaire par rapport à la probabilité d'être battu selon le score attendu. Si ξ est fixée à 400, alors le compétiteur ayant un avantage de 400 sur son opposant

aura une probabilité dix fois plus grande de gagner le match que de le perdre.
En effet, si l'on remplace la différence de notation d_{ij} entre deux compétiteurs i et j par 400, la formule (5) devient :

$$\mu_{ij} = \frac{1}{1 + 10^{-1}} = 0,90909$$

et

$$\mu_{ji} = \frac{1}{1 + 10^{1}} = 0,09090$$

L'équipe i a donc effectivement une probabilité dix fois plus grande de l'emporter par rapport à l'équipe j.

1.1.2.3 h – *home advantage*

Certains sports joués dans des environnements différents prennent cependant en compte un paramètre supplémentaire. Ce paramètre est appelé « home advantage » dans la littérature et correspond à un avantage intangible en faveur d'une équipe lorsqu'elle évolue à domicile. Selon Pollard (2008), ces facteurs sont pour la plupart psychologique, tel que la familiarité avec l'environnement, la territorialité ou encore une influence du public sur l'arbitrage. Ces équipes sont donc avantagées, comme le confirme l'historique des résultats. Les équipes évoluant à domicile sont historiquement avantagées ramenant ainsi, selon notre base de données comprenant les 4 dernières années de football international, 63% des points contre 37% pour l'équipe visiteuse. La correction régulièrement apportée pour neutraliser cet effet au mieux, est d'ajouter une valeur h de 100 points dans le modèle de prédiction pour l'équipe jouant à domicile, ce qui correspond à 64%. Cependant ce paramètre h peut prendre nimporte quelle valeur, pour autant qu'elle témoigne au mieux de l'avantage pour l'équipe évoluant à domicile lors du match en question.
La formule (5) se voit donc modifiée pour l'équipe évoluant à domicile tel que :

$$\mu_{ij} = \frac{1}{1 + 10^{-(d_{ij}+h)/400}} \tag{9}$$

Et pour l'équipe à l'extérieur :

$$\mu_{ji} = \frac{1}{1 + 10^{-(d_{ji}-h)/400}} \tag{10}$$

1.1.2.4 K - Niveau de compétition

Le coefficient K joue un rôle primordial lors de la mise en place de ce type de classement. Il détermine en effet la vitesse à laquelle va évoluer la valeur d'un

joueur/équipe et donc son classement en fonction de ses performances. Un coefficient K trop bas ne changera pas suffisamment rapidement le classement du joueur/équipe enchainant les victoires et à l'inverse, un K trop haut rendra le classement trop volatile et donc inefficace. Dans la plupart des sports et compétitions utilisant le système Elo, ce facteur est régulièrement lié au niveau de la compétition, que ce soit l'expérience du joueur ou l'importance de la compétition lors de laquelle se déroule la rencontre.
A titre d'exemple, on retrouve ci-dessous les valeurs de K dans les classements Elo des fédérations de deux sports que sont les échecs et le football.

K aux échecs selon la FIDE (Fédération internationale d'échecs) :

- $K = 40$ pour un joueur nouveau sur la liste de classement jusqu'à ce qu'il ait terminé des épreuves avec au moins 30 parties.
- $K = 20$ tant que la valeur d'un joueur reste inférieure à 2400.
- $K = 10$ une fois que la cote publiée d'un joueur a atteint 2400 et reste à ce niveau par la suite, même si la cote tombe sous 2400.
- $K = 40$ pour tous les joueurs jusqu'à leur 18ème anniversaire, tant que leur classement reste inférieur à 2300.

The following table shows the difference in importance of the competitions

Match importance	Match importance factor (M)	• Basis factor K= 15
FIFA Women's World Cup Match	4	K = 4 * 15 = 60
Women's Olympic Football Tournament	4	K = 4 * 15 = 60
FIFA Women's World Cup qualifier	3	K = 3 * 15 = 45
Women's Olympic Football Tournament Qualifier	3	K = 3 * 15 = 45
Women's Continental Finals	3	K = 3 * 15 = 45
Women's Continental Qualifier	2	K = 2 * 15 = 30
Women's friendly match	1	K = 1 * 15 = 15
Women's friendly match between two top 10 teams	2	K = 2 * 15 = 30

FIGURE 2 – Valeurs de K dans le classement FIFA féminin.

Le K utilisé par la FIFA dans son classement pour les équipes nationales féminines sert pour sa part à différencier les niveaux et l'importance des rencontres, afin qu'une victoire dans une grande compétition rapporte plus de points qu'une victoire dans un match amical à variables égales.

Cette valeur de K peut également être optimisée de façon à rendre le classement plus prédictif, c'est notamment le cas dans la thèse de Chen (2018) où la paire de paramètres h et K est optimisée afin d'obtenir la meilleure capacité prédictive

possible pour chaque classement Elo développé.

1.1.3 *Propriétés*

Le système d'évaluation Elo est caractérisé par trois propriétés, que chaque modèle, peu importe les modifications apportées, se doit de respecter.

1. La somme des notes de tous les joueurs/équipes participant à une compétition à n'importe quel moment de celle-ci est égale à la somme des notes initiales de tous les joueurs/équipes.

2. La somme des points gagnés et perdus par les deux adversaires lors d'une rencontre est égale à 0.

3. Un joueur/équipe battant un opposant plus fort, ayant donc une note supérieure à la sienne avant leur affrontement, se voit récompenser davantage qu'un compétiteur battant un adversaire moins fort que lui. Dans le cas d'une égalité, le joueur/équipe plus fort ne peut gagner plus de points que le joueur/équipe plus faible.

Les deux premières propriétés sont toujours remplies lorsque les sommes respectives des scores actuels et attendus lors d'une rencontre sont égales à 1. Concernant les scores actuels, qu'ils soient basés sur le résultat -victoire, match nul, défaite- ou le nombre de points/buts marqués par chaque joueur/équipe, la somme des résultats S_{ij} et S_{ji} devra toujours être égale à 1. Si cette règle de $S_{ij} + S_{ji} = 1$ est respectée, alors la somme de toutes les notes Elo des joueurs/équipes $r_i(t)$ à n'importe quel instant $t > 0$ sera toujours égale à la somme de toutes les notes Elo initiales des joueurs/équipes $r_i(0)$, ici égale à σ.

$$\sum_{k=1}^{m} r_k(0) = \sigma \Longrightarrow \sum_{k=1}^{m} r_k(t) = \sigma \tag{11}$$

pour tout $t > 0$.

Cela signifie donc que si la note initiale pour chaque joueur/équipe est fixée à une valeur arbitraire telle que x, la moyenne des notes de tous les participants à n'importe quel instant de la compétition sera toujours égale à x. La propriété numéro 1 est donc validée.

Il en va de même pour la somme de μ_{ij} et μ_{ji}, égale à 1 selon l'équation (5). Ces deux équations prisent ensemble nous donnent une équation valable après chaque rencontre sur le nombre de points gagnés et perdus par chaque adversaire. Ces mises à jour de points pour les deux équipes/joueurs telles que décrites en (1) donnent donc :

$$K(S_{ij} - \mu_{ij}) + K(S_{ji} - \mu_{ji}) = 0 \tag{12}$$

Cette équation confirme la nullité de la somme des points gagnés et perdus et valide la 2ème propriété. Donc, la somme des notes de toutes les équipes m après une rencontre opposant i à j prend la forme suivante :

$$\begin{aligned}\sum_{k=1}^{m} r_k(new) &= \sum_{k\neq i,j}^{m} r_k(old) + [r_i(old) + K(S_{ij} - \mu_{ij})] + [r_j(old) + K(S_{ji} - \mu_{ji})] \\ &= \sum_{k=1}^{m} r_k(old)\end{aligned} \tag{13}$$

Avec la somme totale des notes des équipes inchangée. La 3ème propriété est validée en suivant le raisonnement suivant : Prenons une équipe i plus forte qu'une autre équipe j avant leur rencontre, $r_i(old) > r_j(old)$, donc

$$-(r_i(old) - r_j(old))/400 < 0 \tag{14}$$

$$-(r_j(old) - r_i(old))/400 > 0 \tag{15}$$

$$\implies 10^{-(r_i(old)-r_j(old))/400} < 10^{-(r_j(old)-r_i(old))/400} \tag{16}$$

Ajoutons maintenant 1 de chaque côté de l'équation avant de prendre l'inverse de chacun de ces côtés. Nous obtenons :

$$\frac{1}{1 + 10^{-(r_i(old)-r_j(old))/400}} > \frac{1}{1 + 10^{-(r_j(old)-r_i(old))/400}} \implies \mu_{ij} > \mu_{ji} \tag{17}$$

Si l'équipe la plus faible l'emporte, j, sa nouvelle note devient donc :

$$r_j(new) = r_j(old) + K(1 - \mu_{ji}) \tag{18}$$

A l'inverse, si l'équipe la plus forte gagne, l'équipe i ici, sa nouvelle note devient :

$$r_i(new) = r_i(old) + K(1 - \mu_{ij}) \tag{19}$$

Or, comme le démontre l'équation (17), $\mu_{ij} > \mu_{ji}$, et K étant un paramètre positif, le gain de points est supérieur pour une équipe plus faible, telle que j, par rapport à celui d'une équipe plus forte, telle que i. Le raisonnement est identique lors d'un match nul, où l'on remplacerait la valeur de S d'une victoire égale à 1 par celle d'un match nul égale à $0,5$, dans les équations (18) et (19). Nous obtenons par ailleurs une conclusion identique, avec une équipe plus faible davantage récompensée qu'une équipe plus forte lors d'un match nul entre ces deux équipes.

1.2 *Modèles concurrents*

1.2.1 *Méthode des chaînes de Markov*

Cette méthode de classement est inspirée d'une ancienne technique, appelée aujourd'hui Chaines de Markov et développée en 1906 par A. A. Markov. D'abord utilisée par Markov pour une analyse linguistique des séquences de consonnes et voyelles dans un poème de Pushkin du nom d'Eugene Onegin, elle fut ensuite utilisée dans bien d'autres applications dont les compétitions sportives (Langville et Meyer, 2012).

Popularisée par son utilisation dans la classification de pages lors de recherches sur le web, avec son Google's PageRank développé par Page et al. (1999), la méthode de Markov fut ensuite adaptée à une multitude de domaines, dont le sport. Kvam (2006) l'a adapté au basketball universitaire et la NCAA avant que Govan (2008) l'applique à son tour au football américain avec la NFL et la NCAA.

La méthode a ensuite été adaptée au football par Lasek et al. (2013) en suivant l'idée de base suivante, chaque match est une opportunité pour l'équipe la plus faible de voter pour l'équipe la plus forte. Les équipes votent en effet proportionnellement aux résultats qu'elles ont obtenus contre ces différentes équipes. L'équipe contre laquelle elle obtient les meilleurs résultats aura donc le plus faible pourcentage de ses votes, et à l'inverse l'équipe contre laquelle elle aura obtenu les moins bons résultats se verra affecté le plus haut pourcentage des votes. On prendra ici l'exemple d'un fan sans attaches particulières à une équipe et apportant son support à une des équipes ayant battu celle pour laquelle il tenait jusqu'alors selon le principe discuté de pourcentage de votes.

Ce premier exemple est basé sur le score final des rencontres avec la probabilité que le fan préfère l'équipe j à l'équipe i selon une probabilité $\hat{p}_{ij}$ telle que :

$$\hat{p}_{ij} = \frac{w_{ji} + 1}{w_{ij} + w_{ji} + 2} \tag{20}$$

Avec w_{ji}, élément de la matrice W représentant le nombre de victoires de j contre i et w_{ij}, le nombre de victoires de i sur j. Les matchs nuls sont considérés dans ce modèle comme une demi-victoire et une demi-défaite. La probabilité pour que le fan reste supporter de son équipe actuelle est donc :

$$\hat{p}_{ii} = 1 - \sum_{j \neq i} (\hat{p}_{ij}) \tag{21}$$

Dans un sport comme le football et qui plus est lorsqu'il s'agit d'équipes nationales il est très fréquent que toutes les équipes ne s'affrontent pas pendant un long laps de temps. Afin de remédier à cela, il faut tout d'abord entrer les valeurs de probabilités $\hat{p}_{ij}$ dans une matrice carrée, dont les entrées sont donc égales à $(\hat{p}_{ij})_{i,j=1,2,...,n.}$. Il faut ensuite normaliser les lignes de cette matrice carrée en les divisant par n, afin d'obtenir une matrice stochastique M aux entrées égales à $(p_{ij})_{i,j=1,2,...,n.}$ et modélisant le comportement du fan. Cette matrice stochastique est très importante pour l'élaboration du vecteur d'évaluation des équipes.
La matrice E est composée de $n \times n$ rangs dont l'ensemble des entrées est égal à $\frac{1}{n}$, avec n étant le nombre d'équipes, afin d'obtenir une distribution stationnaire. La combination des deux matrices nous donne une nouvelle matrice $\tilde{M}$:

$$\tilde{M} = \alpha M + (1 - \alpha)E \tag{22}$$

avec le paramètre α compris entre 0 et 1.

Pour déterminer la valeur de chaque équipe il faut ensuite faire une distribution stationnaire de la chaine telle que $\pi = \pi\tilde{M}$ où la coordonnée i du vecteur π nous donne la valeur correspondant à l'équipe i, $r_i = \pi_i$.

Il existe cependant une infinité de possibilités et de matrices M, notamment un modèle basé sur la différence de buts lors des rencontres entre deux équipes.

$$\hat{p}_{ij} = \frac{g_{ji} + 1}{g_{ij} + g_{ji} + 2} \tag{23}$$

Avec ici g_{ji}, élément de la matrice de différence de buts G, représentant le nombre de buts inscrits par l'équipe j contre l'équipe i et g_{ij} le nombre de buts marqués par l'équipe i. La méthode est identique à celle appliquée à la matrice basée sur le résultat du match pour obtenir la valeur de chaque équipe.

1.2.2 *Méthode des moindres carrés de Massey*

Massey, alors étudiant au collège de Bluefield, crée en 1997 une nouvelle méthode permettant de classer les équipes universitaires de football américain. Cette méthode, basée sur les moindres carrés va être appelée ici sous le nom de méthode de Massey, bien que son nom soit associé à d'autres méthodes.

Cette méthode de Massey, basée sur la loi des moindres carrés, peut se synthétiser sous une simple équation :

$$y_k = r_i - r_j + \varepsilon \tag{24}$$

Où y est la différence de buts lors d'un match k entre les équipes i et j avec r étant leurs classements respectifs. Les paramètres de régression correspondant au vecteur r dans la formule suivante, sont obtenus en utilisant la méthode des moindres carrés, minimisant $\sum {e_i}^2$. L'ensemble des rencontres m entre les équipes n peut donc être synthétisé dans un système d'équations sous la forme matricielle X de $m \times n$ rangs tel que :

$$Xr = y \tag{25}$$

Massey a ensuite découvert l'avantage d'une matrice M égale à $M = X^T X$. Il s'est rendu compte qu'en multipliant le membre de droite de l'équation de base (25), il obtenait $X^T y$, qui pouvait également être formé en accumulant la différence de buts des équipes. En effet, l'élément i du vecteur $X^T y$ est la somme des différences de buts lors de chaque match joué par l'équipe i lors de la compétition. $X^T y$ peut donc être défini tel que p et le système des moindres carrés de Massey devient :

$$Mr = p \tag{26}$$

Où $M_{n\times n}$ est la matrice décrite ci-dessus, $r_{n\times 1}$ le vecteur des notes inconnues et $p_{n\times 1}$ le vecteur cumulatif des différences de buts. Il est également possible d'intégrer l'avantage à domicile dans le modèle, comme a pu le faire Massey (1997), en ajoutant à son équation de base (24) les termes r_h et x_h :

$$y_k = r_i - r_j + r_h x_h + \varepsilon \tag{27}$$

avec x_h étant égal à 1 lorsque l'équipe i évolue à domicile et -1 lorsqu'elle évolue à l'extérieur. Le paramètre r_h, définissant l'avantage pour chaque équipe évoluant à domicile, est lui à déterminer avant le début d'une compétition afin d'être utilisé lors de chaque rencontre de celle-ci.

1.2.3 *Modèle de Bradley-Terry*

Le modèle Bradley-Terry est un modèle de comparaison par paires très simple et fréquemment utilisé dans la prédiction de rencontre où des individus, ou des équipes, sont jugées par paires.
Ce modèle suggéré par Bradley et Terry en 1952, suit la formule suivante :

$$p_{ij} = \frac{y_i}{y_i + y_j} \tag{28}$$

Cette formule mesurant la probabilité p_{ij} qu'un compétiteur i s'impose face à un adversaire j est cependant connue depuis longtemps, apparaissant déjà chez Zermelo en 1929. On y retrouve y_i, un paramètre positif représentant la valeur globale associée au compétiteur i, lors de chaque comparaison opposant i à l'adversaire j.

Cependant, à l'instar de nombreux modèles de comparaison par paires, les premiers modèles de Bradley-Terry n'incluaient pas la possibilité d'une égalité entre deux compétiteurs lors d'un match. Ils forçaient donc pour obtenir une préférence entre les deux joueurs/équipes, sinon les ignoraient ou allouaient aléatoirement la victoire à l'un ou l'autre.

Davidson (1970), voulant appliquer le modèle Bradley-Terry à des résultats où la possibilité d'égalité existe, développa le modèle suivant :

$$p_{i1} = \frac{y_i}{y_i + y_j + K\sqrt{y_i y_j}} \tag{29}$$

$$p_{i2} = \frac{K\sqrt{y_i y_j}}{y_i + y_j + K\sqrt{y_i y_j}} \tag{30}$$

$$p_{i3} = \frac{y_j}{y_i + y_j + K\sqrt{y_i y_j}} \tag{31}$$

Avec p_{i1}, p_{i2} et p_{i3} les probabilités respectives de victoire, match nul et défaite pour l'équipe/joueur i face à l'équipe/joueur j. On retrouve dans la formule y_i et y_j, les paramètres positifs déterminant la valeur de chaque équipe/joueur ainsi que K, un index représentant l'effet d'égalité et particulier à chaque problème.

Dans leur application au football Wang et Vandebroek (2013) ont modifié à leur tour le modèle de Davidson afin d'y ajouter l'avantage pour l'équipe évoluant à domicile. Cet avantage est bien connu dans le monde du football et les différents facteurs influençant ou amenant à celui-ci furent décrits par Pollard en 2008. Cet avantage, est inclus tel que le paramètre h dans les équations 29 à 31 :

$$p_{i1} = \frac{hy_i}{hy_i + y_j + K\sqrt{hy_i y_j}} \tag{32}$$

$$p_{i2} = \frac{K\sqrt{hy_i y_j}}{hy_i + y_j + K\sqrt{hy_i y_j}} \tag{33}$$

$$p_{i3} = \frac{y_j}{hy_i + y_j + K\sqrt{hy_i y_j}} \tag{34}$$

h augmente donc la valeur de l'équipe à domicile, l'équipe i, en multipliant sa valeur par un paramètre déterminé auparavant. Ce modèle peut ensuite être développé sous forme de classement comme ont pu le faire Wang et Vandebroek (2013), en utilisant une estimation maximum de vraisemblance. Cette estimation suit l'équation suivante :

$$L = \prod_{i=1}^{N} \prod_{j=1}^{3} p_{ij}^{x_{ij}} \tag{35}$$

Avec x_{ij} égal à 1 lorsque le résultat du match i est égal à j, sinon 0. j étant égal à 1 pour une victoire de l'équipe à domicile, 2 pour un match nul et 3 pour une victoire de l'équipe à l'extérieur. Les probabilités p_{ij} sont identiques à celles décrites plus haut (32 à 34). Les paramètres représentant les valeurs de chaque équipe sont ensuite définis afin de maximiser le logarithme d'estimation de maximum vraisemblance. Ensuite chaque équipe peut être classée selon son paramètre d'estimation de valeur.

1.3 *Choix du modèle*

Ces différents modèles d'évaluation d'équipes ont cependant plusieurs désavantages lorsqu'ils sont appliqués à une compétition de football comme la Coupe du Monde, comparés au modèle Elo décrit dans la section précédente.

Le premier est la meilleure capacité prédictive des modèles Elo par rapport à ces autres modèles, démontrée par Lasek (2013) dans son étude sur la capacité prédictive de différents modèles appliqués aux rencontres de football entre les équipes nationales de 2006 à 2012.
Le second est l'utilisation récurrente de l'Elo en tant que système d'évaluation dans le football, que ce soit au sein de la Fédération internationale de football association (FIFA) pour les classements féminins et masculins des équipes nationales, comme auprès des suiveurs, qui ont développé le site *Eloratings.net* comme alternative à l'ancien système d'évaluation utilisé par la FIFA. Dans un souci de cohérence, et sachant que l'on souhaite avant tout comparer les buts réels aux *expected goals*, l'utilisation de l'Elo est donc sans doute plus juste.
Ajoutons à cela un ajustement du modèle Elo assez intuitif, ce qui va permettre de modifier les paramètres du modèle autant de fois que nécessaire.

Pour ces différentes raisons, le système d'évaluation Elo a été choisi afin de déterminer les différences entre les classements basés sur les buts réels et ceux sur les

expected goals ainsi que pour tester la capacité prédicitive des *expected goals* sur les scores et résultats futurs lors d'une compétition.

2 expected goals

2.1 *Historique et état de l'art*

Bien que le terme « *expected goals* » soit plutôt récent dans le sport, il n'en est pas moins que l'étude de ses composantes a commencé il y a plus de 50 ans. En effet, en 1968, Reep et Benjamin étudiaient le football au travers d'une division du terrain en plusieurs zones distinctes, dans lesquelles ils analysaient la proportion de passes ainsi que de tirs effectués dans ces zones. Ils déterminèrent alors un ratio de buts marqués par rapport au nombre de tirs tentés dans une zone désignée alors comme la zone de tir. Etant donné la faible quantité de données à l'époque, cette zone de tir était déterminée comme étant le dernier quart de terrain de chaque équipe.

Ce n'est ensuite que vers la fin des années 1990 et au début des années 2000 que l'on a pu découvrir de nouveaux modèles très proches de ceux que l'on étudie aujourd'hui dans le football. Pollard et al. (2004) proposaient eux un modèle incluant deux variables augmentant significativement les chances de marquer. Ces variables sont le positionnement du tireur par rapport au but ainsi que son démarquage par rapport au défenseur le plus proche, qui devait être d'au moins un mètre.
L'étude de Ensum et al. (2005) était similaire dans la méthodologie mais comprenait neuf variables allant du positionnement du tireur, à l'endroit où le ballon avait été récupéré par son équipe.
Quelques années plus tôt, Pollard et Reep (1997) concentraient leur étude non pas uniquement sur les tirs mais sur les possessions de balle de chaque équipe. Ces possessions avaient une certaine probabilité d'être ponctuée par un tir qui lui-même avait une certaine probabilité d'être marqué. Cette dernière probabilité suivait 4 variables, l'emplacement du tireur, son démarquage d'au moins un yard (0,91m), si le tireur avait touché plus d'une fois le ballon avant son tir ainsi que si l'occasion provenait d'une phase arrêtée ou d'une action de plein jeu. A cela on notait déjà une différenciation entre les tirs effectués du pied, des têtes, ayant chacun leur propre régression.

Cependant le football n'est pas le seul sport utilisant aujourd'hui ces modèles d'*expected goals*. Ils sont également massivement utilisés par les analystes de l'autre côté de l'Atlantique, dans la Ligue nord-américaine de hockey sur glace, la NHL. Le point commun de ces deux sports, en plus d'être désormais riches en données, est le faible total de buts marqués dans un match. Il est donc très difficile d'évaluer les

performances des équipes étant donné l'importance de la chance et de la réussite face au but sur un intervalle de temps aussi restreint que les 90 minutes d'un match de football ou 60 pour le hockey sur glace (Reep, C., & Benjamin, B. ,1968). Cependant, sur le long-terme, il a été déterminé par Hill (1974) que la technique et le talent des joueurs prenaient le dessus sur ce facteur chance.

C'est donc au début des années 2000 que les premiers modèles de ce que l'on appelle aujourd'hui les *expected goals* font leur apparition en NHL. En 2004, Alan Ryder étudiait les probabilités de buts en hockey sur glace en se basant sur la qualité de chaque tir. Cette qualité du tir était déclinée en deux variables, sa distance par rapport au but et son type (*slap, snap, tip-in*, etc.). Cette étude a démontré une indépendance de cette qualité de tir par rapport au nombre de tirs tentés ainsi qu'une variation significative de la qualité des tirs selon les équipes. Il a ensuite appliqué son modèle au domaine défensif et a obtenu de nouvelles statistiques comme la qualité des tirs concédés par chaque équipe, apportant ainsi une nouvelle information sur la qualité défensive de chaque équipe.

C'est en 2012 et toujours en NHL que l'on voit pour la première fois apparaitre le terme *expected goals*. Macdonald (2012) essayait de trouver une alternative aux deux systèmes de notations alors omniprésents dans le hockey sur glace, celui de Fenwick et de Corsi. Son modèle d'*expected goals* se basait sur plusieurs variables telles que les buts, les tirs, les contacts encaissés et provoqués ainsi que les mises en jeu. En appliquant ces variables dans une régression linéaire et une régression « ridge » sur une base de données de 4 saisons de NHL allant de 2007 à 2011, ces deux modèles ont surperformés par rapport à leurs prédécesseurs que sont Fenwick et Corsi. Autant sur l'erreur quadratique moyenne des résultats passés que la prédiction des résultats à venir, les modèles d'*expected goals* étaient plus performants.

Les modèles se sont ensuite multipliés dans le football sur les blogs spécialisés.

Les variables les plus récurrentes dans ces modèles sont les suivantes :

	Type de tir	Emplacement du tireur	Type de passe décisive	Type de phase de jeu	Grosse occasion
11tegen11.net	X	X	X	X	
Michael Bertin	X	X	X		
Paul Riley		X			
Alex Rathke		X			
Garry Gelade	X	X		X	X
Michael Caley	X	X	X	X	X

TABLE 2 – Variables incluses dans les différents modèles d'*expected goals*

On retrouve donc les variables des différents modèles qui en font leur particularité. Paul Riley (2014) et Alex Rathke (2017) ont développé chacun un modèle le plus épuré possible avec comme seule variable l'emplacement du tireur. Ces modèles sont de bons indicateurs de buts et ont comme principal avantage d'être très simples. Les modèles plus complexes sont légèrement meilleurs indicateurs mais nécessitent un nombre de variables important. C'est le cas notamment du modèle de Michael Caley (2015), qui a essayé de prendre en compte chaque aspect d'une opportunité de but avec un modèle comprenant 6 régressions différentes pour autant de types de tir et un nombre important de variables comprises dans chacune des régressions.

Bien que tous ne soient pas d'accord sur les variables à inclure pour développer le modèle parfait, d'autres essayent déjà de voir plus loin. C'est le cas de Spearman (2018) dont le but est de déterminer la probabilité qu'aurait un joueur n'étant pas en possession du ballon de marquer. Il faut pour cela déterminer la probabilité qu'a chaque joueur de recevoir le ballon lorsqu'il n'en est pas en possession ainsi que la probabilité qu'il aurait de marquer s'il en était à la réception. L'équipe en possession de la balle devrait donc amener la balle avec succès jusqu'à un joueur en position de tir. Les possessions offensives sont donc divisées en trois phases, la transition, le contrôle du ballon, ainsi que la probabilité de marquer. Son modèle est selon lui perfectible, mais cela montre un intérêt dans la recherche au-delà des *expected goals* et un avant-goût de ce que pourraient nous offrir les statistiques dans le futur dans le sport.

2.2 *Modèle de Michael Caley*

Dans cette section, je vais me pencher en détail sur la méthodologie développée par Michael Caley, cité ci-dessus. Le modèle utilisé pour calculer les données d'*expected goals* que j'ai récolté pour les matchs de la Coupe du Monde 2018 est donc basé sur les variables suivantes.

Le type de tir

Pour calculer les *expected goals*, il faut donc séparer la formule en six équations distinctes et autant de types de tir. En effet, pour des raisons liées au football et au jeu, ces six types de tir possèdent leur propre équation :

- Tirs venant d'un coup-franc direct (frappe directe depuis l'endroit où a été commis la faute)
- Les tirs suivant un dribble du gardien
- Les têtes dont la passe décisive est un centre
- Les têtes dont la passe décisive n'est pas un centre
- Les tirs (têtes non comprises) dont la passe décisive est un centre
- Les tirs (têtes non comprises) dont la passe décisive n'est pas un centre (tir « normal »)

On peut en effet faire une première distinction assez claire entre les tirs effectués de la tête et les autres. La probabilité de marquer pour un tir effectué de la tête ne varie pas comme un tir avec le pied, notamment lorsque la distance par rapport au but augmente. L'équation pour une tête effectuée à l'entrée de la surface, soit plus d'une quinzaine de mètres, ne peut donc pas être identique avec les autres types de tir, au vu de la différence de puissance.
Le deuxième point majeur est la passe décisive, et si le centre provient d'un centre ou non. Concernant les têtes il sera beaucoup plus difficile de marquer, plus la distance augmente, mais encore plus si la passe décisive n'est pas un centre. Le ballon qui pourrait être un dégagement ou venant d'un cafouillage aurait significativement moins de vitesse et la puissance donnée à ce tir de la tête serait bien moindre.
A l'inverse, lorsque l'on effectue une tête, l'angle de tir lorsque la passe décisive est un centre a moins d'importance. Dans ce genre de phase, il est en effet plus important pour un joueur de frapper le ballon avec sa tête en ayant un bon contact avec le ballon, ce qui donne une puissance supérieure peu importe l'angle.

Les tirs effectués sur coup-franc ont aussi leur propre équation. Etant donné qu'ils ne proviennent pas de passes décisives, les angles et la distance par rapport au but

varient différemment d'un tir effectué de plein jeu. Identiquement, un tir à la suite d'un dribble du gardien suit sa propre équation. Le gardien ayant été dépassé, la probabilité de but augmente significativement et la difficulté du tir est moindre.

L'emplacement du tir

Bien que les modèles d'*expected goals* soient plutôt récents, l'étude de cette composante, très importante pour définir nos *expected goals* aujourd'hui, n'est pas si récente. En effet, il y a plus de 50 ans déjà, Reep et Benjamin (1968) étudiaient les proportions de passes et de tirs effectués dans des zones distinctes avec déjà un ratio de tir par rapport au nombre de buts marqués dans une zone désignée alors comme la zone de tir. Aujourd'hui, avec la multiplication des données nous pouvons tenter d'être plus précis qu'en déterminant le dernier quart du terrain comme zone de tir. Paul Riley (2014) et Alex Rathke (2017) utilisent oar exemple eux aussi des zones dans leurs modèles d'*expected goals*, 46 et 8 zones respectivement, ayant chacune une probabilité différente en termes d'*expected goals*. Les modèles plus développés comme celui de Caley se basent sur des formules qui permettent de définir la position du tireur par rapport au but, comme expliqué ci-dessous.

On peut voir ci-contre (Figure 3) la probabilité pour un tir (non exécuté de la tête et ne provenant pas d'un centre) d'être marqué suivant sa position.
Le taux de conversion est donc lié à deux variables, la distance à laquelle se trouve le tireur et l'angle par rapport au but. Il est en effet plus compliqué pour un joueur de marquer lorsqu'il il se trouve aux 16 mètres dans l'angle de la surface de réparation, plutôt que centré face au but à la même distance.
Aucun modèle d'*expected goals* n'est encore parvenu à tirer une équation de cette forme. Caley a donc divisé cela en 5 composantes qui sont : la distance par rapport au but, l'inverse de la distance par rapport au but, l'angle relatif par rapport au but, son inverse ainsi que l'inverse du produit de l'angle et de la distance. Ces composantes ont une combinaison spécifique pour chaque type de tir discuté précédemment.

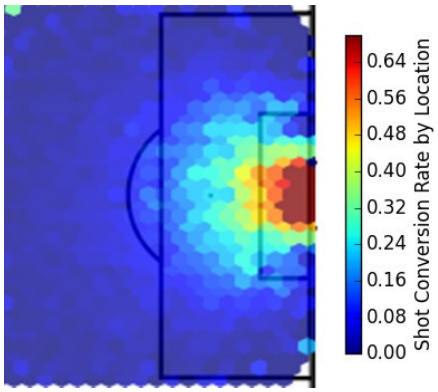


FIGURE 3 – Caley, M. (2015) Taux de conversion d'un tir selon l'emplacement Référence : https://cartilagefreecaptain.sbnation.com/2015/10/19/9295905/premier-league-projections-and-new-expected-goals.

L'angle relatif est calculé de la façon suivante, tout tir dont le point de départ est compris, perpendiculairement par rapport au but, entre les deux poteaux aura un angle égal à 1. Pour le tirs pris à l'extérieur de cette zone, l'angle est pris par rapport au poteau le plus proche. Un tir effectué à 45° par rapport au poteau le plus proche aura donc une valeur de 0,5.

Le type de passe décisive

L'importance des passes décisives dans les modèles d'*expected goals* tient dans une raison majeure, l'élimination de défenseurs adverses. Chaque type de passe décisive a, selon sa particularité, une probabilité plutôt grande ou petite d'éliminer un certain nombre d'adversaires. Les centres étant un type de passe décisive amenant des tirs de moins bonne qualité, ils sont étudiés séparément.

Le premier type de passe et le plus probable d'amener un but est la passe en profondeur. Cette passe est définie selon le glossaire de *Whoscored.com* comme une passe tentée, et réussie dans le cas présent, entre les défenseurs adverses pour trouver un coéquipier dans la course et en direction du but. La notion d'élimination des défenseurs adverses est donc bien présente et cette passe étant donnée entre les défenseurs et arrivant dans le dos des défenseurs est donc très difficile à contester. Une variante de ce type de passe décisive, encore plus dangereuse en termes d'*expected goals*, est la passe décisive à la suite d'une passe en profondeur. La passe en profondeur ayant déjà éliminé les défenseurs adverses, une seconde passe à la suite de celle-ci est souvent synonyme de situation de un-contre-un pour l'attaquant.

Un autre type de passe très efficace lorsqu'elle est réussie est la passe « face au but ». Cette passe décisive est une passe donnée depuis une position excentrée pour un joueur situé face au but. Cette passe est une des plus dangereuses mais aussi une des plus compliquées à réaliser. Cette zone située face au but étant souvent couverte par un ou plusieurs défenseurs il est donc très difficile qu'elle trouve un coéquipier, mais lorsqu'il est trouvé, il est souvent situé dans une position de tir très favorable.

Une passe décisive cette fois-ci impactant négativement le modèle d'*expected goals* est la passe en retrait. Permettant souvent de trouver un joueur démarqué, elle a cependant beaucoup moins de chance d'effacer un adversaire et le tireur se trouvera donc régulièrement dans une position de tir face à un ou plusieurs défenseurs. L'impact négatif est donc léger, mais comparé à une passe face au but ou une passe

en profondeur il est tout de même significativement inférieur.

La dernière donnée à prendre en compte dans ces passes est la zone dans laquelle elle est donnée. Il est ainsi démontré que plus la passe est donnée à proximité du but adverse, plus la valeur d'*expected goals* du tir en résultant est grande. Cela signifie donc qu'une passe donnée par un joueur se trouvant dans ce que l'on considèrerait comme une bonne position de tir est souvent synonyme d'une position encore meilleure pour le joueur recevant la balle. Les défenseurs étant donc concentrés sur le joueur à l'origine de la passe et dans une situation dangereuse, sont donc surpris et peuvent être pris de court pour contester le tir du joueur à la réception. Cette passe est donc très dangereuse tant pour la défense que pour l'attaque, car une perte de balle dans cette position ou une mauvaise exécution de cette dernière passe pourrait être considérée comme du gâchis.

Pour modéliser ces positions de passes, Caley a donc utilisé, tout comme pour les positions de tirs, l'inverse de la distance par rapport au but ainsi que l'angle de la position où a été donnée la passe décisive par rapport au but.

Le type de phase de jeu

La façon dont l'équipe se procure ses occasions est aussi déterminante dans l'élaboration d'un modèle d'*expected goals*. Dans le modèle de Michael Caley on dénombre cinq types d'occasions de buts.

— **Corners**
Ces situations sont considérées comme peu dangereuses. La défense étant regroupée et ayant souvent préparé ses phases ressort très souvent dominatrice de ce type de phase arrêtée. Ils influencent donc négativement les probabilités de buts.

— **Coups francs**
Les coups francs ont, quant à eux, un très léger impact positif sur les occasions de buts. Bien que défendus par un nombre important de défenseurs, ils sont, contrairement aux corners, donné dans des positions constamment changeantes. Cela permet donc une plus grande liberté pour le tireur et peut aussi donner lieu à des combinaisons parfois surprenantes pour la défense.

— **Contre-attaques**

Ces phases sont les plus dangereuses à défendre. Elles sont définies comme des actions où, lors de la récupération d'un ballon de plein jeu, l'équipe en possession se dirige instantanément vers le but sans faire circuler le ballon. Elles prennent souvent la défense à défaut, et sont très compliquées à défendre. L'équipe ayant perdu le ballon lors d'une phase de jeu est en effet souvent désorganisée par cette perte de balle soudaine, et doit défendre soit en reculant et donc en risquant d'être prise de vitesse, soit en courant dos tourné au jeu.

— **Attaques placées**
Ce type d'occasion est légèrement meilleur pour se procurer une occasion de but qu'une attaque de plein jeu « basique ». En effet, ces attaques placées impliquant au moins cinq passes tendent à prouver une certaine domination de l'équipe et évite toute précipitation face à une défense bien en place. Cette domination de l'espace influence donc positivement les chances de buts.

— **Attaques de plein jeu**
Considérée comme l'attaque de base elle n'impacte ni positivement ni négativement la probabilité de buts.

Autres facteurs

Ici nous allons répertorier les cinq autres facteurs inclus dans le modèle. Le premier est ce que l'on appelle une « grosse occasion ». Ces phases sont des occasions où la défense a été battue et dont la probabilité de marquer est d'environ 40%. La pression défensive lors de ces phases est presque nulle ce qui facilite l'attaque adverse. Garry Gelade (2017) utilise également ces « grosses occasions » dans son modèle d'*expected goals* et l'on peut voir sur ces graphes (Figure 4) à la fois la faible pression défensive ainsi que le nombre de joueurs entre le tireur et le but. Dans plus de 60% des « grosses occasions » on ne retrouve qu'un seul joueur, souvent le gardien, entre l'attaquant et le but adverse.

Un deuxième facteur est un tir suivant une erreur défensive. Ces erreurs défensives et pertes de balle dans des positions considérées comme dangereuses, proche de son but par exemple, mettent souvent la défense dans une situation très compliquée et offrent de très belles opportunités à l'adversaire.

Le facteur suivant concerne aussi l'élimination de défenseurs mais cette fois par le dribble. Il est en effet corrélé positivement avec la probabilité d'inscrire le tir qui

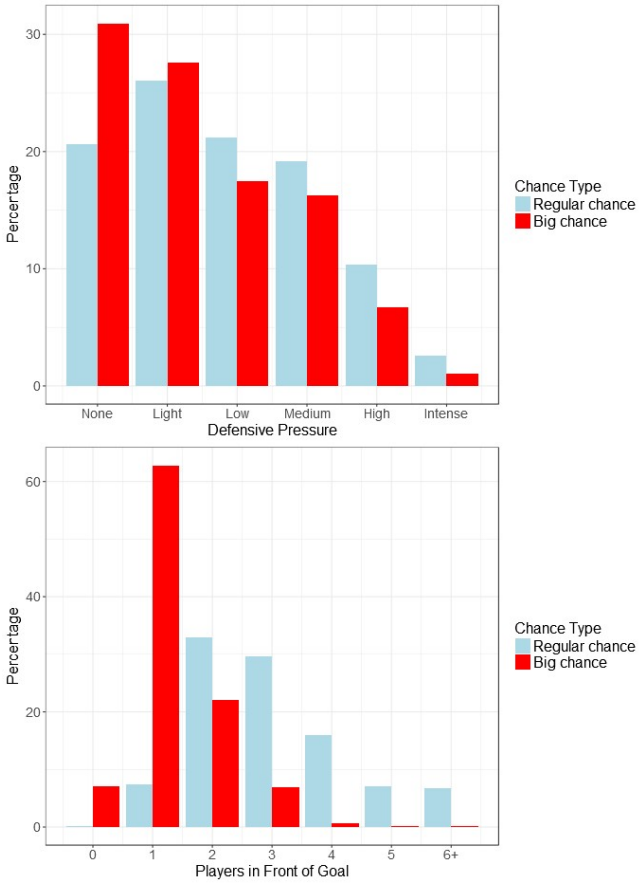


FIGURE 4 – Gelade, G. (2017)
Référence : http://business-analytic.co.uk/blog/assessing-expected-goals-models-part-2-anatomy-of-a-big-chance/.

s'en suit. Un joueur ayant dribblé un adversaire se trouvera très souvent dans une meilleure position de tir, étant donné l'élimination de son opposant le plus proche et le plus à même de contester sa tentative.

Le quatrième facteur est lié au rebond après un tir. Il se trouve qu'un tir suivant un rebond, que ce soit après un tir sur un des montants ou renvoyé par le gardien, a beaucoup plus de chance d'être marqué. La défense étant souvent statique et regardant l'issue du tir, si un attaquant parvient à frapper à la suite du rebond, il aura une tentative très peu défendue et avec un gardien en position désavantageuse à la suite de son arrêt ou tentative d'arrêt précédent.

Le dernier facteur, très léger celui-ci, est lié au résultat en cours. En effet une défense menée de trois buts ou plus sera moins concentrée et déterminée à empêcher l'équipe adverse d'en marquer un de plus, qu'une équipe menant d'un but et s'accrochant à sa victoire. L'effet est cependant très léger.

Finition du joueur

Le dernier point du modèle de Michael Caley est une question primordiale dans un modèle d'*expected goals*. Quel est le joueur ayant effectué le tir ? Il est en effet plus probable qu'un tir ayant à la base la même valeur d'*expected goals* soit marqué s'il est tiré par un attaquant redoutable face au but comme Lionel Messi plutôt qu'un défenseur, comme par exemple, Vincent Kompany. Le modèle est donc pondéré selon l'historique du joueur ayant tenté le tir. Certains joueurs ont prouvé qu'ils avaient pu constamment battre les modèles d'*expected goals* en marquant plus de buts, saison après saison, qu'ils n'auraient dû en marquer selon ces modèles. Afin de ne pas fausser le modèle final avec des joueurs ayant de très bonnes statistiques sur un nombre restreint de matchs, Caley a donc ajouté aux totaux de chaque joueur 75 buts marqués et 75 *expected goals* afin de lisser leurs ratios. Les joueurs parvenant à se démarquer seront alors considérés comme des très bons (ou très mauvais) finisseurs.

Equations des différents types de tir

Les équations détaillées des six types de tir dans le modèle de Michael Caley décrits en 2.2 sont définies en annexes.

3 méthodologie

Afin de déterminer le niveau des équipes lors de la Coupe du Monde nous allons nous baser sur deux indicateurs, d'un côté les résultats et scores des rencontres et de l'autre les *expected goals* lors de ces rencontres.

Pour le classement Elo basé sur les résultats et les scores des rencontres, nous allons suivre la méthodologie utilisée par Chen (2018), qui consiste à comparer les classements Elo utilisant les résultats et les différences de buts comme mesure de performance.
Ensuite nous déterminerons un autre classement basé quant à lui sur le jeu produit, les occasions créées et la qualité des occasions concédées. Le modèle Elo sera également utilisé mais en incluant cette fois les valeurs d'*expected goals* produites par chaque équipe à la place des buts réels.

Dans le but de déterminer les paramètres adéquats pour chaque type de classement, nous allons optimiser les moindres erreurs au carré (MSE) de chaque classement. Cette moyenne des erreurs au carré est utilisée afin de pénaliser plus grandement les lourdes erreurs de prédictions, par rapport à la moyenne des erreurs absolues (MAE) par exemple, qui pénalise "linéairement" les erreurs de prédiction. Hors en football, lorsque l'on évalue la différence de buts ou les scores de matchs, la différence entre une erreur de 2 ou 4 buts ne peut être évaluée du simple au double. En effet, comme on peut le voir dans la figure ci-dessous, les équipes parvenant à s'imposer sur des scores conséquents sont rares, une prédiction fortement erronée se doit donc d'être plus lourdement sanctionnée, afin d'impacter plus fortement le modèle. En effet, un modèle apprendra plus rapidement lorsque les erreurs sont grandes.

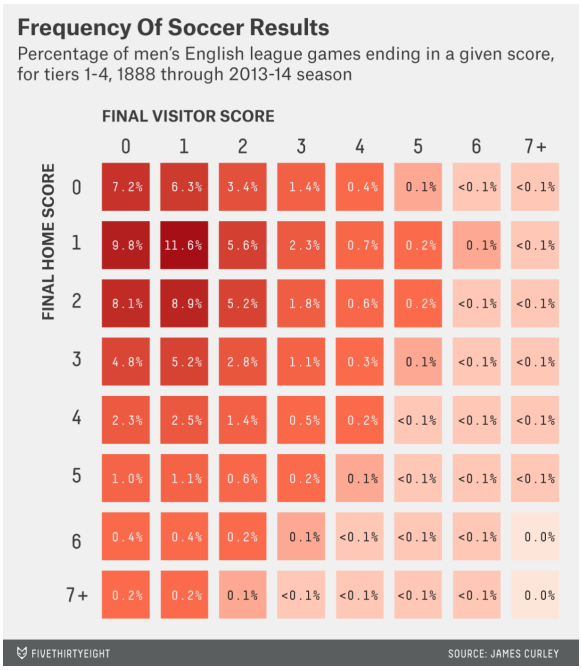


FIGURE 5 – Curley, J. (2014). Distribution des scores de football anglais professionnel de 1988 à 2014.
Référence : https://fivethirtyeight.com/features/in-126-years-english-football-has-seen-13475-nil-nil-draws/.

Nous obtiendrons ainsi un classement optimal évaluant les performances des équipes selon les buts réels et un classement optimal évaluant les performances des équipes ayant créé les meilleures occasions de buts et en ayant concédé peu.

En parallèle, des classements de joueurs basés à la fois sur leurs performances individuelles et collectives seront développés.

3.1 *Elo basé sur les buts réels*

Pour déterminer le système d'évaluation orienté sur les buts des équipes lors de cette Coupe du Monde 2018 nous allons donc utiliser un modèle Elo. Celui-ci possèdera les variables suivantes : une valeur initiale r pour chaque équipe discutée précédemment, le score ou résultat S, le score ou résultat attendu μ ainsi que le facteur K (paramètre fixé).

$$r(new) = r(old) + K(S - \mu)$$

Avec ici la valeur initiale étant $r(old)$ lors du premier match de la compétition pour chaque équipe.

***Valeur initiale** r*

Un système d'évaluation Elo requiert donc un paramètre de base qui est le niveau de chaque équipe au départ de la compétition. Cette note initiale r peut être identique pour toutes les équipes ou différente afin de différencier le niveau de chaque participant. Nous allons déterminer deux paramètres de base différents afin d'avoir une analyse plus complète des performances. L'une ne prenant en compte que les matchs ayant eu lieu durant la compétition et une autre incluant un certain historique de performance des équipes.

Pour la première, nous allons donc déterminer la valeur de chaque équipe à 1000, comme cela est fait dans le système Elo de base lorsqu'un nouveau joueur/équipe prend part à la compétition. Cette valeur de base signifie que l'on ne possède pas d'information sur la valeur de chaque équipe et donc qu'au début de la compétition elles partent toutes sur un pied d'égalité.

Pour la deuxième valeur initiale nous allons nous baser sur le classement publié par la FIFA. La dernière mise à jour de ce classement avant la Coupe du Monde datait d'une semaine avant le début de celle-ci, ce qui permet d'avoir une évaluation assez juste du niveau des équipes avant la compétition. Ce classement est malheureusement très étendu dans ses valeurs, avec notamment 1558 points pour l'Allemagne, meilleure équipe au classement, contre 457 points pour la Russie, dernière équipe au classement parmi celles participante à la Coupe du Monde. Des valeurs aussi dispersées dans un modèle Elo ne témoigneraient donc pas du réel niveau des équipes.
Cependant, la FIFA a effectué un récent changement dans son système de classement durant cette même Coupe du Monde 2018, en passant de leur ancienne méthode à un classement basé sur l'Elo. Afin d'effectuer cette transition et d'adapter les valeurs des équipes à un système d'évaluation Elo, la FIFA s'est basée sur le classement actuel des équipes C, ce qui donne la valeur initiale de chacune d'entre elles $r(init)$ grâce à l'équation suivante :

$$r(init) = 1600 - 4(C - 1) \tag{36}$$

Cela donne donc une nouvelle valeur initiale de 1600 pour l'Allemagne et de 1340 pour la Russie. Admettons maintenant une hypothétique rencontre entre les deux

nations où l'on tiendrait compte d'un avantage à domicile de h égal à 100 pour la Russie. Appliquées au modèle Elo de base et à la formule (9) vue au point 1.1.2.3, ces nouvelles valeurs initiales feraient passer les chances russes de remporter les points mis en jeu de 0,3% dans l'ancienne méthode d'évaluation à 28,5%. Bien que l'Allemagne reste la grande favorite d'une telle confrontation, cela donne des probabilités plus réalistes.

Score du match S

La première variante des modèles Elo étant basée sur les buts réels sous forme de résultat, le résultat S de ces modèles suivra la distribution suivante pour chaque équipe :

$$S = \begin{cases} 1 & \text{si victoire} \\ 0,5 & \text{si égalité} \\ 0 & \text{si défaite} \end{cases}$$

Pour la seconde basée sur les scores des matchs, nous allons nous baser sur la formule du S développée par Langville et Meyer (2012) dans leur livre au chapitre consacré à l'Elo, qui est la suivante :

$$S_{ij} = \frac{G_{ij} + 1}{G_{ij} + G_{ji} + 2}$$

Où G_{ij} sont les buts marqués par l'équipe i lors de la rencontre contre l'équipe j, et inversement pour G_{ji}.

Score attendu μ

La valeur du score attendu est calculée en fonction de deux autres variables que sont les notes de chaque équipe i et j avant leur rencontre et l'avantage à domicile h. Les équations suivantes sont les équations (9) et (4) décrites dans la partie théorique 1.1.2 de l'Elo :

$$\mu_{ij} = \frac{1}{1 + 10^{-(d_{ij}+h)/400}}$$

avec

$$d_{ij} = r_i(old) - r_j(old)$$

h est un paramètre positif qui augmente les probabilités de victoire d'une équipe évoluant à domicile plutôt que sur un terrain neutre ou à l'extérieur, toute chose étant égale par ailleurs. Cependant, afin de respecter la propriété du modèle Elo stipulant que la somme des scores attendus de l'équipe à domicile et à l'extérieur était toujours égale à 1, l'équipe évoluant à l'extérieur se voit donc soustraire cette valeur constante dans sa formule μ. On peut donc en déduire que son score attendu est égal à :

$$\mu_{ji} = 1 - \mu_{ij} \tag{37}$$

Lorsque le match est joué sur un terrain neutre, soit un pays étranger aux deux nations s'affrontant dans le cadre de rencontres internationales, la valeur de h est nulle. Lors de cette Coupe du Monde 2018 se déroulant en Russie, la valeur de h sera donc uniquement applicable lors des matchs de l'équipe nationale russe, mais nulle lors de tous les autres matchs n'impliquant pas la Russie.

Paramétrage de K et h

Afin d'obtenir la formule finale de chacun des modèles, il faut donc déterminer les paramètres h et K de façon la plus optimale possible. En effet, un K trop petit rendra le système de notation trop statique et ne donnerait pas la chance aux équipes d'augmenter leurs notes de façon suffisante à la suite d'un bon résultat. A l'inverse un K trop grand rendrait le système trop volatile et un seul résultat pourrait bouleverser la note d'une équipe. Il en va de même pour h qui doit être déterminé de façon à donner un aperçu réel de l'avantage qu'a une équipe lorsqu'elle joue à domicile. Ainsi, un h trop élevé donnerait à l'équipe locale un avantage présumé trop important, et inversement pour un h trop faible. Il faut donc trouver la paire de h et de K la plus optimale afin d'obtenir le classement le plus juste.

L'avantage du modèle Elo est qu'il possède déjà un terme prédictif, qui est $(S - \mu)$, soit la différence entre le résultat ou le score réel et le résultat ou le score attendu calculé selon la différence de valeur entre les deux équipes avant la rencontre. Cela signifie donc que plus le terme $(S - \mu)$ est faible, plus la prédiction est juste et plus l'évaluation du niveau des équipes par l'Elo est efficace.

La prédiction des erreurs au carré se fait donc comme suit :

$$MSE(i,j) = \frac{1}{M}\sum_{m=1}^{M}(S_{ijm} - \mu_{ijm})^2 + (S_{jim} - \mu_{jim})^2 \tag{38}$$

$$= \frac{1}{M}\sum_{m=1}^{M}(S_{ijm} - \mu_{ijm})^2 + [(1 - S_{ijm}) - (1 - \mu_{ijm})]^2 \tag{39}$$

$$= \frac{1}{M}\sum_{m=1}^{M}(S_{ijm} - \mu_{ijm})^2 + (\mu_{ijm} - S_{ijm})^2 \tag{40}$$

$$= \frac{1}{M}\sum_{m=1}^{M}2(S_{ijm} - \mu_{ijm})^2 \tag{41}$$

Avec i et j les deux équipes se rencontrant lors de chaque match m, matchs allant de 1 à M lors de la compétition. S_{ijm} et μ_{ijm} sont respectivement les résultats ou scores actuels et les résultats ou scores attendus lors de chaque match m opposant i à j. Cette méthode était originellement utilisée par Brier (1950) pour effectuer des prédictions météorologiques. La prédiction la plus juste sera celle dont le MSE sera le plus faible, témoignant du plus faible taux d'erreurs au carré.

Appliqué à nos quatre différents modèles Elo basés sur les buts réels, il permettra donc de trouver la paire de paramètres K et h optimisant le MSE pour chacun d'entre eux.

Récapitulatif des quatre modèles :

- Elo 1 sera donc centré uniquement sur les résultats (victoire, défaite, match nul) lors de la compétition et aura une note initiale de 1000 pour toutes les équipes.
- Elo 2 tiendra compte des scores lors de chaque match avec également une note initiale de 1000.
- Elo 3 se basera sur les résultats de la compétition avec comme note initiale pour chaque équipe, la note qu'elle possède dans le nouveau classement FIFA.
- Elo 4 sera lui basé sur les scores lors de chaque match avec les notes initiales identiques à celles du nouveau classement FIFA.

3.2 *Elo basé sur les expected goals*

Valeur initiale r

Les différents modèles se basant sur les *expected goals* auront également deux variantes basées sur deux valeurs initiales, l'une fixée à 1000 et l'autre sur le classement FIFA tel que décrit précédemment.

Score du match S

Concernant le score du match S, nous allons également utiliser les deux mêmes mesures que sont les résultats et les scores des matchs, mais en se basant sur les *expected goals* lors des rencontres et non les buts réels.
Le système d'évaluation n'étant pas basé sur le résultat du match tel qu'on le connait il est plus compliqué de définir un vainqueur clair en se basant sur les *expected goals*.
Worville (2016) dans son application des *expected goals* à l'Elo avait décidé arbitrairement que pour qu'une équipe soit désignée vainqueur d'une rencontre il fallait qu'elle ait un avantage d'au moins 0,8 *expected goals* par rapport à son adversaire. Nous avons ici préféré, comme décrit dans la partie théorique lorsque le nombre de buts ou de points n'étaient pas des nombres entiers, utiliser la loi de Poisson afin d'obtenir une distribution des *expected goals* de chaque équipe en probabilités de buts. Les paramètres de la loi de Poisson utilisés seront donc la somme des *expected goals* allouées à chacune des deux équipes lors de chaque rencontre. Ensuite, en croisant les probabilités de buts des deux équipes nous allons pouvoir déterminer une probabilité pour chaque score et donc chaque résultat.

On peut voir en Table 3 la distribution des buts selon les *expected goals* du match entre le Brésil et la Belgique. La somme des *expected goals* dans ce match était de 2,5 pour le Brésil contre 0,5 pour la Belgique, ce qui nous donne les probabilités ci-dessous de résultats. Bien que la rencontre ait été gagnée par la Belgique 2 buts à 1, il en ressort qu'au vu des occasions créées par les deux équipes, le Brésil aurait dû l'emporter.

		Belgique									
		0	1	2	3	4	5	6	7	8	9
Brésil	0	0,050	0,025	0,006	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,124	0,062	0,016	0,003	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,156	0,078	0,019	0,003	0,000	0,000	0,000	0,000	0,000	0,000
	3	0,130	0,065	0,016	0,003	0,000	0,000	0,000	0,000	0,000	0,000
	4	0,081	0,041	0,010	0,002	0,000	0,000	0,000	0,000	0,000	0,000
	5	0,041	0,020	0,005	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	6	0,017	0,008	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	7	0,006	0,003	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	8	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	9	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000

TABLE 3 – Distribution de Poisson des *expected goals* lors de Brésil-Belgique

Victoire du Brésil = 81%
Egalité = 13%
Victoire de la Belgique = 6%

Les probabilités de résultat sont ensuite multipliées par un coefficient déjà observé en (2) où la victoire est égale à 1, le match nul à 0,5 et la défaite à 0. Nous obtenons donc un S pour chaque équipe respectant la propriété des sommes S égales à 1.
Concernant la deuxième variante de modèles d'*expected goals*, basée sur le score, elle sera calculée de deux façons différentes. La première méthode suivra également la distribution de poisson expliquée ci-dessus, mais chaque probabilité de score sera ensuite multipliée par sa valeur correspondante dans le tableau (Table 4) ci-dessous.

Buts équipe à domicile \ Buts équipe à l'extérieur	0	1	2	3	4	5	6	7	8	9
0	0,500	0,333	0,250	0,200	0,167	0,143	0,125	0,111	0,100	0,091
1	0,667	0,500	0,400	0,333	0,286	0,250	0,222	0,200	0,182	0,167
2	0,750	0,600	0,500	0,429	0,375	0,333	0,300	0,273	0,250	0,231
3	0,800	0,667	0,571	0,500	0,444	0,400	0,364	0,333	0,308	0,286
4	0,833	0,714	0,625	0,556	0,500	0,455	0,417	0,385	0,357	0,333
5	0,857	0,750	0,667	0,600	0,545	0,500	0,462	0,429	0,400	0,375
6	0,875	0,778	0,700	0,636	0,583	0,538	0,500	0,467	0,438	0,412
7	0,889	0,800	0,727	0,667	0,615	0,571	0,533	0,500	0,471	0,444
8	0,900	0,818	0,75	0,692	0,642	0,600	0,563	0,529	0,500	0,474
9	0,909	0,833	0,769	0,714	0,667	0,625	0,588	0,556	0,526	0,500

TABLE 4 – Valeurs de S selon la formule de différence de buts

Ce tableau est construit selon la formule déjà connue de différence de buts (3), où chaque cellule de la matrice est calculée selon cette formule avec G_{ij} le nombre de buts inscrits par l'équipe à domicile et G_{ji} par celle à l'extérieur.

La valeur du S sera donc calculée selon le produit d'Hadamard, c'est-à-dire en multipliant les deux matrices composante par composante. Cela signifie que la matrice de probabilité P de composante p_{ij} multipliée par la matrice A de différence de buts de composante a_{ij} nous donne le S_i de l'équipe à domicile lors du match m :

$$S_i(m) = \sum_{i,j=0}^{9} p_{ij}.a_{ij} \tag{42}$$

La deuxième méthode est, elle, plus simple. Il suffit de remplacer les buts marqués dans la formule des différences de buts directement par les valeurs d'*expected goals*, ce qui nous donne une nouvelle formule de S pour l'équipe i contre l'équipe j telle que

$$S_{ij} = \frac{xG_{ij} + 1}{xG_{ij} + xG_{ji} + 2} \tag{43}$$

avec xG_{ij} le nombre d'*expected goals* pour l'équipe i lors du match contre l'équipe j et inversement pour xG_{ji}.

Score attendu μ

Contrairement au score du match, le score attendu suit les mêmes équations que dans le modèle basé sur les résultats.

Paramétrage de K *et* h

Les valeurs des paramètres h et K seront calculées de la même manière que dans les modèles Elo basés sur le résultat et les buts marqués, en les optimisant par rapport au MSE.

Récapitulatif des six modèles :

- Elo 5 sera donc centré uniquement sur les résultats - après distribution de Poisson des *expected goals* - lors de la compétition et aura une note initiale de 1000 pour toutes les équipes.
- Elo 6 sera basé sur les *expected goals* distribués selon la formule de différence de buts et avec une note initiale de 1000.
- Elo 6 bis est une variante de l'Elo 6 car il tiendra compte également des *expected goals* mais ils suivront une distribution de Poisson lors de chaque match avec également une note initiale de 1000.
- Elo 7, 8 et 8 bis seront identiques aux trois premiers à la seule différence des notes initiales des équipes qui seront basées sur le nouveau classement FIFA.

3.3 *Analyse des différences de performance*

Après avoir développé et optimisé l'ensemble des modèles, nous obtiendrons un modèle de classement évaluant au mieux les équipes selon leurs résultats ou scores réels, et un autre évaluant au mieux les équipes selon leur capacité à se créer des occasions et à les défendre. Ces modèles seront ceux ayant la plus faible valeur de MSE.

Ces classements étant différents, car basés sur deux indicateurs distincts, ils présenteront très certainement des différences d'évaluation entre les équipes. Ainsi, certaines équipes se verront classées plus haut selon leurs résultats que par rapport aux occasions qu'elles se seront créées.
Il sera donc intéressant d'essayer de comprendre et d'expliquer les différences entre ces systèmes d'évaluations. Corréler ces différences avec des statistiques liées aux

équipes pourraient notamment donner des indications sur des possibles causes de ces surperformances ou sous-performances.

3.4 *Capacité prédictive des expected goals*

Une autre question à laquelle nous souhaitons répondre est relative à la capacité prédictive des *expected goals*.
En utilisant la formule du MSE il serait ainsi possible de comparer un modèle d'évaluation se basant sur les *expected goals* par rapport à sa prédiction des résultats ou scores réels à venir.
Le score du match étant le paramètre avec les meilleures prédictions selon Lasek (2013) et Chen (2017), nous allons comparer plusieurs modèles sur les scores des rencontres pendant la Coupe du Monde. Les différents modèles d'Elo seront toujours calculés et mis à jour de la même façon. La seule différence sera le calcul de leur MSE. Avant chaque match nous allons tester la capacité prédictive de chaque classement en introduisant le μ correspondant à l'évaluation la valeur de chaque équipe afin de le comparer au S du score du match dont la formule (3), décrite en 1.1.2.1 est :

$$S_{ij} = \frac{G_{ij} + 1}{G_{ij} + G_{ji} + 2}$$

Ces valeurs seront ensuite calculées dans la formule suivante :

$$\begin{aligned} MSE(i,j) &= \frac{1}{M} \sum_{m=1}^{M} (S_{ijm} - \mu_{ijm})^2 + (S_{jim} - \mu_{jim})^2 \\ &= \frac{1}{M} \sum_{m=1}^{M} (S_{ijm} - \mu_{ijm})^2 + [(1 - S_{ijm}) - (1 - \mu_{ijm})]^2 \\ &= \frac{1}{M} \sum_{m=1}^{M} (S_{ijm} - \mu_{ijm})^2 + (\mu_{ijm} - S_{ijm})^2 \\ &= \frac{1}{M} \sum_{m=1}^{M} 2(S_{ijm} - \mu_{ijm})^2 \end{aligned}$$

Les différents modèles testés seront les modèles les plus prédictifs parmi ceux se basant sur le même indicateur de performance. Ces modèles seront également optimisés comme vu au point 3.1 avec la paire de paramètres K et h prédisant au mieux le score des rencontres. Ces modèles déjà définis dans la section précédente seront également comparés à un modèle de base, ne se basant pas sur l'Elo et dont le μ dans la formule ci-dessous, évaluant le niveau des équipes, restera fixe avec

une valeur de 0,5. Ainsi nous aurons un aperçu d'un MSE "de base", lorsque lors de chaque match, les équipes sont évaluées à un niveau identique.

Afin de définir si ces différences de prédicition et donc de MSE entre les modèles sont significatives, nous utiliserons le test de Diebold-Mariano. Ce test développé par Diebold et Mariano (2002) est un test d'hypothèse dont les hypothèses nulles peuvent être soit une non-différence de prédiction entre les modèles, soit une meilleure prédiction de l'un ou l'autre modèle sur les valeurs testées. Ce test a pour but de comparer les valeurs de prédiction de deux modèles différents par rapport à un ensemble de valeurs à prédire. Les prédictions peuvent être mesurées dans ce test selon plusieurs types d'erreur, c'est pourquoi nous allons comparer les différentes prédictions des modèles sur les erreurs au carré, mais également sur les erreurs absolues.

Ensuite nous allons comparer les modèles sur leur capacité à prédire le vainqueur des rencontres de la phase finale de la Coupe du Monde, comme ont pu le faire Wang et Vandebroek dans leur étude en 2013 sur le Championnat d'Europe de football 2012. Pour cela, nous utiliserons une méthodologie similaire à celle de Leitner et al. (2010) en prenant comme point de référence en termes de prédiction les entreprises de pari sportif qui seront représentées ici par un consensus de 12 *bookmakers*.

Le vainqueur prédit par chaque classement est désigné selon le terme du score attendu de l'Elo avant chaque match. Ainsi le μ est calculé avant chaque rencontre et l'équipe ayant le μ le plus élevé est désigné comme le vainqueur potentiel selon la prédiction du classement. Le vainqueur prédit selon le consensus de *bookmakers* est l'équipe ayant la cote moyenne la plus basse.

3.5 *Classement des joueurs*

Concernant le classement des joueurs nous allons également nous baser sur les différents classements Elo discutés précédemment et sur les notes des joueurs lors de chaque match.
Les notes proviennent du site spécialisé *Whoscored.com*, référence dans le domaine et utilisé notamment comme le système d'évaluation de référence dans l'étude de Brefeld et al. (2019) sur l'évaluation des joueurs dans le football européen. Le classement se base sur plus de 200 statistiques incluses dans le modèle d'évaluation,

influençant positivement ou négativement la note du joueur.

Notre classement de joueur sera défini à la manière du MVP (*Most Valuable Player*) dans les différents sports US dont les deux critères principaux sont le niveau du joueur, et son implication dans les résultats de son équipe.

Le niveau du joueur sera établi selon la moyenne des notes *Whoscored* d'un joueur durant la compétition. Il témoignera donc du niveau de jeu global qu'aura eu le joueur durant l'ensemble de la compétition.

A cela nous ajouterons l'influence respective sur les performances de leurs équipes en évaluant leur influence sur les différents classements Elo de leur équipe. Les classements Elo utilisés seront ceux ayant la meilleure capacité prédictive telle que définie plus haut.
Pour déterminer cette influence nous allons d'abord extraire les gains ou pertes de points Elo pour chaque équipe lors de chaque match. Ces gains et pertes sont définis dans la formule de mise à jour des notes Elo (1) dans le terme suivant :

$$K(S - \mu) \tag{44}$$

Ensuite nous allons allouer une partie des points à chaque joueur selon sa note lors du match correspondant. Afin de toujours valoriser une meilleure note et donc une meilleure performance, nous allons prendre l'inverse des notes lorsqu'une équipe perd des points. Ainsi, un joueur ayant obtenu une bonne note se verra pénaliser plus légèrement qu'un joueur ayant obtenu une mauvaise note lors d'une défaite ou d'un match nul contre une équipe inférieure.

L'addition de la note globale du joueur et de son apport dans les gains ou pertes de points nous donnera donc deux classements différents, l'un basé sur les buts réels et l'autre sur les *expected goals*.

Troisième partie
Partie empirique

1 jeu de données

Les données sur lesquelles nous allons tester nos différents modèles sont les 64 matchs disputés lors de la Coupe du Monde 2018 de football en Russie. Les données que nous possédons pour l'ensemble de ces matchs sont les dates des rencontres, leur lieu, le score du match ainsi que les *expected goals* pour chaque équipe.
La compétition qui s'est déroulée du 14 juin au 15 juillet 2018 regroupait 32 des meilleures équipes mondiales dont 31 se sont qualifiées à la suite de phases qualificatives dans leur confédération respective. Ces 31 équipes sont accompagnées du pays hôte, la Russie, automatiquement qualifiée (voir Table 5).

Qualification et tirage au sort

La FIFA décide du nombre de places qualificatives allouées à chaque confédération suivant un vote lors de la réunion de son Comité Exécutif. Le nombre de places allouées à chaque confédération est donc resté identique à celui de 2014 suite au vote.
Elles suivent la distribution suivante : 5 équipes pour l'AFC (confédération asiatique), 5 équipes pour la CAF (confédération africaine), 3 équipes pour la CONCACAF (confédération d'Amérique du Nord, d'Amérique centrale et des Caraïbes), 5 équipes pour la CONMEBOL (confédération sud-américaine) et 13 équipes pour l'UEFA (confédération européenne).

Les 32 équipes sont ensuite placées dans 4 « pots » de 8 équipes déterminés selon le classement FIFA de chaque équipe, le pot 1 regroupant les 7 meilleures équipes au classement ainsi que le pays hôte, et le pot 4, les 8 équipes les moins bien classées au classement FIFA. Elles sont ensuite tirées au sort en 8 groupes de 4 équipes, composés d'une équipe de chaque pot.

32 équipes qualifiées	
Pays	**Confédération**
Arabie Saoudite	AFC
Australie	AFC
Corée du Sud	AFC
Iran	AFC
Japon	AFC
Egypte	CAF
Maroc	CAF
Nigéria	CAF
Sénégal	CAF
Tunisie	CAF
Costa Rica	CONCACAF
Mexique	CONCACAF
Panama	CONCACAF
Argentine	CONMEBOL
Brésil	CONMEBOL
Colombie	CONMEBOL
Pérou	CONMEBOL
Uruguay	CONMEBOL
Allemagne	UEFA
Angleterre	UEFA
Belgique	UEFA
Croatie	UEFA
Danemark	UEFA
Espagne	UEFA
France	UEFA
Islande	UEFA
Pologne	UEFA
Portugal	UEFA
Serbie	UEFA
Suède	UEFA
Suisse	UEFA
Russie	*Host Nation*

TABLE 5 – Equipes qualifiées à la Coupe du Monde 2018

Phase de groupe

Les groupes, allant de A à H, sont donc composés de 4 équipes qui vont s'affronter une fois chacune à la manière d'un mini-tournoi en round-robin. La phase de groupe comprend donc un total de 48 matchs. Les matchs sont joués avec un temps réglementaire de 90 minutes, sans prolongations, ni tirs au but, ce qui signifie qu'une égalité est possible durant cette phase de compétition. Les victoires rapportent 3 points, les égalités 1 point et les défaites ne rapportent aucun point aux équipes.

Les critères afin de déterminer les 2 équipes se qualifiant pour la phase finale de compétition sont au nombre de 8 (Fifa.com, 2018). Le premier est assez logiquement le nombre de points engrangés par les équipes, suivent la différence de buts, le nombre de buts marqués, les résultats en face-à-face, la différence de buts en face-à-face, le nombre de buts marqués en face-à-face, le fair-play déterminé selon le nombre de cartons reçus et finalement un tirage au sort.
Le critère le plus avancé utilisé dans cette édition de la Coupe du Monde est le 7ème et avant-dernier, le fair-play. Le Japon a ainsi devancé le Sénégal à la deuxième place du groupe H grâce à son nombre de cartes jaunes (4) inférieur à celui des sénégalais (6).

Les classements finaux des 8 groupes sont les suivants :

Rang	**Groupe A**	**Groupe B**	**Groupe C**	**Groupe D**
1	*Uruguay*	*Espagne*	*France*	*Croatie*
2	*Russie*	*Portugal*	*Danemark*	*Argentine*
3	Arabie Saoudite	Iran	Pérou	Nigéria
4	Egypte	Maroc	Australie	Islande
Rang	**Groupe E**	**Groupe F**	**Groupe G**	**Groupe H**
1	*Brésil*	*Suède*	*Belgique*	*Colombie*
2	*Suisse*	*Mexique*	*Angleterre*	*Japon*
3	Serbie	Corée du Sud	Tunisie	Sénégal
4	Costa Rica	Allemagne	Panama	Pologne

TABLE 6 – Classement de la phase de groupe

Les 16 équipes qualifiées pour la phase suivante de compétition et les huitièmes de finale sont les deux premières équipes de chaque groupe dans les classements ci-dessus (Table 6).

Phase finale

La phase finale de la compétition se déroule en matchs à élimination directe, en allant des huitièmes de finale, en passant par les quarts et les demies jusqu'en finale. Les vainqueurs de chaque rencontre sont désormais décidés par 90 minutes de temps réglementaire, suivies de 30 minutes de prolongations en cas d'égalité, et de tirs au but si nécessaire. L'entièreté de la phase finale sera décidée selon cette formule. Cette phase compte au total 16 rencontres.

Les équipes qualifiées pour les huitièmes de finale s'affrontent de façon croisée, le 1er du groupe A affronte le 2ème du groupe B et inversement, le 1er du groupe C affronte le 2ème du groupe D, etc. (voir Table 7).

Huitièmes de finale				
Match	**Label**	**Equipe A**	**Equipe B**	**Label**
I	1A	*Uruguay*	Portugal	2B
J	1B	Espagne	*Russie*	2A
K	1C	*France*	Argentine	2D
L	1D	*Croatie*	Danemark	2C
M	1E	*Brésil*	Mexique	2F
N	1F	*Suède*	Suisse	2E
O	1G	*Belgique*	Japon	2H
P	1H	*Colombie*	Angleterre	2G

TABLE 7 – Rencontres des huitièmes de finale

Les quarts de finale (voir Table 8) regrouperont les huit vainqueurs des huitièmes de finale, ici en italique, qui s'affronteront à nouveau en croisant les groupes, avec le vainqueur du huitième I affrontant le vainqueur de K, et le vainqueur de J contre le vainqueur de L.

Quarts de finale				
Match	**Label**	**Equipe A**	**Equipe B**	**Label**
Q	I	Uruguay	*France*	K
R	J	Russie	*Croatie*	L
S	M	Brésil	*Belgique*	O
T	N	Suède	*Angleterre*	P

TABLE 8 – Rencontres des quarts de finale

Les demi-finales sont l'occasion de se voir croiser les groupes A, B, C, D et E, F, G, H pour la première fois dans la compétition. Les deux affiches seront France-Belgique et Croatie-Angleterre (voir Table 9).

Demi-finales			
Label	**Equipe A**	**Equipe B**	**Label**
Q	*France*	Belgique	S
R	*Croatie*	Angleterre	T

TABLE 9 – Rencontres des demi-finales

A l'issue des demi-finales nous connaissons donc les deux équipes qui accèderont à la grande finale, les deux autres se disputeront quant à elles la 3ème place.

Finale	
Equipe A	**Equipe B**
France	Croatie
Petite finale	
Equipe A	**Equipe B**
Belgique	Angleterre

TABLE 10 – Finale et rencontre pour la 3ème place

La France est donc couronnée championne du monde à la suite de sa victoire sur la Croatie, qui prendra la deuxième place. La Belgique complètera le podium grâce à sa victoire sur l'Angleterre (voir Table 10).

2 application des modèles elo

2.1 *Evaluation des équipes selon leurs buts réels*

Comme expliqué auparavant, les quatre modèles d'Elo basés sur les résultats et scores réels ont été optimisé de façon à obtenir le MSE le plus faible possible. Pour cela il a fallu chercher le h optimal pour une certaine valeur de K et inversement jusqu'à obtenir la paire optimale. Cette paire est également celle ayant la plus faible valeur de MSE.
Afin de commencer nos recherches pour l'optimisation des différentes paires optimales des modèles, il a fallu déterminer la valeur d'un des deux paramètres de départ, K et h. Il a été choisi arbitrairement de prendre K et la valeur, fixée à 60, est également celle utilisée par la FIFA et le site *Eloratings.net* dans leurs modèles d'Elo.

Les détails d'optimisation de chacun des modèles se trouve en annexes, les résultats finaux des modèles optimisés sont les suivants :

Paramètres	Elo 1	Elo 2	Elo 3	Elo 4
r_{init}	1000	1000	FIFA	FIFA
S (réel)	résultat	score	résultat	score
K	76,99	54,88	43,07	92,69
h	0,00	31,41	0,00	0,00
MSE	0,3861	**0,0588**	0,3457	0,0656

TABLE 11 – Valeurs de MSE des classements Elo optimisés basés sur les buts réels

On constate premièrement que les valeurs de K optimales pour les différents modèles ne sont pas éloignées outre-mesure de la valeur initiale fixée à 60, ce qui signifie que le niveau de compétition est très proche de celui utilisé par les systèmes d'évaluations comme celui de la FIFA ou d'*Eloratings.net*.
Concernant l'avantage à domicile et h, pour trois modèles sur quatre sa valeur peut être considérée comme nulle, et le dernier modèle possède un h relativement faible, aux alentours de 30. La Russie n'aurait donc pas bénéficié d'un avantage significatif sur ses résultats, étant donné que la valeur de h est régulièrement définie à 100, synonyme d'un avantage pour l'équipe à domicile de gagner 64% des points mis en jeu.

Concernant les différences de MSE entre les modèles, la première remarque est la différence entre les deux modèles basés sur les résultats -victoire, égalité, défaite-

et ceux sur les scores des rencontres. Il est en effet beaucoup plus facile pour l'Elo de prédire un résultat se basant sur les scores plutôt que sur les résultats, comme cela avait déjà été démontré par Lasek (2013) et Chen (2017).
L'utilisation d'un historique de notes d'équipes, ici le classement FIFA, n'a pas forcément amélioré la capacité prédictive des modèles. En effet, il améliore légèrement le modèle basé sur les résultats, mais est moins prédictif qu'une valeur initiale fixe pour les scores des matchs. On ne peut donc pas en déduire qu'il rend les modèles plus prédictifs.

Nous pouvons désormais répondre à la première partie de notre première question de recherche qui était « Quels sont les modèles d'Elo se basant sur les buts réels ainsi que sur les *expected goals* les plus performants ? ». le modèle le plus performant sur les buts réels et ayant le MSE le plus faible est donc l'Elo 2, un modèle basé sur les scores avec une valeur initiale fixe pour toutes les équipes. Il est également le seul à avoir une valeur de h éloignée de 0, avec un K très proche de celui défini par la FIFA. Il est donc le modèle développé ici prédisant au mieux les buts réels des matchs de cette Coupe du Monde.

Le classement des équipes à l'issue de la Coupe du Monde 2018 suivant ce modèle est défini ci-dessous.
On y retrouve l'équipe championne du monde en tête, la France. La Belgique et la Croatie, respectivement troisième et deuxième du mondial sont également très bien classées. Ce classement étant basé sur les scores des matchs, il n'est pas illogique de retrouver l'Angleterre qui perd trois de ses sept matchs à la 10ème place, loin de sa quatrième place finale lors de la compétition.
Ainsi, nous avons un aperçu avec ce classement des valeurs des équipes à la fin de la compétition en se basant sur les buts réels.

Classement Elo 2

Rang	Pays	Note
1	France	1046,37
2	Belgique	1042,48
3	Brésil	1031,49
4	Croatie	1028,29
5	Uruguay	1023,23
6	Colombie	1018,94
7	Suède	1012,44
8	Danemark	1010,22
9	Espagne	1009,76
10	Angleterre	1007,76
11	Portugal	1004,41
12	Corée du Sud	1001,40
13	Russie	1000,14
14	Iran	1000,00
15	Pérou	997,59
16	Sénégal	996,77
17	Suisse	996,06
18	Nigéria	994,30
19	Japon	993,29
20	Tunisie	991,24
21	Argentine	990,06
22	Serbie	990,01
23	Pologne	989,36
24	Maroc	984,59
25	Mexique	984,51
26	Allemagne	983,20
27	Islande	983,16
28	Arabie Saoudite	982,11
29	Australie	980,94
30	Costa Rica	980,09
31	Egypte	979,91
32	Panama	965,88

TABLE 12 – Classement des équipes suivant Elo 2

2.2 *Evaluation des équipes selon leurs expected goals*

La méthode utilisée afin de déterminer les modèles ayant la meilleure capacité prédictive reste inchangée, cependant ici les modèles ne seront pas jugés sur leur capacité à prédire les résultats ou scores réels des équipes mais leurs résultats ou scores selon les *expected goals*. Ainsi, plus un modèle aura un MSE faible, au mieux il prédira les *expected goals* à venir.

Paramètres	Elo 5	Elo 6	Elo 6 bis	Elo 7	Elo 8	Elo 8 bis
r_{init}	1000	1000	1000	FIFA	FIFA	FIFA
S (xG)	résultat*	score	score*	résultat*	score	score*
K	100,17	101,09	100,58	65,06	155,07	182,36
h	254,11	110,01	93,94	0,00	0,00	0,00
MSE	0,0997	**0,0273**	**0,0210**	0,0897	0,0358	0,0314

TABLE 13 – Valeurs de MSE des classements Elo optimisés basés sur les *expected goals*

Les valeurs de K sont ici plus élevées, aux alentours de 100, sans que cela soit pour autant surprenant. Les valeurs de h diffèrent elles significativement entre les modèles basés sur le classement FIFA pour les valeurs initiales, de ceux basés sur des valeurs fixes et identiques. En effet, si l'on prend en compte le classement FIFA, les valeurs de h sont presque nulles, ce qui signifierait que l'avantage d'évoluer à domicile était également nul en termes d'*expected goals* lors de la Coupe du Monde. Cependant, pour les modèles avec une valeur initiale identique, l'impact de l'avantage à domicile est très significatif, et encore plus pour le modèle basé sur les résultats après distribution, qui atteint une valeur supérieure à 250.

Comme pour l'Elo basé sur le résultat réel on peut observer que les deux modèles les moins prédictifs sont également ceux basés sur le résultat -victoire, égalité, défaite- après distribution de Poisson. Tout comme les valeurs initiales fixées à 1000 nous donnent des modèles plus prédictifs pour les modèles incluant les scores mais pas pour les résultats.

Pour répondre à la deuxième partie de notre première question de recherche qui était « Quels sont les modèles d'Elo se basant sur les buts réels ainsi que sur les *expected goals* les plus performants ? », nous avons obtenu deux modèles basés sur les *expected goals* ayant des valeurs de MSE beaucoup plus faibles que les autres. Ces deux modèles, les plus prédictifs, sont l'Elo 6 et l'Elo 6 bis. Ils sont tous deux

basés sur les *expected goals* sous forme de scores, avec un avantage pour celui utilisant la distribution de Poisson au préalable, défini comme score* dans la Table 13.

Concernant les classements des équipes, ils sont très similaires, à quelques exceptions près, telles que la Croatie ou la Tunisie. On remarque cependant des différences notoires avec le classement de l'Elo 2 basé sur les buts réels, ce qui indique un réel décalage entre les *expected goals* et les buts réellement marqués.
Les classements basés sur les *expected goals* sont très différents des classements basés sur les buts réels appliqués à cette Coupe du Monde 2018.

Classement Elo 6

Rang	Pays	Note
1	Brésil	1069,39
2	Uruguay	1048,00
3	France	1040,98
4	Croatie	1040,72
5	Espagne	1036,08
6	Belgique	1032,98
7	Suède	1018,57
8	Angleterre	1017,76
9	Portugal	1011,51
10	Sénégal	1008,24
11	Arabie Saoudite	1006,75
12	Australie	1005,96
13	Suisse	1004,17
14	Islande	1002,65
15	Allemagne	998,60
16	Nigéria	996,37
17	Pologne	993,75
18	Iran	991,73
19	Colombie	987,90
20	Egypte	987,02
21	Mexique	985,72
22	Japon	985,58
23	Argentine	985,13
24	Serbie	984,84
25	Tunisie	982,89
26	Maroc	982,72
27	Danemark	981,77
28	Pérou	979,45
29	Costa Rica	976,83
30	Corée du Sud	976,78
31	Panama	947,20
32	Russie	931,95

Classement Elo 6 bis

Rang	Pays	Note
1	Brésil	1061,40
2	Uruguay	1041,14
3	Croatie	1035,68
4	France	1035,65
5	Espagne	1031,94
6	Belgique	1027,11
7	Suède	1016,77
8	Angleterre	1015,18
9	Portugal	1009,79
10	Sénégal	1007,49
11	Arabie Saoudite	1006,60
12	Australie	1004,54
13	Suisse	1004,16
14	Islande	1002,37
15	Allemagne	998,81
16	Nigéria	996,95
17	Pologne	994,16
18	Iran	992,71
19	Colombie	989,01
20	Tunisie	988,51
21	Egypte	987,98
22	Mexique	986,71
23	Japon	986,40
24	Serbie	986,31
25	Argentine	986,20
26	Maroc	984,32
27	Danemark	984,25
28	Pérou	982,78
29	Costa Rica	979,68
30	Corée du Sud	979,64
31	Panama	953,96
32	Russie	941,81

TABLE 14 – Classement des équipes suivant Elo 6 et Elo 6 bis

3 analyse des performances

Afin d'obtenir le niveau de performance des équipes et ainsi déterminer quelles équipes auraient pu surperformer ou à l'inverse sous-performer durant la compétition, nous allons déduire du classement final basé sur les buts réels la valeur de chaque équipe dans le classement basé sur les *expected goals*. Nous obtenons le classement en Table 15.

Les équipes se situant au-dessus de zéro étant les équipes en surperformance, avec une valeur dans le classement basé sur les buts réels supérieure à leur valeur dans le classement des *expected goals*. Les équipes avec une différence inférieure à zéro sont à l'inverse les équipes avec une valeur supérieure dans le classement des *expected goals*.

Quelles sont donc les facteurs qui font qu'une équipe parvienne mieux à transformer ses occasions en buts marqués ? Ces facteurs ont-ils une réelle influence ou la chance est-elle, comme on a pu le lire précédemment, le seul facteur déterminant ?

Nous avons donc regroupé un certain nombre de variables pouvant influencer le parcours et les performances d'une équipe en Coupe du Monde :

- Age moyen des joueurs sélectionnés
- Nombre de participations à la Coupe du Monde du pays
- Joueurs évoluant à l'étranger
- Valeur marchande des joueurs
- Salaire du sélectionneur
- Temps passé par le sélectionneur à la tête de la sélection

Ces données proviennent du site *transfermarkt.com* et les salaires des sélectionneurs ont été récupérés du *site mirror.co.uk*.

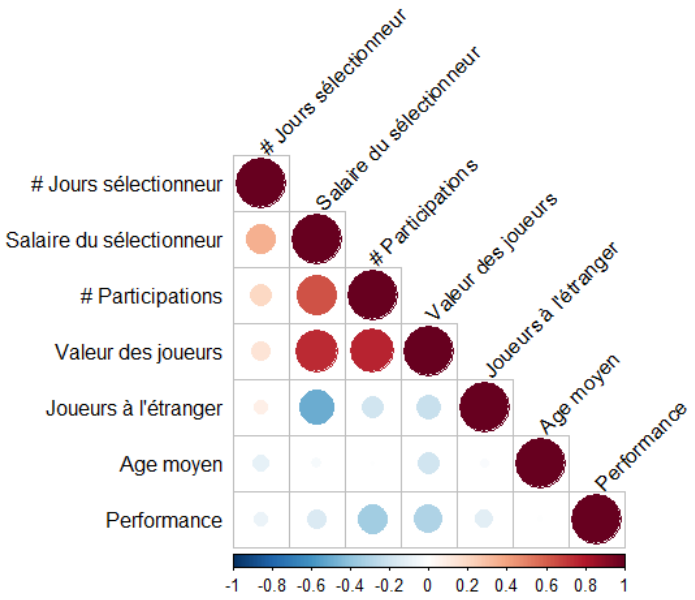
Classement de performance

Rang	Pays	Performance
1	Russie	68,193
2	Colombie	31,042
3	Danemark	28,444
4	Corée du Sud	24,612
5	Panama	18,677
6	Pérou	18,140
7	Belgique	9,495
8	Tunisie	8,343
9	Iran	8,274
10	Japon	7,699
11	France	5,396
12	Serbie	5,173
13	Argentine	4,938
14	Costa Rica	3,260
15	Maroc	1,874
16	Mexique	-1,214
17	Nigeria	-2,076
18	Pologne	-4,392
19	Suède	-6,124
20	Portugal	-7,097
21	Egypte	-7,114
22	Suisse	-8,111
23	Angleterre	-10,004
24	Sénégal	-11,466
25	Croatie	-12,422
26	Allemagne	-15,402
27	Islande	-19,488
28	Arabie Saoudite	-24,638
29	Uruguay	-24,767
30	Australie	-25,025
31	Espagne	-26,321
32	Brésil	-37,900

TABLE 15 – Performance des équipes

Les données sélectionnées peuvent influer sur les performances des différentes sélections, que ça soit de l'expérience de l'équipe, avec l'âge moyen et son nombre de participations à la Coupe du Monde, ou au contraire de son inexpérience. Le nombre de joueurs évoluant à l'étranger peut être synonyme de patriotisme s'il est faible ou à l'inverse d'une qualité des joueurs s'exportant très bien à l'étranger et donc très demandés. Les valeurs marchandes des joueurs ainsi que le salaire des sélectionneurs sont de bons indicateurs de leur niveau et de leur valeur intrinsèque. Finalement, le temps passé par le sélectionneur à la tête de sa sélection peut témoigner d'un travail effectué sur le long-terme ou alors de la fraicheur et des nouvelles idées d'un coach récemment nommé.

Afin de confirmer ou infirmer ces différentes hypothèses nous allons comparer la corrélation entre ces différentes valeurs pour chaque équipe avec leur indice de performance lors de la Coupe du Monde. Cela nous permettra d'avoir un aperçu de l'impact que ces facteurs pourraient avoir sur les performances des équipes et par la suite si des liens existeraient entre les différentes équipes ayant sous-performées et les autres ayant surperformées.

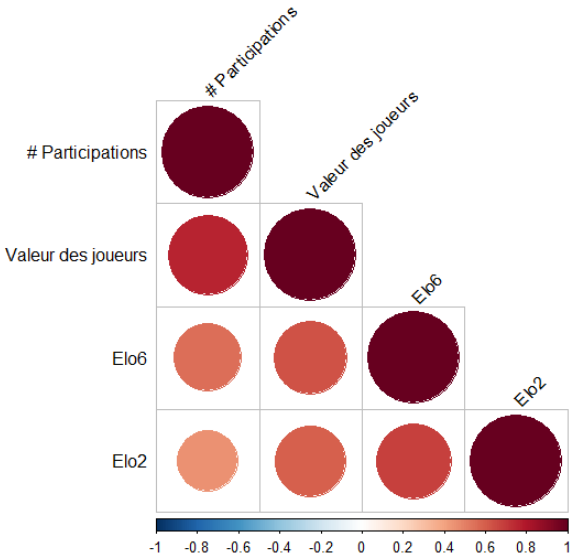


Corrélation Performance		
	r-value	p-value
Age moyen	-0,01	0,960
# Participations	-0,34	0,061
Joueurs à l'étranger	-0,13	0,490
valeur des joueurs	-0,31	0,081
Salaire du sélectionneur	-0,16	0,370
# Jours sélectionneur	-0,08	0,650

TABLE 16 – Corrélation entre performance et différentes statistiques d'équipe

Les corrélations entre notre index de performance et les autres facteurs ne sont pas très fortes, la r-valeur étant inférieure à 0,5 (voir Table 16). Néanmoins, deux des facteurs sont significatifs pour une p-valeur de 0,1 ou un indice de confiance de 90%. Ces valeurs sont les participations à la Coupe du Monde et les valeurs marchande des joueurs sélectionnés, toute deux faiblement corrélées négativement par rapport aux performances. Cela signifie que ces nations historiques avec des joueurs de valeur supérieure tendent à sous-performer légèrement dans leur capacité à transformer des occasions en buts.

Les corrélations entre ces deux variables et les valeurs d'Elo 6 et Elo 2 (voir Table 17) montrent également que ces équipes se procurent un plus grand nombre d'occasions, au vu de la relative forte corrélation entre Elo 6 et les participations ainsi que la valeur des équipes. Elles ont néanmoins une corrélation un peu plus faible avec le nombre de buts marqués, ce qui indique que les joueurs évalués assez haut et jouant dans des sélections historiquement fortes produisent un grand nombre d'occasions et en concèdent peu mais ne parviennent pas forcément à traduire cela en but marqués ou à en empêcher avec la même constance.



	Corrélation Elo 2		**Corrélation Elo 6**	
	r-value	p-value	r-value	p-value
# Participations	0,45	0,00980	0,55	0,00120
valeur des joueurs	0,59	0,00034	0,63	0,00012

TABLE 17 – Corrélation entre Elo 2, Elo 6 et différentes statistiques d'équipe

A l'inverse, on peut en conclure que les niveaux se lissent lors d'une compétition aussi importante et toutes les équipes, peu importe le nombre d'occasions qu'elles vont se créer vont jouer leur coup à fond. Les équipes considérées comme plus fortes ne vont pas concrétiser leurs nombreuses occasions là où des équipes plus faibles réussiront à tirer leur épingle du jeu malgré un nombre parfois inférieur d'occasions.

Cependant, à l'inverse de facteurs comme l'âge moyen de l'équipe ou le temps passé à la tête de sa sélection par le sélectionneur, ces deux variables ne sont pas forcément modifiables outre-mesure. Le nombre de participations aux précédentes éditions ne peut être changé et la valeur des joueurs est difficilement quantifiable. Un sélectionneur devrait-il donc prendre un joueur un peu moins coté au détriment d'un autre plus « bankable » sur le marché ? Pour quelles raisons ?

Il est donc très difficile de trouver des facteurs sur lesquels une équipe nationale pourrait travailler ou modifier afin d'améliorer cet aspect d'efficacité dans les rectangles. Cependant d'autres pistes pourraient être envisagées, telles que le nombre de matchs

où l'équipe a affiché les mêmes 11 joueurs au coup d'envoi, la qualité des joueurs aux postes clés tels que l'avant-centre ainsi que les défenseurs centraux et le gardien.

Une donnée intéressante concerne les sélectionneurs limogés dans l'année suivant la Coupe du Monde. D'un côté on aurait pu penser que les sélectionneurs dont les résultats n'ont pas été à la hauteur des attentes malgré de nombreuses occasions créées et peu d'occasions concédées auraient pu en faire les frais, et qu'à l'inverse, des bons résultats malgré un fond de jeu plus faible auraient pu en sauver plus d'un. Le résultat est qu'il n'y aucune corrélation entre départ ou limogeage du sélectionneur et les performances ainsi que l'efficacité de son équipe durant la compétition. Sur les 15 sélectionneurs démis de leurs fonctions, 9 ont même connu une surperformance face aux buts lors de la Coupe du Monde. Les résultats d'une équipe, même dans la plus grande compétition mondiale, ne peuvent donc plus protéger les hommes à leur tête.

Nous avons donc tenté de répondre à la deuxième de nos questions de recherche, qui est « Quelles sont les différences entre les évaluations des deux types de modèle (buts réels et *expected goals*) et comment peut-on les expliquer ? ».
Nous avons donc d'abord défini les différences d'évaluations entre les deux types de modèle dans la Table 15. Ensuite, Les différences constatées entre les classements basés sur les *expected goals* et les buts réels ont été testées sur leur corrélation avec des statistiques liées aux équipes, mais les écarts n'ont pas pu être suffisamment expliqués par ces différentes statistiques.

4 prédiction des scores et résultats futurs

Maintenant que nous connaissons les modèles prédisant au mieux les buts réels et les *expected goals* nous allons pouvoir les comparer entre eux sur leur faculté commune à prédire les résultats et scores à venir. En effet, nous avons obtenu les modèles Elo prédisant au mieux les *expected goals* à venir, mais sont-ils capables de prédire avec justesse les buts réels à venir ?

C'est ce que nous avons essayé de savoir en modifiant la formule de MSE (41 à 44) des modèles d'*expected goals* comme expliqué dans la section méthodologie. Cela nous permettra de comparer les performances prédictives des modèles d'*expected goals* sur les scores à venir par rapport à l'Elo 2 ayant la meilleure capacité prédictive sur les buts réels.

Nous avons donc comparé trois modèles, l'Elo 2, 6 et 6 bis, sur leur capacité à prédire les scores à venir lors de la compétition. Les modèles possèdent tous une même valeur initiale fixe et identique pour toutes les équipes et ils se basent sur la différence de buts ou d'*expected goals* suivant la même formule (3). La seule différence vient du modèle 6 bis, qui applique sur les valeurs d'*expected goals* une distribution de Poisson avant de les appliquer à la formule de différence de buts (3), là où l'Elo 2 et l'Elo 6 appliquent respectivement les buts marqués et les *expected goals* directement dans la formule.

Afin de mettre les différences de prédiction des modèles Elo en perspective, nous avons également ajouté un modèle de base qui est une valeur fixe de 0,5 pour u_{ij} et u_{ji} dans la formule de MSE (41 à 44). Cela signifie que dans ce modèle, toutes les équipes sont considérées égales en valeur avant chaque match de la compétition. Les résultats de MSE pour les prédictions des scores à venir sont les suivants :

Paramètres	Fixe	Elo 2	Elo 6	Elo 6 bis
r_{init}	-	1000	1000	1000
S calculé	-	score réel	score xG	score xG*
K	-	54,88	110,56	126,63
h	-	31,41	0,00	0,89
S comparé	score réel	score réel	score réel	score réel
MSE	0,0599	0,0588	**0,0580**	0,0581

TABLE 18 – Valeurs de MSE des différents classements Elo optimisés sur leur prédiction des scores réels

Nous observons donc que les deux modèles se basant uniquement sur les *expected goals* ont une meilleure prédiction des différences de buts à venir que le modèle basé sur les buts réels. Leurs valeurs de K et h sont plus ou moins identiques et celles de MSE également. Les différences de MSE entre les deux modèles d'*expected goals* ne sont pas très grandes, mais celui basé sur les buts réels est légèrement moins bon en termes de prédiction des buts à venir. La différence entre les *expected goals* et les buts réels est presque aussi grande que celle entre les buts réels et le modèle fixe où toutes les équipes possèdent la même valeur.

Afin de déterminer si ces différences sont significatives, nous avons effectué un test de Diebold-Mariano, dont le but est de comparer les prédictions de différents modèles. Appliqué à nos modèles d'*expected goals* et de buts réels en comparant leurs erreurs au carré ainsi que leurs erreurs absolues de prédiction, nous obtenons une meilleure capacité prédicitve pour le modèle d'*expected goals*. Cependant elle n'est pas suffisamment significative, la p-valeur n'étant que de 0,29 pour les erreurs au carré et de seulement 0,44 pour les erreurs absolues. Ce léger avantage prédictif est le même que celui du modèle basé sur les buts réels par rapport au modèle fixe où les équipes sont toutes considérées comme égales en valeur à chaque match, où la p-valeur est de 0,27 pour les erreurs au carré et de 0,37 pour les erreurs absolues. Face à ce modèle fixe, le modèle basé sur les *expected goals* est lui légèrement plus prédictif pour une p-valeur de 0,16 sur les erreurs au carrée mais un moins bon 0,42 sur les erreurs absolues.
On ne peut donc pas en déduire que les *expected goals* peuvent prédire correctement les scores à venir, étant donné qu'ils ne performent pas significativement mieux que les buts réels et surtout qu'un modèle fixe. Nous ne pouvons également pas affirmer avec certitude que le modèle basé sur les *expected goals*, bien que légèrement supérieur dans notre exemple, prédit plus efficacement les scores à venir que le modèle basé sur les buts réels.
Il n'est donc pas possible de répondre avec certitude à la première partie de notre dernière question de recherche qui est, « Un modèle se basant sur les *expected goals* peut-il prédire efficacement les scores et résultats futurs d'une compétition ? ».

Nous allons maintenant comparer ces modèles sur leur capacité à prédire le vainqueur uniquement de la rencontre à venir, lors de la phase finale de le Coupe du Monde. Le classement Elo 2 représentera les classements basés sur les résultats réels et le classement Elo 6, les *expected goals*. Le classement Elo 6 est ici préféré à l'Elo 6

bis car il est à la fois plus prédictif sur les résultats réels et est identique dans sa conception à l'Elo 2 si ce n'est qu'il est basé sur les *expected goals* et non sur les buts réels.
Ces modèles seront comparés, comme ont pu le faire Leitner et al. (2010) à un consensus de *bookmakers*, afin de déterminer la capacité prédictive des modèles face à une valeur établie dans le domaine de la prédiction de résultats.
Le format de phase finale nous permet notamment d'éviter les matchs nuls, car chaque rencontre se doit de déterminer le vainqueur qualifié pour le prochain tour. Ainsi les équipes avec un astérisque (*) sont celles ayant gagné après une séance de tirs au but. Nous pouvons donc comparer les prédictions des deux classements Elo et du consensus de *bookmakers* sur l'entièreté des rencontres de la phase finale (voir Table 19).

On constate que sur la prédiction des vainqueurs, le classement Elo 6 basé sur les *expected goals* performe à nouveau légèrement mieux que le classement Elo 2 basé sur les buts réels, et aussi bien que le consensus de *bookmakers*. Les différences majeures entre les deux classements Elo en termes de prédiction sont l'estimation des parcours russes et anglais.

On peut ici conclure que le modèle se basant sur les *expected goals* peut prédire efficacement les résultats futurs d'une compétition, étant donné qu'il performe aussi bien que notre référence prédictive que sont les *bookmakers*. Cela répond donc positivement à la deuxième partie de la question suivante, « Un modèle se basant sur les *expected goals* peut-il prédire efficacement les scores et résultats futurs d'une compétition ? ».

Prédiction phase finale

Equipe 1	Score équipe 1	Score équipe 2	Equipe 2	Elo 2	Elo 6	*Bookmakers*	Résultat réel
France	4	3	Argentine	France	France	France	France
Uruguay	2	1	Portugal	Uruguay	Uruguay	Portugal	Uruguay
Russie	1	1	Espagne	Russie	Espagne	Espagne	Russie*
Croatie	1	1	Danemark	Croatie	Croatie	Croatie	Croatie
Brésil	2	0	Mexique	Brésil	Brésil	Brésil	Brésil
Belgique	3	2	Japon	Belgique	Belgique	Belgique	Belgique
Suède	1	0	Suisse	Suède	Suède	Suisse	Suède
Colombie	1	1	Angleterre	Colombie	Angleterre	Angleterre	Angleterre*
Uruguay	0	2	France	Uruguay	Uruguay	France	France
Brésil	1	2	Belgique	Brésil	Belgique	Brésil	Belgique
Suède	0	2	Angleterre	Suède	Angleterre	Angleterre	Angleterre
Russie	2	2	Croatie	Russie	Croatie	Croatie	Croatie*
France	1	0	Belgique	Belgique	Belgique	France	France
Angleterre	1	2	Croatie	Croatie	Angleterre	Angleterre	Croatie
Belgique	2	0	Angleterre	Belgique	Belgique	Belgique	Belgique
France	4	2	Croatie	France	France	France	France
Pourcentage de prédictions correctes				63%	**69%**	**69%**	-

TABLE 19 – Comparaison des prédictions de l'Elo 2 et l'Elo 6 sur les résultats de la phase finale

5 classement des joueurs

Comme décrit dans la section méthodologie, nous allons tout d'abord récupérer les différentes notes des joueurs lors de la Coupe du Monde ainsi que les notes attribuées à leur performance collective.

	Classement Joueurs note moyenne		
Rang	Joueur	Pays	Note
1	Hazard	Belgique	8,52
2	Mina	Colombie	8,37
3	Isco	Espagne	8,30
4	Neymar	Brésil	8,26
5	Hadary	Egypte	8,20
6	Trippier	Angleterre	8,05
7	Messi	Argentine	7,88
8	Haddadi	Tunisie	7,80
9	Ndidi	Nigéria	7,77
10	Mbappé	France	7,76
11	Cavani	Uruguay	7,75
12	D. Costa	Brésil	7,75
13	Kane	Angleterre	7,75
14	Ronaldo	Portugal	7,75
15	Januzaj	Belgique	7,70
16	Kalinic	Croatie	7,70
17	Casemiro	Brésil	7,68
18	Schmeichel	Danemark	7,68
19	De Bruyne	Belgique	7,67
20	Modric	Croatie	7,61

TABLE 20 – Classement des joueurs selon leur note moyenne

Il a donc fallu récupérer les notes de chaque joueur lors de chaque match joué par celui-ci pour obtenir le classement ci-dessus (Table 20), avec ici le top 20. Ce classement prend en compte le niveau de jeu global de chaque joueur durant la compétition. Certains joueurs, tels que Hadary (Egypte), Haddadi (Tunisie), Januzaj (Belgique) ou encore Kalinic (Croatie) sont des anomalies avec seulement un match joué. Néanmoins, leur classement final sera tout de même impacté par ce faible temps de jeu étant donné le faible impact qu'ils ont pu avoir sur les résultats

collectifs de leur sélection.

Les classements des apports de chaque joueur sur les scores (Elo 2) ou sur les scores d'*expected goals* (Elo 6) de leur sélection sont les suivants :

	Classement Joueurs Elo 2			**Classement Joueurs Elo 6**		
Rang	Joueur	Pays	Note	Joueur	Pays	Note
1	Griezmann	France	3,77	Neymar	Brésil	5,58
2	Mbappé	France	3,68	Coutinho	Brésil	5,33
3	Pogba	France	3,65	Miranda	Brésil	5,07
4	Varane	France	3,54	T. Silva	Brésil	5,01
5	Pavard	France	3,44	Paulinho	Brésil	4,88
6	L. Hernández	France	3,39	Willian	Brésil	4,80
7	Kanté	France	3,36	Jesus	Brésil	4,79
8	Lloris	France	3,35	Fagner	Brésil	4,64
9	Umtiti	France	3,35	Alisson	Brésil	4,47
10	Giroud	France	3,33	Fernandinho	Brésil	4,40
11	Courtois	Belgique	3,21	Firmino	Brésil	4,31
12	E. Hazard	Belgique	3,16	Mandzukic	Croatie	4,17
13	Meunier	Belgique	3,14	Marcelo	Brésil	4,10
14	Cavani	Uruguay	3,02	Rebic	Croatie	4,00
15	Casemiro	Brésil	2,86	Vrsaljko	Croatie	3,99
16	De Bruyne	Belgique	2,78	Vida	Croatie	3,95
17	Neymar	Brésil	2,75	Godín	Uruguay	3,84
18	Tolisso	France	2,67	Mertens	Belgique	3,82
19	Mertens	Belgique	2,58	Brozovic	Croatie	3,79
20	Lukaku	Belgique	2,55	Subasic	Croatie	3,78

TABLE 21 – Classement des joueurs selon leur influence sur les classements Elo 2, Elo 6 de leur équipe

On peut constater que ces deux classements, avec ici leur top 20 (voir Table 21), sont très largement dominés par les joueurs évoluant dans les sélections en tête des classements d'équipe Elo 2 et Elo 6. Ainsi, on dénombre onze joueurs français parmi les vingt joueurs ayant eu le plus d'influence sur les résultats positifs de leur équipe et douze joueurs brésiliens parmi les vingt ayant eu le plus d'influence positive sur les *expected goals* crées et concédés par leur équipe.

En additionnant à ces deux classements la note moyenne de chaque joueur (voir Table 20), nous obtenons les deux classements finaux, regroupant à la fois leur apport collectif et leur note personnelle, ci-dessous :

	Classement final Joueurs Elo 2			**Classement final Joueurs Elo 6**		
Rang	Joueur	Pays	Note	Joueur	Pays	Note
1	Hazard	Belgique	11,68	Neymar	Brésil	13,84
2	Mbappé	France	11,44	Coutinho	Brésil	12,93
3	Griezmann	France	11,29	T. Silva	Brésil	12,31
4	Pogba	France	11,14	Miranda	Brésil	12,23
5	Neymar	Brésil	11,01	Hazard	Belgique	11,96
6	Cavani	Uruguay	10,77	Willian	Brésil	11,74
7	Varane	France	10,68	Jesus	Brésil	11,69
8	Pavard	France	10,54	Mandzukic	Croatie	11,68
9	Casemiro	Brésil	10,53	Paulinho	Brésil	11,68
10	Meunier	Belgique	10,52	Fagner	Brésil	11,67
11	De Bruyne	Belgique	10,45	Mbappé	France	11,44
12	Mina	Colombie	10,42	Cavani	Uruguay	11,44
13	Kanté	France	10,40	Isco	Espagne	11,43
14	L. Hernández	France	10,38	Casemiro	Brésil	11,41
15	Courtois	Belgique	10,34	D. Costa	Brésil	11,38
16	Giroud	France	10,13	Marcelo	Brésil	11,28
17	Umtiti	France	10,13	Rebic	Croatie	11,15
18	Coutinho	Brésil	10,07	Godin	Uruguay	11,14
19	Lloris	France	9,97	Vida	Croatie	11,08
20	Modric	Croatie	9,91	Vrsaljko	Croatie	11,02

TABLE 22 – Classement final des joueurs selon leur note moyenne et leur influence sur les classements Elo 2, Elo 6 de leur équipe

La prépondérance de joueurs français et brésiliens dans les deux classements (Table 22) reste d'actualité mais les dix premières places ne leur sont plus réservées. Dans le classement final d'Elo 2 basé sur les résultats, le belge Eden Hazard prend même la première place du classement. Griezmann qui était auparavant en tête se fait doubler par son compatriote Mbappé. Du côté de l'Elo 6 et des *expected goals*, l'écart des autres nations avec la sélection brésilienne se fait toujours ressentir. Neymar déjà cinquième dans le classement d'Elo 2 survole celui de l'Elo 6, suivi par trois de ses

coéquipiers, et Hazard qui se trouve également bien classé, prend la cinquième place.

Fait intéressant, le prix de meilleur joueur du tournoi a été attribué par la FIFA au croate Luka Modric, qui n'est que 20ème dans notre classement basé sur le résultat et 22ème dans celui basé sur les *expected goals*. Il paye notamment ses prestations en demi-teinte en demi-finale contre l'Angleterre (note de 6,7) et en finale contre la France (6,8) dans les deux classements. La FIFA avait déjà récompensé le joueur phare de l'équipe finaliste perdante en 2014, le capitaine argentin Lionel Messi. Elle a réitéré la chose ici, en donnant ce titre à Modric, la star croate, lui aussi capitaine de sa sélection et défait en finale.

Du côté des gardiens de but, c'est le belge Thibaut Courtois qui a obtenu les faveurs de la FIFA en se voyant remettre le *Golden Glove* de meilleur gardien du tournoi. Ce choix est un peu moins contestable par rapport à nos classements, étant donné qu'il occupe la place de premier gardien de but dans l'Elo 2 et malgré une quatrième place dans l'Elo 6, il est le plus performant en moyenne sur les deux classements.

6 conclusion, limites et travail supplémentaire

6.1 *Conclusion*

Ce mémoire nous a permis de développer deux modèles optimaux de classements Elo. Le premier, basé sur les buts réels et plus particulièrement le score des rencontres, est optimal pour un K de 54, ce qui est assez similaire à celui utilisé par la FIFA qui varie entre 50 et 60 pour les matchs de Coupe du Monde. L'avantage à domicile, h d'une valeur de 30, est tangible bien que relativement peu élevé. Le classement final de ce modèle représente de façon assez juste le déroulement de la compétition en termes de résultats. Nous avons donc répondu à la première partie de la première question de recherche qui est, « Quels sont les modèles d'Elo se basant sur les buts réels ainsi que sur les *expected goals* les plus performants ? ».

Concernant le jeu et les occasions créées, deux classements optimisés se démarquaient et tous deux se basaient sur la formule de score en y incluant les *expected goals* à la place des buts marqués. Le modèle utilisant la distribution de Poisson au préalable, l'Elo 6 bis, possède néanmoins un MSE légèrement plus faible que l'Elo 6. Cependant, lorsqu'on a testé ces deux modèles sur les scores à venir de la Coupe du Monde, ils obtenaient des valeurs de MSE identiques. Ces deux modèles répondent, eux, à la seconde partie de notre première question de recherche. Les valeurs de K pour ses deux classements étaient très élevées avec 155 et 182 respectivement ce qui témoigne d'une volatilité importante du classement, et ne comportaient pas d'avantage pour la Russie, l'équipe à domicile. Les deux classements finaux étaient très éloignés du déroulement de la compétition, avec trois des cinq équipes les mieux classées n'ayant même pas dépassé le stade des quarts de finale, dont un Brésil ultra-dominateur dans ces deux classements.

Les classements basés sur les buts réels et les *expected goals* sont très différentes les uns des autres.
Les différences entre les deux types de classement ont donc permis de mettre une valeur sur la réussite ou non face au but de chacune des équipes. Certaines équipes ont montré une réussite certaine, à l'image de la Russie qui possède 68 points de plus dans le classement des résultats. A l'inverse, le Brésil est l'équipe ayant pâti le plus de son manque de réussite devant les cages avec une différence négative de 37 points entre les deux classements.

La conversion des occasions en buts peut donc avoir un impact très important sur le classement d'une équipe.

Afin d'expliquer, au moins partiellement, cette différence entre les classements, sa corrélation avec des statistiques liées aux différentes équipes a été testée. Les seules corrélations significatives, bien que relativement faibles, liaient les performances dans la conversion d'occasions en buts négativement avec le nombre de participations du pays ainsi que la valeur marchande des joueurs. Les équipes considérées comme historiques et plus talentueuses étaient donc plus souvent moins réalistes devant le but durant la compétition, malgré un nombre plus élevé d'occasions de buts, comme le démontre la corrélation entre les *expected goals* et ces deux statistiques.
Cependant, il n'est toujours pas possible d'établir de liens significatifs et forts entre les différences de performance des équipes dans la conversion d'occasions avec d'autres variables. Nous n'avons donc pas pu répondre totalement à notre deuxième question de recherche qui est, « Quelles sont les différences entre les évaluations des deux types de modèle (buts réels et *expected goals*) et comment peut-on les expliquer ? ».

Les *expected goals* nous ont donc permis d'obtenir un nouvel indicateur d'évaluation d'équipe qui ne prend plus en compte le facteur du réalisme face au but. Cette valeur est, qui plus est, plus prédictive sur les résultats à venir d'une compétition que les buts réels, et aussi prédictive qu'un ensemble de *bookmakers*. Elle n'est cependant pas significativement supérieure aux buts réels sur les scores à venir, bien que légèrement plus prédictive appliquée aux matchs de la Coupe du Monde 2018. Il est donc très intéressant de voir que, bien que les *expected goals* produisent des classements assez différents des résultats et scores basés sur les buts réels, leur évaluation n'en est pas moins pertinente. Cela répond donc globalement positivement à notre dernière question de recherche qui est, « Un modèle se basant sur les *expected goals* peut-il prédire efficacement les scores et résultats futurs d'une compétition ? ».

6.2 *Limites et travail supplémentaire*

Les principales limites dans ce mémoire étaient liées à la faible quantité de données relatives aux *expected goals*. La valeur initiale des équipes au départ de la compétition ne pouvait ainsi pas être basée sur les *expected goals* étant donné qu'aucune

base de données n'était disponible sur les *expected goals* des rencontres avant la Coupe du Monde pour ces équipes. Concernant les données d'*expected goals* lors de la compétition, seul le détail du modèle ainsi que les valeurs d'*expected goals* de Michael Caley étaient disponible librement. Bien que son modèle soit reconnu par de nombreux pairs, une plus grande variété de modèles d'*expected goals* aurait permis de les comparer entre eux et de définir le plus adapté à notre étude.

Les *expected goals* utilisés lors de ce mémoire étant directement agrégés par équipe, il n'a pas été possible d'obtenir les détails relatifs aux joueurs impliqués dans ces occasions de but. Il serait donc intéressant de connaitre l'historique complet de conversion d'occasion en buts d'un joueur depuis un certain laps de temps, et plus spécifiquement pour les attaquants. Cette position étant mise à contribution régulièrement dans la finition d'une offensive, la corrélation entre le taux de conversion de l'attaquant ou de la ligne d'attaque d'une sélection et les performances de l'équipe face au but lors d'une compétition pourrait être significative.

Une autre limite est le format d'une Coupe du Monde qui comporte un nombre assez faible de rencontres. 64 matchs dont 7 au maximum par équipe, c'est en effet un échantillon assez faible pour tirer des conclusions définitives sur les différentes comparaisons et déductions développées. Réitérer ce travail sur d'autres compétitions permettrait ainsi d'avoir une plus grande fiabilité sur les différents enseignements tirés au cours de ce mémoire.

Quatrième partie
Annexes

a *Equations des différents types de tir selon le modèle de Caley*

— $RegularShots : (-3.19 - 0.095 * distance + 3.18 * inverse.distance + 1.88 * relative.angle + 0.24 * inverse.angle - 2.09 * inverse.dist * angle + 0.45 * throughball.assist + 0.64 * throughball.2nd.assist + 0.31 * assist.across.face - 0.15 * cutback.assist + 2.18 * inverse.assist.distance + 0.12 * assist.angle + 0.23 * fast.break + 0.18 * counterattack + 0.09 * established.possession - 0.18 * following.corner + 1.2 * big.chance + 1.1 * following.error + 0.39 * following.dribble + 0.14 * dribble.distance + 0.37 * rebound + 0.03 * game.state + 0.07 * Bundesliga - 0.1 * EPL - 0.09 * LaLiga - 0.07 * SerieA)$

— $HeadedShotsAssistedbyCrosses : (-2.88 - 0.21 * distance + 2.13 * relative.angle + 4.31 * inverse.assist.distance + 0.46 * assist.angle + 0.2 * fastbreak + 0.11 * counterattack + 0.12 * set.play - 0.24 * corner - 0.18 * otherbodypart + 1.2 * big.chance + 1.1 * following.error + 0.18 * EPL + 0.15 * LaLiga)$

— $Non - HeadedShotsAssistedbyCrosses : (-2.8 - 0.11 * distance + 3.52 * inverse.distance + 1.14 * angle + 0.14 * assist.across.face + 6.94 * inverse.assist.distance + 0.59 * assist.angle - 0.12 * corner + 0.24 * fastbreak + 0.11 * counterattack + 1.25 * big.chance + 1.1 * following.error - 0.2 * EPL)$

— $HeadedShotsNotAssistedbyCrosses : (-3.85 - 0.1 * distance + 2.56 * inverse.distance + 1.94 * relative.angle + 0.51 * throughball.assist + 0.44 * fastbreak + 0.26 * counterattack + 0.7 * rebound + 0.44 * established.possession + 1.14 * otherbodypart + 1.3 * big.chance + 1.1 * following.error - 0.29 * EPL - 0.24 * LaLiga - 0.26 * SerieA)$

— $ShotsfromDirectFreeKicks : (-3.84 - 0.1 * distance + 98.7 * inverse.distance + 3.54 * inverse.angle - 91.1 * inverse.distance * angle)$

— $ShotsFollowingaDribbleoftheKeeper : (-0.61 - 0.09 * distance + 7.4 * inverse.distance + 1.04 * angle - 3.2 * inverse.distance * angle + 1.1 * big.chance + 0.67 * following.error)$

b *Optimisation des différents modèles Elo*

Prédiction Elo 1 ($r_{init} = 1000$, $S =$ résultat)		
K (niveau de compétition)	h (avantage à domicile)	MSE
$K = 60$	$H = 8,387935 \cdot 10^{-9}$	0,3866
$K_{opti} = 76,98534$	$H_{opti} = 8,387935 \cdot 10^{-9}$	**0,3861**
Prédiction Elo 2 ($r_{init} = 1000$, $S =$ score)		
K (niveau de compétition)	h (avantage à domicile)	MSE
$K = 60$	$H = 30,69749$	0,0589
$K = 54,88164$	$H = 30,69749$	0,0588
$K_{opti} = 54,88164$	$H_{opti} = 31,40754$	**0,0588**
Prédiction Elo 3 ($r_{init} = FIFA$, $S =$ résultat)		
K (niveau de compétition)	h (avantage à domicile)	MSE
$K = 60$	$H = 8,387935 \cdot 10^{-9}$	0,3463
$K_{opti} = 43,06698$	$H_{opti} = 8,387935 \cdot 10^{-9}$	**0,3457**
Prédiction Elo 4 ($r_{init} = FIFA$, $S =$ score)		
K (niveau de compétition)	h (avantage à domicile)	MSE
$K = 60$	$H = 8,387935 \cdot 10^{-9}$	0,0660
$K_{opti} = 92,68889$	$H_{opti} = 8,387935 \cdot 10^{-9}$	**0,0656**
Prédiction Elo 5 ($r_{init} = 1000$, $S =$ résultat *expected goals* (Poisson))		
K (niveau de compétition)	h (avantage à domicile)	MSE
$K = 60$	$H = 1000$	0,1005
$K = 83,89989$	$H = 1000$	0,1000
$K_{opti} = 100,1694$	$H_{opti} = 254,1108$	**0,0997**
Prédiction Elo 6 ($r_{init} = 1000$, $S =$ score *expected goals*)		
K (niveau de compétition)	h (avantage à domicile)	MSE
$K = 60$	$H = 196,6504$	0,0275
$K = 84,09586$	$H = 196,6504$	0,0274
$K_{opti} = 101,0926$	$H_{opti} = 110,0072$	**0,0273**
Prédiction Elo 6 bis ($r_{init} = 1000$, $S =$ score *expected goals* (Poisson))		
K (niveau de compétition)	h (avantage à domicile)	MSE
$K = 60$	$H = 166,1335$	0,0212
$K = 83,76929$	$H = 166,1335$	0,0211
$K_{opti} = 100.5764$	$H_{opti} = 93.94369$	**0,0210**
Prédiction Elo 7 ($r_{init} = FIFA$, $S =$ résultat *expected goals* (Poisson))		
K (niveau de compétition)	h (avantage à domicile)	MSE
$K = 60$	$H = 8,387935 \cdot 10^{-9}$	0,0897
$K_{opti} = 65,056$	$H_{opti} = 8,387935 \cdot 10^{-9}$	**0,0897**

Prédiction Elo 8 ($r_{init} = FIFA$, $S =$ score *expected goals*)		
K (niveau de compétition)	h (avantage à domicile)	MSE
$K = 60$	$H = 8,387935 \cdot 10^{-9}$	0,0376
$K_{opti} = 155,0651$	$H_{opti} = 8,387935 \cdot 10^{-9}$	**0,0358**
Prédiction Elo 8 bis ($r_{init} = FIFA$, $S =$ score *expected goals* (Poisson))		
K (niveau de compétition)	h (avantage à domicile)	MSE
$K = 60$	$H = 8,387935 \cdot 10^{-9}$	0,0342
$K_{opti} = 182,3622$	$H_{opti} = 8,387935 \cdot 10^{-9}$	**0,0314**

TABLE 24 – Optimisation et valeur de MSE des classements Elo basés sur les buts réels et les *expected goals*

Prédiction Elo 2 ($r_{init} = 1000$, $S =$ score)		
K (niveau de compétition)	h (avantage à domicile)	MSE
$K = 60$	$H = 30,69749$	0,0589
$K = 54,88164$	$H = 30,69749$	0,0588
$K_{opti} = 54,88164$	$H_{opti} = 31,40754$	**0,0588**
Prédiction Elo 2 avec μ Elo 6 ($r_{init} = 1000$, $S =$ score)		
K (niveau de compétition)	h (avantage à domicile)	MSE
$K = 60$	$H = 4,319338$	0,0584
$K = 110,4486$	$H = 4,319338$	0,0580
$K_{opti} = 110,56$	$H_{opti} = 8,387935 \cdot 10^{-9}$	**0,0580**
Prédiction Elo 2 avec μ Elo 6 bis ($r_{init} = 1000$, $S =$ score)		
K (niveau de compétition)	h (avantage à domicile)	MSE
$K = 60$	$H = 7,9422$	0,0585
$K = 126,461$	$H = 7,9422$	0,0581
$K_{opti} = 126,6315$	$H_{opti} = 0,8938322$	**0,0581**

TABLE 25 – Optimisation et valeur de MSE de différents classements Elo sur leur prédiction des scores réels

C *Code R*

Voici ci-dessous le code R pour le calcul des modèles Elo basé sur les buts réels et les *expected goals* des rencontres. Il s'agit du code principal pour l'élaboration

de ce mémoire. Si vous souhaitez le code détaillé ainsi que les données utilisées, veuillez envoyer un e-mail à *thibault.gerard8@hotmail.com*.

```
function (data, K, initratings, H) {
    # data: "Date", "Team1", "Team2", "Team1.Score", "Team2.Score",
    "Team1.xg","Team2.xg", "Team1.Home", "Team2.Home", "Winner", "Loser",
    "Draw"
    # K: the K-factor
    # initratings: initial ratings
    # H: home-field advantage boost

    # 32 teams for the tournament
    Teams <- c(unique(dat$Team1))

    # Team1 expected score function
    E1 <- function(r1, r2) (10^(r1/400)) / (10^(r1/400) + 10^(r2/400))

    # Team2 expected score function
    E2 <- function(r1, r2) 1 - E1(r1, r2)

    # Elo updating formula
    RatingNew <- function (RatingOld, K, S, E) RatingOld + K*(S - E)

    # Teams' ratings
    ratings <- matrix(NA, nrow = dim(data)[1] + 1,
                      ncol = length(Teams))
    colnames(ratings) <- Teams
    rownames(ratings) <- paste("game",
                               0:dim(data)[1],
                               sep = "")

    # generate S for Team1 (1-S for Team2), according to game results
     S1 <- rep(NA, dim(data)[1])
     S1[which(as.character(data$Winner) ==
              as.character(data$Team1))] <- 1
     S1[which(is.na(data$Winner) )] <- 1/2
     S1[which(as.character(data$Winner) !=
              as.character(data$Team1))] <- 0

    # generate S for Team1 (1-S for Team2), incorporating game scores
    # S1 <- rep(NA, dim(data)[1])
    # for (i in 1 : dim(data)[1])
```

```
# S1[i] <- (data$Team1.Score[i] + 1) /
# (data$Team1.Score[i] + data$Team2.Score[i] + 2)

# generate S for Team1 (1-S for Team2), according to Poisson
 expected goals results
# S1 <- rep(NA, dim(data)[1])
# for (i in 1 : dim(data)[1])
# S1[i] <- 1*home[i]+(1/2)*draw[i]

# generate S for Team1 (1-S for Team2), incorporating expected
 goals scores
# S1 <- rep(NA, dim(data)[1])
# for (i in 1 : dim(data)[1])
# S1[i] <- (data$Team1.xg[i] + 1) /
# (data$Team1.xg[i] + data$Team2.xg[i] + 2)

# generate S for Team1 (1-S for Team2), according to Poisson
 expected goals scores
# S1 <- rep(NA, dim(data)[1])
# for (i in 1 : dim(data)[1])
#    S1[i] <- newm[i]

ratings[1, ] <- initratings

for (i in 1:dim(data)[1]) {

    r1 <- if (data[i, ]$Team1.Home == 1)
        ratings[i, as.character(data[i, ]$Team1)] + H else
            ratings[i, as.character(data[i, ]$Team1)]

    r2 <- if (data[i, ]$Team2.Home == 1)
        ratings[i, as.character(data[i, ]$Team2)] + H else
            ratings[i, as.character(data[i, ]$Team2)]

    ratings[i + 1, as.character(data[i, ]$Team1)] <-
        RatingNew(ratings[i, as.character(data[i, ]$Team1)],
                  K,
                  S = S1[i],
                  E = E1(r1, r2)
        )
```

```
        ratings[i + 1, as.character(data[i, ]$Team2)] <-
            RatingNew(ratings[i, as.character(data[i, ]$Team2)],
                      K,
                      S = 1 - S1[i],
                      E = E2(r1, r2)
            )

        ratings[i + 1, ][is.na(ratings[i + 1, ])] <-
            ratings[i, ][is.na(ratings[i + 1, ])]

    }

    # MSE (Brier score)
    mse <- (sum((S1 -
                    sapply(1:dim(data)[1], function(i)
                        E1(
                            ratings[i, as.character(data[i, ]$Team1)],
                            ratings[i, as.character(data[i, ]$Team2)]
                        )
                    )
    )^2,
    ((1 -S1) -
         sapply(1:dim(data)[1], function(i)
             E2(
                 ratings[i, as.character(data[i, ]$Team1)],
                 ratings[i, as.character(data[i, ]$Team2)]
             )
         )
    )^2
    )/64)

    return(list(ratings = ratings,
                mse = mse))
}
```

Cinquième partie

Bibliographie

(1) 11tegen11.net. (2014). *expected goals* 2.0 – Some light in the black box. Récupéré sur : http://11tegen11.net/2014/08/07/expected-goals-2-0-some-light-in-the-black-box/.

(2) Alrababah, A., Marble, W., Mousa, S., & Siegel, A. (2019). Can Exposure to Celebrities Reduce Prejudice ? the Effect of Mohamed Salah on Islamophobic Behaviors and Attitudes.

(3) Bertin, M. (2015). The third-t-last thing I'll ever write about *expected goals*. Récupéré sur : http://michaelbertin.com/2015/08/28/the-third-to-last-thing-ill-ever-write-about-expected-goals/.

(4) Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs : I. The method of paired comparisons. *Biometrika, 39*(3/4), 324-345.

(5) Brefeld, U., Davis, J., Van Haaren, J., & Zimmermann, A. (2019). Machine Learning and Data Mining for Sports Analytics, 5th International Workshop, MLSA 2018, Co-located with ECML/PKDD 2018, Dublin, Ireland, September 10, 2018, Proceedings.

(6) Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthey Weather Review, 78*(1), 1-3.

(7) Caley, M. (2015). Premier League Projections and New *expected goals*. Récupéré sur : https://cartilagefreecaptain.sbnation.com/2015/10/19/9295905/premier-league-projections-and-new-expected-goals.

(8) Cecco, L. (2019). Toronto Raptors' success gives Canadian sports fans a rare feeling : hope. Récupéré sur : https://www.theguardian.com/sport/2019/jun/09/toronto-raptors-nba-finals-canada-fans.

(9) Chen, C. H. (2018) Elo Rating System for UEFA Women's Euro 2017.

(10) Davidson, R. R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association, 65*(329), 317-328.

(11) Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics, 20*(1), 134-144.

(12) Duseaux, G. (2017). GOAT – les plus grands sportifs de tous les temps. Récupéré sur : https://lasueur.com/goat-plus-grands-sportifs.

(13) Dyte, D., & Clarke, S. R. (2000). A ratings based Poisson model for World Cup soccer simulation. *Journal of the Operational Research society, 51*(8), 993-998.

(14) Elo, A. E. (1978). The rating of chessplayers, past and present. *Arco Pub.*

(15) Ensum, J., Pollard, R., & Taylor, S. (2005). Applications of logistic regression to shots at goal at association football. *In Science and football V : the proceedings of the Fifth World Congress on Science and Football* (p. 214).

(16) FE Online (2018). France, Croatia, Belgium receive heroes' welcome after World Cup glory ; see pics. Récupéré sur : https://www.financialexpress.com/photos/business-gallery/1247753/france-croatia-belgium-receive-heroes-welcome-after-world-cup-glory-see-pics/.

(17) FIDE. Calculators : Ratings Change Calculator. Récupéré sur : https://ratings.fide.com/calculator_rtd.phtml.

(18) FIFA.com.(2011) FIFA/Coca-Cola Women's World Ranking. Récupéré sur : https://img.fifa.com/image/upload/rxqyxdjhbs2qdtstluy6.pdf.

(19) FIFA.com (2017). Tirage au sort, mode d'emploi. Récupéré sur : https://fr.fifa.com/worldcup/news/tirage-au-sort-mode-d-emploi-2921493.

(20) FIFA.com (2018). En cas d'égalité dans les groupes.... Récupéré sur : https://fr.fifa.com/worldcup/news/en-cas-d-egalite-dans-les-groupes.

(21) Gelade, G. (2017). Assessing *expected goals* Models. Part 1 : Shots. Récupéré sur : http://business-analytic.co.uk/blog/evaluating-expected-goals-models/.

(22) Gelade, G. (2017). Assessing *expected goals* Models. Part 2 : Anatomy of a Big Chance. Récupéré sur : http://business-analytic.co.uk/blog/assessing-expected-goals-models-part-2-anatomy-of-a-big-chance/.

(23) Glickman, M. E. (1995). A comprehensive guide to chess ratings. *American Chess Journal, 3*(1), 59-102.

(24) Glickman, M. E. (1995). The glicko system. Boston University.

(25) Glickman, M. E. (2012). Example of the Glicko-2 system. Boston University, 1-6.

(26) Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal of Forecasting, 21*, 331–340.

(27) Govan, A. Y., & Meyer, C. D. (2006). Ranking national football league teams using google's pagerank. *In AA Markov Anniversary Meeting.*

(28) Graepel, T., & Herbrich, R. (2006). Ranking and matchmaking. *Game Developer Magazine, 25*, 34.

(29) Hill, I. D. (1974). Association football and statistical inference. *Journal of the Royal Statistical Society : Series C (Applied Statistics), 23*(2), 203-208.

(30) Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models. *The annals of statistics, 32*(1), 384-406.

(31) Hvattum, L. M., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of forecasting, 26*(3), 460-470.

(32) Kvam, P., & Sokol, J. S. (2006). A logistic regression/Markov chain model for NCAA basketball. *Naval Research Logistics (NrL), 53*(8), 788-803.

(33) Langville, A. N., & Meyer, C. D. (2012). Who's♯1? : the science of rating and ranking. Princeton University Press.

(34) Lasek, J., Szlávik, Z., & Bhulai, S. (2013). The predictive power of ranking systems in association football. *International Journal of Applied Pattern Recogni-*

tion, 1(1), 27-46.

(35) Leitner, C., Zeileis, A., & Hornik, K. (2010). Forecasting sports tournaments by ratings of probabilities : A comparison for the EURO 2008. *International Journal of Forecasting, 26*(3), 471-481.

(36) Leitner, C., Zeileis, A., & Hornik, K. (2010). Forecasting the winner of the FIFA World Cup 2010.

(37) Macdonald, B. (2012). An *expected goals* model for evaluating NHL teams and players. *In Proceedings of the 2012 MIT Sloan Sports Analytics Conference*, http://www.sloansportsconference.com.

(38) Massey, K. (1997). Statistical models applied to the rating of sports teams. Bluefield College.

(39) Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking : Bringing order to the web. Stanford InfoLab.

(40) Pollard, R. (2006). Worldwide regional variations in home advantage in association football. *Journal of sports sciences, 24*(3), 231-240.

(41) Pollard, R. (2008). Home advantage in football : A current review of an unsolved puzzle. *The open sports sciences journal, 1*(1).

(42) Pollard, R., & Reep, C. (1997). Measuring the effectiveness of playing strategies at soccer. *Journal of the Royal Statistical Society : Series D (The Statistician), 46*(4), 541-550.

(43) 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.

(44) Rathke, A. (2017). An examination of *expected goals* and shot efficiency in soccer. Journal of Human Sport and Exercise, 12(2proc), S514-S529.

(45) Reep, C., & Benjamin, B. (1968). Skill and chance in association football. Journal of the Royal Statistical Society. *Series A (General), 131*(4), 581-585.

(46) Riley, P. (2014). A Shooting Model – An Exp(G)lanation and Application. Récupéré sur : https://differentgame.wordpress.com/2014/05/19/a-shooting-model-an-expglanation-and-application/.

(47) Ryder, A. (2004). Shot Quality. Récupéré sur : http://hockeyanalytics.com/\Researchfiles/ShotQuality.pdf.

(48) Silver, N. Boice, J. Paine, N. (2019). How Our NBA Predictions Work. Récupéré sur : https://fivethirtyeight.com/methodology/how-our-nba-predictions-work/.

(49) Smith, K. (2018). Let the World Have Soccer. Récupéré sur : https://www.nationalreview.com/2018/06/soccer-corrupt-hyper-regulated-low-scoring-boring/.

(50) Spearman, W. (2018). Beyond *expected goals*. In Proceedings of the 12th MIT sloan sports analytics conference (pp. 1-17).

(51) Sprigings, T. (2016). Game Changers : Greece's Underdog Victory at Euro 2004. Récupéré sur : https://www.vice.com/en_uk/article/78gzeg/game-changers-greeces-underdog-victory-at-euro-2004.

(52) Stanton, J. (2017). Expected goals : What are we learning from new metric used on Match of the Day ?. Récupéré sur : https://www.bbc.com/sport/football/41822455.

(53) Sullivan, C., & Cronin, C. (2016). Improving Elo Rankings For Sports Experimenting on the English Premier League. Virginia Tech CSx824/ECEx424 technical report.

(54) Wang, C., & Vandebroek, M. L. (2013). A model based ranking system for soccer teams. Available at SSRN 2273471.

(55) Whoscored.com. Glossaire. Récupéré sur : https://fr.whoscored.com/Glossary.

(56) Worville, T. (2016). xGELO : Combining *expected goals* And ELO Ratings. Récupéré sur : https://medium.com/@worville/xgelo-combining-expected-goals-and-elo-ratings-6aa987481479.

(57) Zermelo, E. (1929). Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift, 29*(1), 436-460.

LOUVAIN-LA-NEUVE | BRUXELLES | MONS | TOURNAI | CHARLEROI | NAMUR
Place des Doyens, 1 bte L2.01.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/lsm**