Projet Scientifique Collectif
Les Paris Sportifs

Dugarzhapov, Hippolyte, Hoang, Jacquemart, Meyer, Sellami, Watine

Lundi 4 mai 2009

Introduction

Le marché des paris sportifs est en plein essor. Bien qu'étant déjà populaires depuis de nombreuses années dans les pays anglo-saxons, ces paris connaissent depuis quelques années une expansion considérable dans plusieurs pays, portés par la multiplication des sites de pari en ligne. En France, ce marché sera libéralisé le 1e janvier 2010, alors que la Française des Jeux en détenait jusqu'ici le monopole.

Bien que le marché des jeux d'argent ait déjà été largement étudié, il semble que le nombre d'études concernant les paris sportifs soit encore relativement faible, en raison de leur développement récent. Nous avons donc décidé d'étudier ce marché dans le cadre de notre PSC. Cette démarche a pour but de mettre à profit nos profils scientifiques dans le but d'explorer un sujet nouveau, et jusqu'ici peu abordé sous l'angle scientifique. En effet, il ne semble pas évident à première vue qu'il soit possible d'envisager de manière rationnelle un environnement aussi incertain que celui des événements sportifs. C'est toutefois l'enjeu de notre démarche, dans laquelle nous utiliserons en particulier nos connaissances en matière de microéconomie et de mathématiques appliquées. Notre problématique est la suivante : " Quels procédés utilisent les bookmakers pour fixer les cotes des matchs de football afin de maximiser leurs gains ? "

Pour y répondre, nous avons divisé notre étude en deux approches principales. Dans une première partie, nous étudierons le marché des paris sportifs en analysant les modèles existants concernant les jeux d'argent ; il s'agira en particulier d'analyser les comportements du bookmaker et des parieurs, ainsi que leurs interactions éventuelles. Ces études soulèveront principalement des problèmes de nature économique et sociologique. Notre deuxième partie consistera quant à elle à modéliser les matchs de football, pour permettre de fixer les cotes d'une manière rationnelle. Enfin, nous répondrons à notre problématique initiale en confrontant ces deux approches, et en nous appuyant sur des résultats expérimentaux.

Résumé

Dans la première partie de notre PSC, nous modélisons le marché des paris sportifs. Après avoir présenté les différents modèles existants concernant les jeux d'argent (casino, PMU), nous modélisons les comportements des deux agents de ce marché : le vendeur (bookmaker) d'une part, et l'acheteur d'autre part (parieur).

L'étude du comportement du bookmaker est très théorique, puisqu'il est impossible d'obtenir des données réelles auprès des bookmakers eux-mêmes. Nous décrivons donc formellement le comportement d'un bookmaker cherchant à maximiser son profit en fonction des probabilités des matchs.

L'étude du comportement des parieurs se prête quant à elle plus à l'étude expérimentale. Cette étude consiste principalement à dresser la courbe d'utilité du parieur moyen. Nous donnons tout d'abord quelques résultats généraux obtenus au cours d'études précédentes, en particulier ceux d'un article de Jullien et Salanié paru en 2000. Puis nous réfléchissons à de nouveaux modèles permettant de représenter les parieurs. Nous mettons en particulier au point deux nouvelles formes de fonctions d'utilité envisageables : notre approche consiste en particulier à modéliser un comportement qui n'a jusqu'ici pas été pris en compte. Selon nous, il ne convient pas seulement de mesurer les gains et pertes possibles pour le parieur pour déterminer la manière dont il va parier, mais il convient également de prendre en compte l'excitation liée au fait-même de parier : si tant de gens jouent au Loto alors que l'espérance de gain est négative, c'est pour l'excitation que procure l'espoir de gagner de grandes sommes. Nous présentons enfin les expériences que nous avons mises au point afin de perfectionner ces modèles ; nous en donnerons les résultats dans la troisième partie.

Notre deuxième partie constitue l'aspect le plus technique de notre étude ; il s'agit de modéliser les matchs de football eux-mêmes. Cela correspond à la mission première du bookmaker. En effet, si notre première partie nous permet de donner les cotes maximisant le gain du bookmaker pour des probabilités données, il s'agit ici de pouvoir effectivement évaluer ces probabilités. Il est vrai que la détermination réelle des cotes des matchs par les bookmakers existants s'appuie en bonne partie sur des aspects très techniques et ponctuels, qui constituent d'innombrables paramètres quasi-impossibles à modéliser efficacement (blessures, fatigue, météo, enjeu des matchs, psychologie des joueurs, hasard). Nous proposons ici de manière théorique une manière rationnelle de déterminer ces cotes.

Nous mettons au point dans cette partie différentes méthodes, de complexité croissante. Ces méthodes font appel à nos connaissances en mathématiques appliquées, et leur mise en oeuvre s'appuie surtout sur des programmes informatiques. Nous modélisons en particulier la probabilité pour les équipes de marquer un but à l'aide de lois de Poisson, et en prenant en compte les résultats passés des équipes. Nous prenons également en compte le fait que les équipes jouent à domicile ou à l'extérieur, facteur ayant un impact non négligeable. Enfin, nous proposons une dernière approche dans laquelle nous prenons en compte la forme actuelle des équipes, en donnant un plus grand poids à leurs résultats les plus récents.

Notre troisième partie est très expérimentale. Elle consiste à présenter les résultats obtenus par l'exploitation de nos modèles. Tout d'abord, nous consolidons les modèles de fonctions d'utilité imaginés dans notre première partie en leur donnant une valeur expérimentale. Nous

voulions prendre en compte l'effet de l'excitation liée au fait-même de parier. Pour cela, nous avons organisé un sondage auprès des étudiants de l'école, afin de mettre en évidence cet effet, et d'en mesurer le poids.

Ensuite, nous présentons les résultats obtenus par informatique concernant l'évaluation des probabilités. Enfin, nous mettons en pratique les résultats de notre étude en mêlant les deux approches précédentes de notre sujet : en utilisant les données réelles du championnat de France 2008/2009, nous simulons pour chaque match de la journée du 2 mai le calcul des cotes que doit fixer le bookmaker pour maximiser son gain. Nous effectuons cette simulation pour nos deux modèles envisagés et commentons leur validité, en comparant les cotes calculées à celles proposées sur les principaux sites de paris en ligne.

Table des matières

1 Modélisation du marché **7**
1.1 Comportement du bookmaker . . . 7
1.1.1 Introduction au pari sportif . . . 7
1.1.2 Etude du gain . . . 8
1.2 Description du parieur . . . 9
1.2.1 Observations du comportement des parieurs . . . 9
1.2.2 Explications des phénomènes rencontrés . . . 11
1.2.3 Notre modèle . . . 12
1.2.4 Calcul des paramètres des utilités . . . 13
1.3 Calcul des cotes . . . 14
1.3.1 Principe général du calcul des cotes . . . 14
1.3.2 Etude théorique des utiliés u_1 etu_2 . . . 14
1.3.3 Etude numérique . . . 15

2 Modélisation des matchs de football **17**
2.1 Loi de Poisson . . . 18
2.2 Maximum de vraisemblance . . . 19
2.2.1 Principe . . . 19
2.2.2 Méthode par itérations . . . 20
2.2.3 Méthode de Newton . . . 24
2.3 Variation des paramètres au cours de la saison . . . 26
2.3.1 Intérêt . . . 26
2.3.2 Pondération exponentielle . . . 27
2.3.3 Pondération polynômiale . . . 28

3 Résultats de l'étude et commentaires **31**
3.1 Résultats numériques de la modélisation des matchs de football . . . 31
3.1.1 Calcul des coefficients . . . 31
3.1.2 Simulation de la fin du championnat . . . 32
3.2 Calcul numérique des cotes . . . 34
3.2.1 Résultats du sondage . . . 34
3.2.2 Etude d'un pari à une issue . . . 35
3.2.3 Comparaison des cotes obtenues en utilisant u_1 et $u2$ . . . 36
3.3 Les cotes et les probabilités de la 34ème journée . . . 37
3.3.1 Remarques particulières sur les résultats de l'annexe 1 . . . 38
3.3.2 Remarques générales sur les résultats de l'annexe 1 . . . 39

4 Remarques et remerciements **41**
4.1 Organisation du travail 41
4.2 Problèmes rencontrés 42
4.3 Bibliographie 42
4.4 Remerciements 43

5 Conclusion **45**

6 Annexes **47**
6.1 Annexe 1 : comparaison des cotes 47
6.2 Annexe 2 : exploitation du sondage 47
6.3 Annexe 3 : calcul des cotes 47
6.4 Annexe 4 : comparaison des p et λ 47

Chapitre 1

Modélisation du marché

Commençons par une étude du marché des paris sportifs. Il convient en premier lieu de modéliser le comportement des différents acteurs, c'est-à-dire leurs différents intérêts, pour comprendre comment les paramètres des paris sont fixés. Nous allons donc réaliser des études microéconomiques, en commençant par l'agent économique qui est à l'origine des paris : le bookmaker, qui fixe les cotes. Ensuite, nous verrons ce qu'il en est du comportement de l'autre acteur essentiel : le parieur. Enfin, en confrontant les points de vue de ces différents agents économiques, nous essaierons de déterminer comment sont fixées les cotes.

1.1 Comportement du bookmaker

Le bookmaker est le principal agent économique de ce marché. En effet, c'est celui qui fixe les cotes, qui correspondent à des prix du jeu. Pour comprendre ses choix, nous commencerons par introduire la notion de pari sportif, avant de voir deux déterminations polaires des cotes. Pour finir cette approche du bookmaker, nous ferons une étude plus générale de son espérance de gain.

1.1.1 Introduction au pari sportif

Commençons par expliquer le déroulement d'un pari sportif. Un match entre une équipe A et une équipe B a trois issues possibles : la victoire de A (1), le match nul (2) et victoire de B (3). A chaque événement (i), on associe la cote C_i.

Définissons les probabilités p_i des événements, les sommes d'argent m_i misées sur les événements. On posera la somme totale misée $M = \sum m_i$ et les proportions d'argent misé $f_i = \dfrac{m_i}{M}$.

Un pari sportif se déroule en deux phases. Avant le match, le bookmaker choisit ses cotes C_i, et les joueurs misent les sommes m_i que récolte le bookmaker. Après le match, ce dernier paie les vainqueurs $m_i C_i$.

On peut remarquer que cela correspond tout à fait à l'achat d'un bon de loterie, dont le prix est $\dfrac{1}{C_i}$, qui raporte 1 avec une probabilité p_i et 0 avec une probabilité $1 - p_i$. Quelle que soit la vision adoptée, le gain du bookmaker vérifie $G_\lambda = M\big[1 - \sum f_i C_i 1_{\mathcal{S}_i}\big]$.

Remarquons que le principe d'arbitrage (les joueurs ne doivent pas pouvoir gagner de façon certaine) nécessite que $1 \leq \sum \frac{1}{C_i} = \frac{1}{1-\gamma}$, avec $\gamma \geq 0$. Posons ensuite $\lambda_i = \frac{1-\gamma}{C_i}$, qui correspond aux probabilités induites par les cotes, dont on peut remarquer que $\sum \lambda_i = 1$. Le gain s'écrit alors $G_\lambda = M[1-(1-\gamma)\sum \frac{f_i}{\lambda_i} 1_{\mathcal{S}_i}]$.

Il vient $E[G_\lambda] = M[1-(1-\gamma)\sum \frac{p_i f_i}{\lambda_i}]$ et $\operatorname{Var} G_\lambda = M^2(1-\gamma)^2\big[\sum \big(\frac{f_i}{\lambda_i}\big)^2 p_i - \big(\sum \frac{f_i}{\lambda_i} p_i\big)^2\big]$.

1.1.2 Etude du gain

Au PMU, le bookmaker rassemble toutes les mises, en récupère une portion γ et répartit la somme totale misée entre les vainqueurs, proportionnellement au rapport de la somme qu'ils ont pariée sur la somme totale misée sur l'issue victorieuse. Ainsi si un joueur k mise m_i^k sur le cheval i, et si ce cheval gagne, il récupère $(1-\gamma)M\frac{m_i^k}{m_i} = \frac{1-\gamma}{f_i} m_i^k$, ce qui est équivalent au modèle précédent, avec des cotes $\lambda_i^{\mathrm{PMU}} = f_i$. On a alors $E[G^{\mathrm{PMU}}] = \gamma M$ et $\operatorname{Var} G^{\mathrm{PMU}} = 0$. On voit alors que γ est toujours la fraction des mises que récolte le bookmaker, ou plutôt le casino. La variance quant à elle est nulle, ce qui est cohérent avec la description précédente.

Au casino, le bookmaker fixe donc ses cotes uniquement par rapport aux probabilités p_i. Il pose alors $\lambda_i^{\mathrm{Cas}} = p_i$. On a alors $E[G^{\mathrm{Cas}}] = \gamma M$ et $\operatorname{Var} G^{\mathrm{Cas}} = M^2(1-\gamma)^2\big[\sum \big(\frac{f_i^2}{p_i}\big) - 1\big]$. On voit alors que γ est la fraction des mises que récolte le casino. La variance quant à elle est non nulle, et on voit qu'elle peut même être importante pour une mauvaise répartition des mises, c'est-à-dire, par exemple, lorsque l'issue (3) rencontre tous les preneurs, auquel cas, $\operatorname{Var} G^{\mathrm{Cas}} = M^2(1-\gamma)^2(\frac{1}{p_3} - 1)$ est élevée. Cependant, la loi des grands nombres assure au casino une variance quasi-nulle.

Ces deux exemples correspondent à une même espérance. Est-il possible de faire mieux ? Maximisons $E[G_\lambda]$ en fonction de $\lambda \in P = \{\lambda \in \mathbb{R}_+^{*\,3} | \sum \lambda_i = 1\}$. A un extrémal, la tangente de la variété $E[G_\lambda] = cste$ est égale au plan P. Les deux vecteurs orthogonaux aux plans tangents sont alors proportionnels, ce qui implique que $\nabla E[G_\lambda] \propto (1,1,1)$, qui se traduit également par $\exists A / \forall i, \partial_i E[G_\lambda](\lambda^*) = A$. On en déduit alors $\exists A' / \forall i, \frac{f_i p_i}{\lambda_i^{*2}} = A'$, d'où $\exists A'' / \forall i, \lambda_i^* = A''\sqrt{f_i p_i}$. La condition $\sum \lambda_i^* = 1$ permet ensuite de conclure. On a donc $\forall i, \lambda_i^* = \frac{\sqrt{f_i p_i}}{\sum \sqrt{f_j p_j}}$. Par concavité de l'espérance de gain, ou par unicité du problème de recherche d'extrema réalisé, on peut dire que c'est nécessairement l'unique maximum.

Ainsi, l'espérance de gain optimal s'écrit $E[G^*] = M[1-(1-\gamma)\big(\sum \sqrt{f_i p_i}\big)^2]$ alors que la variance est $\operatorname{Var} G^* = M^2(1-\gamma)^2\big(\sum \sqrt{f_i p_i}\big)^2[1 - \big(\sum \sqrt{f_i p_i}\big)^2]$. On peut ainsi remarquer que si les f_i sont grands lorsque les p_i sont faibles et inversement, l'espérance de gain est plus grande. Ainsi, le bookmaker se doit d'espérer (ou de faire en sorte) que les joueurs parient beaucoup sur les événements peu probables.

Par ailleurs, on remarque que l'espérance de gain est positive si $\sum \frac{f_i p_i}{\lambda_i} \leq \frac{1}{1-\gamma}$. Comme on l'a vu dans les modèles du PMU et du casino, γ représente la marge réalisée par le bookmaker. On voit très bien ici que plus γ est grand, plus $g^{-1}(\mathbb{R}_+) \cap P$ est grand. Encore une fois, cet ensemble est également d'autant plus grand que les joueurs parient beaucoup sur les issues à faible probabilité.

En ce qui concerne les paris sportifs portant sur les matchs de football, les valeurs des probabilités p_i sont difficiles à déterminer et les répartitions des mises f_i sont fixées a posteriori des cotes, c'est-à-dire que les f_i sont des fonctions des λ_i, ce qui rend le problème bien plus complexe. Pour déterminer la relation entre les f_i et les λ_i, intéressons-nous donc aux parieurs.

1.2 Description du parieur

Nous allons maintenant modéliser le comportement du parieur. Cette étude est particulièrement délicate, à cause de diverses raisons que nous allons expliciter, mais elle a pu progresser grâce à différentes expériences que nous détaillerons. Nous donnerons qualitativement les principales explications aux résultats de ces expériences, avant d'introduire les modèles que nous étudierons par la suite.

1.2.1 Observations du comportement des parieurs

Il apparaît que le consommateur comporte des aspects très particuliers dans le cadre des jeux d'argent, et ce pour diverses raisons. Tout d'abord, les informations sur le bien proposé sont incomplètes puisque les issues des matchs sont incertaines et dépendent de facteurs parfois inconnus. De plus, les parieurs ont un accès inégal à ces informations, et se différencient notamment par leurs croyances personnelles, ou la subjectivité de leurs pensées. Ensuite, il est évident que le goût ou l'aversion pour le risque différent beaucoup d'un consommateur à l'autre et influent sur le type de paris qu'ils réalisent.

Jullien et Salanié[1] ont publié en 2000 l'article " Empirical evidence on the preferences of racetrack bettors ". Ils y étudient le comportement des parieurs et détaillent différents modèles ayant déjà été proposés. Mais des études sur le sujet sont délicates car les expériences réalisées reproduisent mal le marché des paris : en effet, les gains ou pertes d'argent réelles sont difficilement comparables avec les gains ou pertes simulées lors d'expériences. Un individu participant à une simulation sera par exemple plus enclin à prendre de grands risques dès lors que son patrimoine réel n'est pas menacé.

Malgré cela, des quantités significatives de données ont pu être obtenues à partir des courses hippiques, et permettent d'étudier les théories de préférence sous risque. Remarquons que l'étude des marchés financiers fournirait également de grandes quantités de données, mais celles-ci seraient plus difficiles à exploiter car elles mettent en jeu des périodes plus longues et des paris plus complexes, dont le revenu ex-post est parfois peu lisible. Les courses hippiques ont quant à elles deux avantages : le résultat est obtenu à court terme et permet une évaluation exacte de l'issue du pari.

Une particularité se dégage des données étudiées. Les parieurs ont tendance à miser anormalement trop d'argent sur les cotes élevées (outsiders), et pas assez sur les faibles cotes (favoris).

Deux expériences principales ont montré empiriquement l'existence de ce biais. Détaillons ces deux expériences.

Première expérience

En 1949, Griffith[1] étudie 1386 courses hippiques. Il divise toutes les cotes en classes C, et compare deux données en fonction des classes C :
- E_c le nombre total de chevaux ayant une cote dans la classe C.
- N_c le produit du nombre de gagnants dans cette classe par leur cote C.

En pariant 1 sur chaque cheval de la classe C (soit une somme totale E_c), on récupère N_c. Il y a alors bénéfice si $N_c > E_c$. En superposant les deux courbes, on constate qu'elles ont une allure similaire. Cependant, la courbe Nc se situe au-dessus lorsque C est faible, et en-dessous lorsque C est élevé. Autrement dit, un parieur ayant un comportement rationnel devrait miser tout son argent sur le favori.

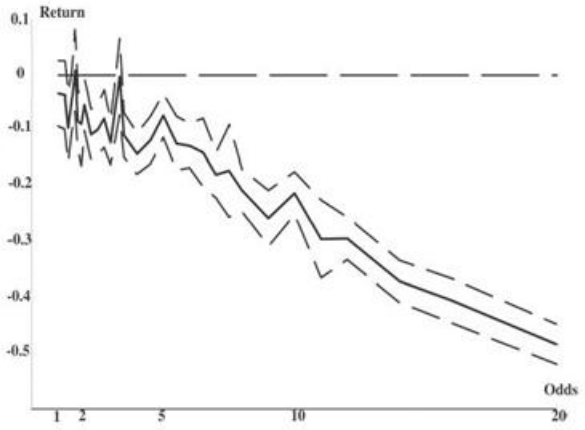


Figure 1: Observed Expected Return

Le fait qu'il n'en soit pas ainsi témoigne alors du biais des parieurs, qui ont une tendance malheureuse à trop parier sur les cotes élevées.

Deuxième expérience

Jullien et Salanié[1] ont collecté les données de 34443 courses hippiques en Grande-Bretagne entre 1986 et 1995. Pour chaque classe C, ils calculent la probabilité empirique de victoire $p(C)$. L'espérance de gain pour une mise de 1 est alors : $EC(C) = p(C)(C-1) - (1-p(C))$. En traçant $EC(C)$ en fonction de C, ils obtiennent des valeurs toujours négatives avec une courbe quasi-linéaire et décroissante. Ainsi, un individu rationnel et neutre au risque (qui maximise donc simplement son espérance de gain) ne parie pas, et s'il doit parier, mise tout son argent sur le

favori.

Ces deux expériences mettent donc en évidence l'existence d'un biais qui détourne les parieurs des choix rationnels. C'est le *favorite-longshot bias*.

1.2.2 Explications des phénomènes rencontrés

Reprenons. Alors que l'espérance de gain du bookmaker est positive, et que celle de la plupart des parieurs est en conséquence négative, de nombreux joueurs parient malgré tout, en dépit de leur espérance de gain négative. On a également souligné le "favorite longshot bias", qui correspond à la surévaluation des grosses cotes correspondant aux phénomènes de faible probabilité. On peut avancer trois explications importantes à ces deux phénomènes.

La raison généralement avancée dans le cas d'une loterie pour expliquer le choix des personnes est la forme de la fonction d'utilité qui leur est associée. Souvent concave, ce qui correspond à une aversion au risque, elle devient convexe pour ceux qui aiment le risque ; en d'autres termes, l'utilité marginale de l'argent est pour eux croissante, et il est ainsi beaucoup plus intéressant d'avoir 2000 plutôt que 1000 . Modélisant l'ensemble des parieurs par un parieur moyen que nous définirons ultérieurement, cette étude réalisée par Jullien et Salanié[1] conduit à caractériser les joueurs par un goût du risque. Cette théorie corrobore le phénomène de " favorite longshot bias " évoqué précédemment : le fait de parier sur les grandes cotes, qui dans le cas des courses hippiques, correspondent pourtant à des espérances plus faibles, conduit à augmenter significativement la variance du gain. Ainsi, la possibilité de gagner beaucoup plus correspond à des joueurs dont le goût du risque est grand.

Remarquons enfin la mauvaise perception des probabilités par les parieurs. En 1985, Henery[1] affirme que les parieurs ont tendance à sous-estimer les probabilités de défaite des outsiders ; si la vraie probabilité est q, le parieur pense qu'elle vaut $Q = fq$, avec $0 < f < 1$. Q est alors la valeur de la probabilité qui donne une espérance de gain nulle, et vaut alors $1 - 1/c$, tandis que q vaut $1 - p(c)$. En régressant Q sur q, il trouve $f = 0,975$, qui mesure le biais apporté par la mauvaise perception des parieurs.

Cependant, cette raison conduit à considérer que tous les parieurs aiment le risque ou l'estiment mal, ce qui est en fait une hypothèse discutable. En effet, il semble que les paris sportifs ne sont pas seulement motivés par un simple goût du risque, et que certains parieurs jouent même s'ils ont conscience que leur espérance de gain est négative. On peut certainement expliquer cela par le plaisir et l'excitation causés par le fait-même de parier. En quelque sorte, les parieurs sont prêts à " payer " pour la montée d'adrénaline que leur offre le pari. Ainsi, l'utilité tirée du simple fait de parier peut augmenter avec la variance du gain, et donc indirectement avec la cote ou la somme misée, car le divertissement est dès lors plus intense. On retrouve alors le fait que l'espérance de gain est plus élevée pour les petites cotes : c'est l'excitation liée au jeu qui compense la perte d'espérance d'utilité liée au gain.

C'est à partir de cette dernière remarque, très peu étudiée dans les articles portant sur les paris sportifs que nous réaliserons notre étude. En effet, on va maintenant introduire des modèles de parieurs averses au risque, dotés une vision juste des probabilités, mais possédant une envie de jouer, une excitation due aux paris sportifs.

1.2.3 Notre modèle

On va maintenant chercher à créer un modèle. Le but est de définir, étant donné un match, avec des probabilités de résultat bien définies, et des cotes fixées, comment les parieurs répartissent leurs mises, pour en déduire quelle est l'espérance de gain du bookmaker.

La difficulté à représenter l'ensemble des parieurs réside entre autre dans le fait qu'ils ont des profils hétérogènes. Il existe ainsi une distribution sous-jacente de leurs caractéristiques (préférences, croyances). Pour surmonter ce problème, notre approche consiste à représenter l'ensemble des parieurs par un unique individu, dont on construit la fonction d'utilité à partir des sommes misées par l'ensemble des parieurs sur les différentes cotes. Cet individu fictif représente l'ensemble de la population des parieurs ; il a ainsi une fonction d'utilité compatible avec l'ensemble des paris réalisés. En d'autres termes, s'il avait misé seul une somme d'argent égale à celle misée par l'ensemble des parieurs, il aurait pu réaliser les mêmes paris qu'eux.

En nous inspirant des études préexistentes, nous avons déterminé un certain nombre d'utilités "'type"'. L'idée à chaque fois est de définir une utilité u dépendant à la fois du gain G et de la mise m, traduisant ainsi l'excitation liée au pari. L'utilité $u(G, m)$ doit donc être une fonction croissante de G et m.

Remarquons d'autre part qu'il est nécessaire que l'effet de gain soit plus important que l'effet de mise, c'est-à-dire que le joueur préfèrera ne pas jouer plutôt que jouer et être sûr de perdre sa mise. Cela équivaut à l'inégalité $u(0,0) > u(-m, m)$ pour tout $m > 0$. Mieux, il faut que la mise avec perte sûre ait une utilité plus faible lorsque la mise est grande, ce qui conduit à une fonction $m \mapsto u(-m, m)$ décroissante. Cette hypothèse est incontournable, puisqu'elle justifie la mise finie des parieurs. En effet, dans le cas contraire, les joueurs auraient tout intérêt à faire des mises infinies.

On en vient à considérer l'utilité $u_1(G, m) = -\exp(-\alpha G - \beta m)$ pour $\alpha > \beta$, une fonction négative dont l'effet de mise est traduite par une diminution de la valeur absolue de l'utilité, et dont l'effet de gain est une fonction CARA croissante et concave. Cela correspond à la remarque selon laquelle les parieurs n'ont pas forcément le goût du risque.

Remarquons un défaut majeur de cette fonction d'utilité simple. Le parieur préfèrera parfois miser pour une cote $C = 1$, notamment lorsque la probabilité de gain p est grande. En effet, pour $C = 1$, $E[u] = -p\exp(-m(\alpha+\beta)) - (1-p)\exp((\alpha-\beta)m)$, qui pour une mise m fixée, est supérieure à -1 pour p assez grand, car $E[u]$ tend vers $\exp(-m(\alpha+\beta))$ si p tend vers 1. Ainsi, la fonction d'utilité a le défaut d'envisager le cas de parieurs misant alors qu'ils sont sûrs de perdre.

Pour résoudre ce problème, on remarque que l'effet d'excitation ne dépend ici que de la somme misée. On en vient donc à considérer un effet d'excitation liant la somme misée à la somme que la mise peut rapporter, c'est-à-dire la mise et la cote. La fonction d'utilité $u_2(G, m) = -\exp(-\alpha G - (\beta R/C)m)$, où $R = C - 1$, répond aux exigences soulevées si $\alpha > \beta$. En effet, lorsque $C = 1$ l'effet d'excitation n'existe pas, et le joueur se retrouve face à un choix entre perdre sa mise avec une probabilité $1 - p$ et ne rien gagner avec une probabilité p. Ainsi, il préfèrera ne rien miser.

1.2.4 Calcul des paramètres des utilités

Pour déterminer les paramètres α et β qui caractérisent les joueurs, nous réalisons un sondage nous permettant d'interroger les individus sur leurs façons de parier et d'analyser les cotes proposées au cours du pari. Nous le proposons à un maximum d'individus, parieurs ou non, afin d'avoir un échantillon assez important de réponses et d'évaluer avec plus de précision nos paramètres.

Pour un pari donné à une cote fixée, on propose aux individus interrogés une somme d'argent à miser sur le pari. Il lui est alors demandé de nous donner la probabilité minimale p à partir de laquelle il désire miser sur le pari proposé.

Pour réaliser ce sondage, nous avons envisagé plusieurs possibilités de pari, ce qui nous permet d'étudier le plus de cas possibles : petite ou grosse mise avec petite ou grande cote. Cette diversité de cas nous permettra par la suite d'avoir le maximum de points afin de trouver le plus précisément possible nos deux paramètres.

Voici, le sondage qui nous avons mis en place :

Mise m (en)	Gain potentiel (en)	Probabilité minimale de pari $\bar{p}$
2	4	
2	6	
2	8	
10	15	
10	30	
10	60	
50	70	
50	100	
50	200	
50	300	
100	200	
100	500	

Il s'agit d'une méthode classique qui fournira un grand nombre de données à partir du moment où une cinquantaine de personnes seront interrogées. On espère que les sondés répondront intelligemment aux douze questions, sans aucun calcul d'espérance, mais sans choisir la réponse de façon inconsidérée.

Par l'intermédiaire de notre sondage, nous avons demandé aux personnes interrogées de choisir la probabilité minimale à partir de laquelle ils désiraient parier la mise m donnée. Ainsi, pour la probabilité $\bar{p}$, l'individu est indifférent au fait de miser ou pas sur le pari. De façon mathématique, cela se traduit par $\bar{p}u(Rm, m) + (1-\bar{p})u(-m, m) = u(0)$. En remplaçant la fonction u par nos fonctions d'utilité, on obtient une équation en α et en β. On cherche ensuite à minimiser la somme des $|\bar{p}u(Rm, m) + (1-\bar{p})u(-m, m) = u(0) - u(0)|^2$ pour tous les sondages en fonction des paramètres.

1.3 Calcul des cotes

Considérons un match sur lequel il faut parier, avec des probabilités de résultats p_i, pour i correspondant à la victoire, le match nul et la défaite. On supposera que le bookmaker se laisse une marge et que les parieurs peuvent être modélisés par un representative bettor, dont l'utilité est u. Celle-ci est croissante et vérifie $u(-m,m)$ décroissante. On cherche à déterminer les cotes C_i qui maximisent l'espérance de gain du bookmaker. On notera $R_i = C_i - 1$, et on introduira les cotes réduites $\lambda_i = \dfrac{1-\gamma}{C_i}$, qui vérifient $\sum \lambda_i = 1$. On négligera la distortion des probabilités. Il s'agit de déterminer les λ_i qui maximisent le gain du bookmaker.

1.3.1 Principe général du calcul des cotes

On définit l'espérance d'utilité du parieur pour un pari dont la probabilité de gain est p_i, dont la cote réduit est λ_i et dont la marge est γ, et ce pour une mise m_i, par la fonction $f_{\gamma,\lambda_i,p_i}(m_i) = E(u) = pu^{\gamma,\lambda_i}((\dfrac{1-\gamma}{\lambda_i} - 1)m_i, m_i) + (1-p)u^{\gamma,\lambda_i}(-m,m)$. La mise optimale du parieur s'écrit alors $m^*(\gamma,\lambda_i,p_i) = f^{-1}_{\gamma,\lambda_i,p_i}(\sup_{\mathbb{R}^+} f_{\gamma,\lambda_i,p_i})$, qui, si f_{γ,λ_i,p_i} est concave, s'écrit $(f'_{\gamma,\lambda_i,p_i})^{-1}(0)$.

On définit $g_{\gamma,p}(\lambda) = E(G_{\text{bookmaker}}) = \sum (1 - p_i\dfrac{1-\gamma}{\lambda_i})m^*(\gamma,\lambda_i,p_i)$ l'espérance de gain du bookmaker. On cherche alors à maximiser $g_{\gamma,p}$ par rapport à λ sous la condition $\sum \lambda_i = 1$. On sait que si λ^* est le maximum recherché de $g_{\gamma,p}$ sur $\{\lambda / \sum \lambda_i = 1 \text{et} \forall i, \lambda_i \geq 0\}$, alors $\nabla g_{\gamma,p}(\lambda^*)$ est proportionnel au vecteur normal de la variété dans laquelle se trouvent les λ, c'est-à-dire le vecteur $(1,1,1)$.

On définit donc les dérivées partielles simplifiées $h_i^{\gamma,p}(\lambda_i) = \dfrac{\partial g_{\gamma,p}}{\partial \lambda_i}$, dont on remarque qu'il ne dépend que de λ_i. Toutes les dérivées $h_i^{\gamma,p}$ doivent maintenant être égales et il faut que $\sum \lambda_i = 1$. On suivra la démarche suivante.

En fixant λ_1, on détermine la valeur des $h_i^{\gamma,p}$, qui sera égale à $h_1^{\gamma,p}(\lambda_1)$. En admettant l'injectivité des $h_i^{\gamma,p}$, on en déduit $\lambda_i = (h_i^{\gamma,p})^{-1} \circ h_1^{\gamma,p}(\lambda_1)$, puis la somme des λ_i. Celle-ci est alors une fonction de λ_1, qu'il suffit finalement d'ajuster pour que cette somme vaille 1, et on aura trouvé λ_1^*, puis on pourra en déduire les λ_i^*.

Traduisons maintenant tout cela mathématiquement. On définit donc la somme déduite $S_{\gamma,p}(\lambda_1) = \sum (h_i^{\gamma,p})^{-1} \circ h_1^{\gamma,p}(\lambda_1)$, qui est bien une fonction de λ_1. On a alors $\lambda_1^*(\gamma,p) \in S^{-1}_{\gamma,p}(1)$. Enfin, on a les cotes réduites optimales $\lambda_i^*(\gamma,p) = (h_i^{\gamma,p})^{-1} \circ h_1^{\gamma,p}\lambda_1^*(\gamma,p)$.

1.3.2 Etude théorique des utiliés u_1 etu_2

Rappelons que l'utilité u_1 est définie par $u_1(G,m) = -\exp(-\alpha G - \beta m)$. On rappelle que $R = C - 1$. Posons maintenant $E[u_1] = -p\exp(-\alpha R m - \beta m) - (1-p)\exp(\alpha m - \beta m)$. En annulant la dérivée cette expression par rapport à m, ce qui revient à dire que le joueur maximise son utilité en fonction de sa mise, on a $p(\alpha R+\beta)\exp(-(\alpha R+\beta)m) - (1-p)(\alpha-\beta)\exp((\alpha-\beta)m) = 0$,

ce qui conduit à $m^*(\lambda,p,\gamma)=\frac{1}{\alpha C}\ln(\frac{p}{1-p}\frac{\alpha R+\beta}{\alpha-\beta})=\frac{\lambda}{\alpha(1-\gamma)}\ln(\frac{p}{1-p}\frac{-\frac{1-\gamma}{\lambda}+1+\beta/\alpha}{1-\beta/\alpha})$.

En reprenant les notations de 1.3.1, on a alors $g_{\gamma,p}(\lambda)=\sum(1-p\frac{1-\gamma}{\lambda_i})m^*(\lambda_i,p_i,\gamma)$. On a alors $h_i^{\gamma,p}(\lambda_i)=\frac{1}{\lambda_i}m^*(\lambda_i,p_i,\gamma)+(1-p\frac{1-\gamma}{\lambda_i})\frac{1-\gamma}{1-\gamma+\lambda_i(1+\beta/\alpha)}$.

L'étude d'une telle fonction est délicate, mais on peut tout de même faire quelques remarques. On voit en effet que les fonction $h_i^{\gamma,p}(\lambda_i)$ ne dépendent que du rapport β/α. Ainsi la valeur de λ^* ne dépendra que du rapport entre ces deux coefficients. C'est ce qu'on pourrait appeler l'équilibre de jeu, le rapport entre l'envie de jouer et la peur de perdre, entre l'excitation du jeu et l'aversion au risque.

Par ailleurs, m^* n'est ici définie que si l'expression du logarithme est supérieure à 1. En effet, sinon, la cote n'intéresse pas le joueur qui préfèrera alors ne pas jouer. Cette condition équivaut à $R\geq\frac{1-p}{p}(1-\beta/\alpha)-\beta/\alpha$. Ainsi la cote doit être suffisamment grande et l'équilibre de jeu du parieur doit être suffisamment petite. Il faut aussi que la probabilité de succès du pari soit assez grande. Cependant, on peut noter à nouveau le défaut de ce modèle, puisque le joueur peut vouloir miser, alors que $R=0$, c'est-à-dire qu'il est sûr de perdre de l'argent.

Venons en à l'étude de l'utilité $u_2(G,m)=-\exp(-\alpha G-\beta R/Cm)$. On a alors l'espérance d'utilité du parieur $E[u_2]=-p\exp(-\alpha Rm-\beta R/Cm)-(1-p)\exp(\alpha m-\beta R/Cm)$. En annulant la dérivée, on a $p(\alpha R+\beta R/C)\exp(-(\alpha R+\beta R/C)m)-(1-p)(\alpha-\beta R/C)\exp((\alpha-\beta R/C)m)=0$, d'où $m^*(\lambda,p,\gamma)=\frac{1}{\alpha C}\ln(\frac{p}{1-p}\frac{\alpha R+\beta R/C}{\alpha-\beta R/C})=\frac{\lambda}{\alpha(1-\gamma)}\ln(\frac{p}{1-p}\frac{\frac{1-\gamma}{\lambda}-1+(1-\frac{\lambda}{1-\gamma})\beta/\alpha}{1-(1-\frac{\lambda}{1-\gamma})\beta/\alpha})$.

Comme précédemment, on voit que la formule de $\alpha g_{\gamma,p}$ ne dépend que du rapport α/β. Ainsi, puisque maximiser $\alpha g_{\gamma,p}$ revient à maximiser $g_{\gamma,p}$, le maximum sera atteint par un λ^* qui ne dépendra que de γ, p et du rapport α/β.

La condition de positivité de m^* est ici plus difficile à expliciter. Cependant, en remarquant que $R/C=1-1/C$ est une fonction croissante de C, on voit que le terme à l'intérieur du logarithme est croissant. De plus, pour $\lambda=(1-\gamma)p$, on a ce terme qui est égal à $\frac{p}{q}\frac{1/p-1+(1-p)\beta/\alpha}{1-(1-p)\beta/\alpha}=\frac{1+p\beta/\alpha}{1-q\beta/\alpha}$ qui est bien strictement supérieur à 1 pour $p>0$. Ainsi, quelque soit le match, la mise optimale est strictement positive pour une cote bien choisie. En fait, on peut même dire mieux en remarquant que $(1-\gamma)p<p$ et que le terme à l'intérieur du logarithme est positif, on sait que pour $\lambda=p$, toutes les mises sont strictement positives, et le bookmaker est sûr de pouvoir assurer la positivité de toutes les mises et de son espérance de gain. Celle-ci vaut alors $\sum\gamma m_i^*$.

1.3.3 Etude numérique

Pour mieux comprendre les phénomènes observés et pour faire face à la complexité des formules, nous allons nous servir de scilab pour obtenir des solutions numériques. Nous disposons déjà d'une formule littérale pour l'espérance de gain du bookmaker en fonction des différentes cotes.

Dans un premier temps, nous étudierons un pari à une issue victorieuse. On se donne donc un

bookmaker qui n'offre comme possibilité que le pari sur une issue d'un match. Il ne peut agir que sur les cotes, mais il connaît à la fois la probabilité de cette issue du match et la forme des utilités des joueurs, qui sera d'abord u_1, puis u_2. Son problème est alors de maximiser son espérance de gain en fonction de C. Sur scilab, on se contentera de prendre des valeurs de C régulièrement réparties entre 1 et 50, et pour plusieurs valeurs de p dans $]0,1[$, on choisira la valeur de C qui maximisera l'espérance de gain, puis on tracera la courbe pC en fonction de p. Remarquons au passage que $p \leq pC \leq 1$.

Dans un second temps, on considérera le cadre des paris sportifs sur les matchs de football. On supposera connues de tous les probabilités p_i de victoire des différentes équipes. Par ailleurs, le bookmaker, surveillé par les joueurs ne peut excéder une marge γ. Celle-ci est en effet calculable par les joueurs (elle vaut $1 - 1/(\sum 1/C_i)$) et peut amener ceux-ci à changer de bookmaker s'ils s'aperçoivent que la concurrence en offre une meilleure. De plus, prendre une marge trop faible peut amener à avoir des espérances de gain en fait négatives, à cause des imprécisions sur les connaissances des différents paramètres caractérisant le marché des paris sportifs. Il s'agit en effet d'une sécurité.

Etant donnés les p_i et γ fixés, le bookmaker doit alors maximiser son espérance de gain en fixant les λ_i optimaux. Il s'agit alors pour lui de maximiser une fonction à trois variables, qui évoluent dans le plan d'équation $\sum \lambda_i = 1$ et ayant toutes les coordonnées positives. Sur scilab, on se contentera de prendre un nuage fini de points dans cet espace et de choisir celui qui maximise l'espérance de gain. Pour des problèmes de complexité en temps, on se limitera à une précision de 10^{-2}.

Enfin, et ce sera l'aboutissement de cette étude, on calculera les cotes des matchs de la 34ème journée. Cependant, si on pourra fixer γ de la même manière que précédemment, il nous faudra estimer les probabilités des issues de ces matchs. Il s'agit d'un problème extrêmement difficile, vu les innombrables paramètres entrant en jeu dans un match de football, qui peuvent aller de la météo au nombre de spectateurs, en passant par les risques de blessures des joueurs clés.

Dans le second chapitre de ce rapport, nous allons donc modéliser les matchs de football, et mettre en place des méthodes pour calculer les paramètres de ces modèles.

Chapitre 2

Modélisation des matchs de football

Les bookmakers, pour déterminer au mieux les cotes qu'ils proposent aux parieurs, cherchent à évaluer le plus précisément possible les probabilités des différents résultats des matchs de football. De nombreuses modélisations peuvent être proposées pour approcher ces probabilités. Cependant, un obstacle majeur est constitué par le nombre considérable de facteurs pouvant influer sur le résultat d'un match de football. La difficulté de la modélisation repose sur le choix des paramètres à prendre en compte. Le modèle utilisé ne doit pas être trop simpliste pour pouvoir obtenir des résultats pertinents. Néanmoins, il faut savoir restreindre le nombre de paramètres pour rendre possible une résolution numérique du problème.

Pour une modélisation cohérente, certaines propriétés évidentes des compétitions sportives en général, et des matchs de football en particulier, doivent être prises en considération dans les calculs. Notamment, il semble nécessaire de :

- faire intervenir les caractéristiques des 2 équipes opposées ;

- distinguer les forces offensives (capacité à marquer des buts) et défensives (capacité à ne pas encaisser de buts) des équipes qui s'affrontent ;

- traduire l'observation empirique d'un avantage pour l'équipe qui joue à domicile par rapport à celle qui joue à l'extérieur ;

- considérer les résultats récents des équipes considérées (périodes de forme ou de méforme) ;

- analyser ces derniers résultats en tenant compte du niveau des équipes rencontrées.

Nous allons, dans cette partie, présenter les modèles que nous avons utilisés, en nous appuyant sur les documents que nous avons consultés. Dans un premier temps, nous considérerons que les facteurs influant sur les résultats des matchs sont constants au cours de la saison. Ensuite, nous nous intéresserons à des modélisations pour lesquelles ces mêmes facteurs évoluent avec le temps.

2.1 Loi de Poisson

Notre base de travail a été le modèle de Maher datant de 1982 et présenté dans l'article "Modelling association football scores and inefficiencies in the football betting market" rédigé par Dixon et Coles[4]. Cette modélisation repose sur l'hypothèse que les nombres de buts inscrits par les équipes évoluant à domicile et à l'extérieur sont des variables de Poisson indépendantes. Les moyennes de ces variables sont déterminées à partir des caractéristiques offensives et défensives des deux équipes qui s'affrontent. Il faut de plus prendre en considération l'avantage représenté par le fait de jouer à domicile.
Pour un match opposant les équipes i et j, on note $X_{i,j}$ et $Y_{i,j}$ les nombres de buts respectivement inscrits par l'équipe jouant à domicile et l'équipe évoluant à l'extérieur.
On a alors :

$$X_{i,j} \equiv \text{Poisson } (\alpha_i \beta_j \gamma)$$
$$Y_{i,j} \equiv \text{Poisson } (\alpha_j \beta_i)$$

où $X_{i,j}$ et $Y_{i,j}$ sont des variables indépendantes.
Les coefficients α_i et α_j représentent respectivement les forces offensives des équipes i et j.
Les β_i et β_j sont les faiblesses défensives de ces mêmes équipes.
Le paramètre γ traduit l'avantage de jouer à domicile.

On en déduit la probabilité que l'équipe à domicile marque x buts :

$$P(X_{i,j} = x) = \frac{\lambda^x exp(-\lambda)}{x!}$$

et la probabilité que l'équipe à l'extérieur marque y buts :

$$P(Y_{i,j} = y) = \frac{\mu^y exp(-\mu)}{y!}$$

L'indépendance des variables $X_{i,j}$ et $Y_{i,j}$ assure que

$$P(X_{i,j} = x, Y_{i,j} = y) = P(X_{i,j} = x)\ P(Y_{i,j} = y).$$

Ainsi, pour un match opposant les équipes i et j sur le terrain de l'équipe i, la probabilité pour que le score soit de x buts inscrits par l'équipe i et y buts inscrits par l'équipe j est :

$$P(X_{i,j} = x, Y_{i,j} = y) = \frac{\lambda^x exp(-\lambda)}{x!} \frac{\mu^y exp(-\mu)}{y!}$$

où

$$\lambda = \alpha_i \beta_j \gamma$$
$$\mu = \alpha_j \beta_i$$

avec γ l'avantage domicile,
α_i, α_j les forces respectives des attaques des équipes i et j,
β_i, β_j les forces respectives des défenses des équipes i et j.

2.2 Maximum de vraisemblance

2.2.1 Principe

Le modèle développé repose alors sur les paramètres $\{\alpha_1, ..., \alpha_n\}$, $\{\beta_1, ..., \beta_n\}$ et γ qu'il nous faut déterminer.
On considère un championnat à n équipes. Nous avons ainsi 2n+1 coefficients à calculer : les forces offensives des n équipes $\{\alpha_1, ..., \alpha_n\}$, les faiblesses défensives des n équipes $\{\beta_1, ..., \beta_n\}$ et l'avantage domicile-extérieur γ.
Pour obtenir ces paramètres, on s'appuye sur la méthode du maximum de vraisemblance.
On cherche à maximiser la fonction de vraisemblance suivante :

$$F(\{\alpha_1, ..., \alpha_n\}, \{\beta_1, ..., \beta_n\}, \gamma) = \prod_{k=1}^{N} \frac{\lambda_k^{x_k} exp(-\lambda_k)}{x_k!} \frac{\mu_k^{y_k} exp(-\mu_k)}{y_k!}$$

où

$$\lambda_k = \alpha_{i(k)} \beta_{j(k)} \gamma$$
$$\mu_k = \alpha_{j(k)} \beta_{i(k)}$$

Ici, les indices $i(k)$ et $j(k)$ se réfèrent aux équipes évoluant à domicile et à l'extérieur lors du match k.
Pour obtenir les paramètres qui nous intéressent, nous devons imposer une condition supplémentaire, sans laquelle nous manquons d'équations pour faire aboutir la résolution numérique du problème.

La condition choisie dans les articles que nous avons consultés est la suivante :

$$\frac{\sum_{i=1}^{n} \alpha_i}{n} = 1$$

Elle consiste à contraindre les coefficients α_i à avoir une moyenne égale à 1. Nous avons, pour notre part, préféré choisir une équipe i et fixer sa force offensive α_i à 1 (en pratique, nous avons opté pour la dernière équipe du classement).

On souhaite maximixer la fonction de vraisemblance, ce qui revient à maximiser son logarithme. On égalise à 0 les dérivées partielles de $ln(F)$ par rapport aux divers coefficients que nous cherchons, d'où :

$$\frac{\partial(lnF)}{\partial\gamma} = 0$$
$$\frac{\partial(lnF)}{\partial\alpha_i} = 0, \text{ pour i} = 1, ..., n$$
$$\frac{\partial(lnF)}{\partial\beta_i} = 0, \text{ pour i} = 1, ..., n$$

Ceci fournit un sytème d'équations dont la résolution mène aux paramètres recherchés.

Pour nous familiariser avec les calculs, nous avons voulu expérimenter différentes méthodes de résolution sur ce modèle relativement simple où les paramètres ne dépendent pas du temps.

Les valeurs calculées informatiquement ont été déterminées à partir des données du championnat de ligue 1 de la saison 2007/2008.
La résolution numérique nous a amené à fixer la valeur de l'un des α_i afin de permettre la convergence des calculs. Nous avons choisi d'imposer $\alpha = 1$ pour la dernière équipe au classement (Metz).
Nous avons employé deux méthodes distinctes de résolution numérique pour mener à bien les calculs : la méthode de Newton et une autre méthode reposant sur un processus d'itérations. Nous les détaillons ci-dessous.

2.2.2 Méthode par itérations

Voici un descriptif rapide de la résolution numérique mise en oeuvre :
- on initialise tous les α et β à 1
- on calcule les α à partir de ces valeurs initiales
- on détermine les β à partir de ces nouvelles valeurs de α
- on réitère ce processus 10000 fois.

Dans un premier temps, on pose $\gamma = 1$. Cela revient à considérer qu'il n'y a aucun avantage pour l'équipe qui joue à domicile par rapport à celle qui joue à l'extérieur.

$\frac{\partial(lnF)}{\partial\alpha_i} = 0,$ pour i $= 1, ..., n$ donne

$$\alpha_i = \frac{\sum_{j\neq i}(x_{i,j} + y_{j,i})}{2\sum_{j\neq i}\beta_j},$$

avec $\sum_{j\neq i}(x_{i,j} + y_{j,i})$ = total des buts marqués par l'équipe i au cours de la saison.

De même, $\frac{\partial(lnF)}{\partial\beta_i} = 0,$ pour i $= 1, ..., n$ donne

$$\beta_i = \frac{\sum_{j\neq i}(x_{j,i} + y_{i,j})}{2\sum_{j\neq i}\alpha_j},$$

avec $\sum_{j\neq i}(x_{j,i} + y_{i,j})$ = total des buts encaissés par l'équipe i au cours de la saison.

Les résultats obtenus sont récapitulés dans le tableau suivant.

Equipe	α	β
Lyon	2.476	0.693
Bordeaux	2.143	0.703
Marseille	1.958	0.827
Nancy	1.457	0.541
Saint-Etienne	1.565	0.616
Rennes	1.548	0.778
Lille	1.494	0.578
Nice	1.159	0.536
Le Mans	1.560	0.887
Lorient	1.066	0.623
Caen	1.636	0.963
Monaco	1.354	0.863
Valenciennes	1.408	0.720
Sochaux	1.143	0.767
Auxerre	1.122	0.927
Paris	1.247	0.806
Toulouse	1.209	0.751
Lens	1.463	0.938
Strasbourg	1.160	0.982
Metz	1	1.136

On peut alors, à partir de ces coefficients déterminer les probabilités de résultats de tous les matchs de la saison. Il est dès lors possible de simuler le championnat en calculant les espérances de points obtenus par chacune des équipes, ce qui fournit le classement ci-dessous.

	Classement réel	Pts	Classement simulé	Espérance de pts
1	Lyon	79	Lyon	75.8889776725218
2	Bordeaux	75	Bordeaux	69.5552866465969
3	Marseille	62	Nancy	62.089118623988
4	Nancy	60	Lille	61.2739512009293
5	Saint-Etienne	58	Saint-Etienne	61.2241739402477
6	Rennes	58	Marseille	60.9442760989173
7	Lille	57	Nice	54.8072167153191
8	Nice	55	Rennes	54.0000177242932
9	Le Mans	53	Valenciennes	53.0608648804903
10	Lorient	52	Le Mans	50.0404960361977
11	Caen	51	Caen	48.9796030201192
12	Monaco	47	Lorient	48.4122825503966
13	Valenciennes	45	Toulouse	46.9107366656643
14	Sochaux	44	Monaco	46.1824557975569
15	Auxerre	44	Lens	45.9714319304502
16	Paris	43	Paris	45.7434120297212
17	Toulouse	42	Sochaux	44.5995528970637
18	Lens	40	Auxerre	38.2692914713799
19	Strasbourg	35	Strasbourg	37.3732469387946
20	Metz	24	Metz	28.9656673477273

On peut également calculer l'espérance de différence de buts (buts marqués - buts encaissés)

pour chaque équipe, grâce aux α et β qui déterminent les probabilités de tous les scores possibles pour chacun des matchs se déroulant au cours du championnat.

	Equipe	Différence de buts	Espérance de différence de buts
1	Lyon	+37	+36.999997882289
2	Bordeaux	+27	+26.0000004060364
3	Marseille	+13	+13.0000009730452
4	Nancy	+14	+13.9999993548958
5	Saint-Etienne	+13	+12.9999994010965
6	Rennes	+3	+2.99999787546079
7	Lille	+13	+13.0000013719479
8	Nice	+5	+4.99999880100142
9	Le Mans	−3	−2.99999920523265
10	Lorient	−3	−2.99999764333521
11	Caen	−5	−4.99999846791355
12	Monaco	−8	−7.99999991230347
13	Valenciennes	+2	+2.00000278647357
14	Sochaux	−9	−9.00000022748738
15	Auxerre	−19	−19.0000007537003
16	Paris	−8	−7.99999709172313
17	Toulouse	−6	−6.00000120709041
18	Lens	−9	−8.9999996838552
19	Strasbourg	−21	−21.0000027122837
20	Metz	−36	−35.0000019473215

On prend désormais en compte l'avantage domicile-extérieur.
On a alors :

$$\alpha_i = \frac{\sum_{j\neq i}(x_{i,j} + y_{j,i})}{(\gamma+1)\sum_{j\neq i}\beta_j}$$

et

$$\beta_i = \frac{\sum_{j\neq i}(x_{j,i} + y_{i,j})}{(\gamma+1)\sum_{j\neq i}\alpha_j}.$$

Par ailleurs, l'équation $\frac{\partial(lnF)}{\partial\gamma} = 0$ nous permet d'obtenir γ :

$$\gamma = \frac{\sum_{i,j}(x_{i,j} + y_{i,j})}{\sum_{i\neq j}\alpha_i\beta_j}.$$

Le calcul donne :

$$\gamma = 1.246.$$

Equipe	α	β
Lyon	2.476	0.617
Bordeaux	2.143	0.626
Marseille	1.958	0.736
Nancy	1.457	0.482
Saint-Etienne	1.565	0.548
Rennes	1.548	0.693
Lille	1.494	0.515
Nice	1.159	0.477
Le Mans	1.559	0.790
Lorient	1.066	0.554
Caen	1.636	0.857
Monaco	1.354	0.768
Valenciennes	1.408	0.642
Sochaux	1.143	0.683
Auxerre	1.122	0.825
Paris	1.247	0.718
Toulouse	1.209	0.669
Lens	1.463	0.836
Strasbourg	1.160	0.874
Metz	1	1.012

	Classement réel	Pts	Classement simulé	Espérance de pts
1	Lyon	79	Lyon	75.5796114098147
2	Bordeaux	75	Bordeaux	69.3502615218128
3	Marseille	62	Nancy	62.0091716300348
4	Nancy	60	Lille	61.2019007119839
5	Saint-Etienne	58	Saint-Etienne	61.1510486192699
6	Rennes	58	Marseille	60.8657683789047
7	Lille	57	Nice	54.80507613283
8	Nice	55	Rennes	54.0212311520975
9	Le Mans	53	Valenciennes	53.0894883375865
10	Lorient	52	Le Mans	50.1215144469087
11	Caen	51	Caen	49.0831455249998
12	Monaco	47	Lorient	48.4762560403243
13	Valenciennes	45	Toulouse	47.0075041163795
14	Sochaux	44	Monaco	46.3044735381267
15	Auxerre	44	Lens	46.1081007972635
16	Paris	43	Paris	45.8602435852942
17	Toulouse	42	Sochaux	44.7197500432291
18	Lens	40	Auxerre	38.4738353437997
19	Strasbourg	35	Strasbourg	37.5978150530296
20	Metz	24	Metz	29.2743480502660

	Equipe	Différence de buts	Espérance de différence de buts
1	Lyon	+37	+36.9999986506995
2	Bordeaux	+27	+25.999996832861
3	Marseille	+13	+13.0000027072299
4	Nancy	+14	+14.0000027428733
5	Saint-Etienne	+13	+13.0000007039101
6	Rennes	+3	+2.99999877099087
7	Lille	+13	+12.9999967722968
8	Nice	+5	+4.99999952331815
9	Le Mans	−3	−3.00000084718429
10	Lorient	−3	−3.00000039485386
11	Caen	−5	−4.99999723973073
12	Monaco	−8	−7.99999814999758
13	Valenciennes	+2	+2.00000319188020
14	Sochaux	−9	−8.99999635682712
15	Auxerre	−19	−18.9999982872925
16	Paris	−8	−8.00000102446256
17	Toulouse	−6	−5.99999790309383
18	Lens	−9	−9.000004057119
19	Strasbourg	−21	−21.0000051628133
20	Metz	−36	−35.0000004726852

2.2.3 Méthode de Newton

De même que pour la méthode précédente on initialise les α et β à 1.
On dispose de 40 équations portant sur les 40 variables α et β.
On note f le vecteur de taille 40 de coordonnées f_i telles que :

$$f_i = \begin{cases} \alpha_i \sum_{j \neq i} \beta_j - \dfrac{\sum_{j \neq i}(x_{i,j} + y_{j,i})}{2} & \text{pour i = 1,...,20} \\ \beta_i \sum_{j \neq i} \alpha_j - \dfrac{\sum_{j \neq i}(x_{j,i} + y_{i,j})}{2} & \text{pour i = 21,...,40} \end{cases}$$

L'utilisation de la méthode de Newton repose sur la résolution de l'équation $f(x + dx) = 0$. Ceci donne $dx = -f(x)grad(f)^{-1}$. On réitère le processus en remplaçant x par x + dx, et ainsi de suite jusqu'à 10000 itérations.

Comme précédemment, dans un premier temps, on ne tient pas compte du facteur domicile-extérieur ($\gamma = 1$).

Equipe	α	β
Lyon	3.268	0.526
Bordeaux	2.829	0.533
Marseille	2.585	0.627
Nancy	1.924	0.411
Saint-Etienne	2.065	0.467
Rennes	2.043	0.590
Lille	1.972	0.439
Nice	1.530	0.406
Le Mans	2.059	0.673
Lorient	1.406	0.465
Caen	2.159	0.730
Monaco	1.787	0.655
Valenciennes	1.858	0.546
Sochaux	1.509	0.582
Auxerre	1.481	0.703
Paris	1.647	0.611
Toulouse	1.596	0.570
Lens	1.931	0.712
Strasbourg	1.531	0.745
Metz	1	0.854

En ce qui concerne l'application de la méthode de Newton au cas $\gamma \neq 1$, nous n'avons pas encore obtenu de résultats cohérents. Cette seconde méthode étant de plus d'un maniement plus complexe, nous nous limiterons par la suite à l'utilisation de la méthode par itérations.

Cependant, il est intéressant de comparer les résultats obtenus par les deux méthodes de calcul, pour le cas où l'on néglige l'avantage domicile-extérieur ($\gamma = 1$). On constate que les valeurs des α et β diffèrent selon l'algorithme utilisé. Cependant, il est essentiel de noter l'égalité entre les produits $\alpha_i\beta_j$ dans les deux cas. En effet, seul ce produit représente un intérêt concret dans les calculs probabilistes effectués. Les valeurs des α et β ne revêtent, dans nos modèles, aucune importance en elles-mêmes.

Par ailleurs, on peut remarquer que, pour la première méthode, lorsque γ n'est pas fixé, il prend une valeur strictement supérieure à 1. Ainsi, l'espérance des buts marqués par les équipes à domicile augmente et celle des buts inscrits à l'extérieur diminue. Ceci traduit donc bien l'avantage empirique observé pour les équipes qui jouent sur leur terrain.

Dans les différents cas présentés ci-dessus, on note que les classements simulés a posteriori ne correspondent pas tout à fait au classement réel. En revanche, les espérances de buts sont nettement plus précises (elles sont en fait identiques aux données réelles à plus ou moins un but près). Ceci provient du fait que la méthode de calcul repose sur la prise en compte des scores de tous les matchs joués et non des seuls résultats (victoire, match nul ou défaite).

Le principal inconvénient de ce modèle est qu'il ne repose que sur des paramètres d'attaque et de défense constants au cours du temps (que ce soit au cours d'un même match ou au sur la durée d'une saison).

2.3 Variation des paramètres au cours de la saison

On se réfère dans cette partie au document "Modelling association football scores and inefficiencies in the football betting market" de Dixon et Coles[4].

2.3.1 Intérêt

La limite du modèle précédent est qu'il prend en compte toutes les journées de la même façon. Par exemple, si nous sommes à une journée proche de la fin du championnat, la configuration suivante peut apparaître :

$$\begin{cases} \text{Equipe A : bonnes performances au début de la saison et mauvaises à la fin} \\ \text{Equipe B : mauvaises performances au début de la saison et bonnes à la fin} \end{cases}$$

On voit immédiatemment que les équipes A et B auront des paramètres α et β proches, ce qui ne rend pas compte du fait que B est en forme et que A connaît une période difficile. Ce modèle n'est donc pas satisfaisant pour l'obtention de probabilités réalistes. Dans notre deuxième modèle, nous allons tenir compte de cela : l'importance des résultats d'une journée de championnat diminuera avec l'éloignement dans le temps. Ainsi, une équipe sera peu pénalisée par un mauvais début de saison si elle se rattrape ensuite. A l'inverse, une équipe restant sur une mauvaise série verra sa force diminuer.

Pour remédier à la limitation détaillée ci-dessus, nous allons donc faire varier les coefficients α_i et β_i (correspondant aux caractéristiques offensives et défensives des différentes équipes engagées) au cours de la saison.

Cette amélioration peut être apportée au modèle présenté plus haut en modifiant la fonction de vraisemblance afin de faire intervenir les résultats des matchs précédents dans la détermination des paramètres. L'idée consiste à inclure les résultats passés avec une importance croissante avec la proximité dans le temps avec le match considéré. Le poids apporté à cette caractéristique temporelle peut être plus ou moins fort selon les fonctions correctives choisies. On passe alors à une évolution dynamique des paramètres au cours de la saison.

On fait intervenir une fonction de poids qui donne une importance supérieure aux résultats des matchs récents dans le calcul du maximum de vraisemblance.

La fonction de vraisemblance prend alors la forme suivante :

$$F(\{\alpha_1,...,\alpha_n\},\{\beta_1,...,\beta_n\},\gamma) = \prod_{A_t}\left(\frac{\lambda_k^{x_k}exp(-\lambda_k)}{x_k!}\frac{\mu_k^{y_k}exp(-\mu_k)}{y_k!}\right)^{\frac{\phi(t-t_k)}{\sum_k \phi(t-t_k)}}$$

avec t_k le moment de la saison auquel le match k a eu lieu, ϕ une fonction non-croissante de t et $A_t = \{k; t_k < t\}$. On a par ailleurs toujours $\lambda_k = \alpha_{i(k)}\beta_{j(k)}\gamma$ et $\mu_k = \alpha_{j(k)}\beta_{i(k)}$

Plaçons-nous à un instant t.

La fonction F prend en considération tous les résultats de tous les matchs ayant eu lieu avant ce temps t. L'application du principe du maximum de vraisemblance sur cette fonction fournit

alors des valeurs pour les différents paramètres $\{\alpha_1, ..., \alpha_n\}$ et $\{\beta_1, ..., \beta_n\}$ à cet instant t de la saison.

On peut donc calculer ces paramètres à tout instant de la saison. Ils connaissent désormais une évolution dynamique. Avec cette nouvelle modélisation, la force des équipes varie au cours du championnat, ce qui est plus cohérent avec la réalité. En effet, les périodes de forme et de méforme des différentes équipes en compétition sont ainsi prises en compte.

Le choix de la fonction ϕ détermine l'importance plus ou moins grande accordée par les résultats récents, par rapport aux résultats plus anciens, dans la détermination des α_i et β_i. On peut jouer sur le poids plus ou moins grand à apporter aux précédents résultats en faisant varier cette fonction.

2.3.2 Pondération exponentielle

Dans un premier temps, nous avons suivi la méthode étudiée par Dixon et Coles dans leur article. Elle consiste à poser une fonction $\phi(t)$ de la forme $\phi(t) = exp(-\xi t)$.

On obtient alors :

$$ln(F(\{\alpha_1, ..., \alpha_n\}, \{\beta_1, ..., \beta_n\}, \gamma)) = \frac{\sum_{A_t} \phi(t - t_k) ln(\frac{\lambda_k^{x_k} exp(-\lambda_k)}{x_k!} \frac{\mu_k^{y_k} exp(-\mu_k)}{y_k!})}{\sum_{A_t} \phi(t - t_k)}$$

Tous les paramètres (force d'attaque α et faiblesse de défense β de chaque équipe, avantage local γ et coefficient ξ) sont ensuite calculés par la méthode du maximum de vraisemblance. Comme précédemment, on égalise les dérivées par rapport des différents paramètres à 0 pour obtenir un système d'équations.

$\frac{\partial(lnF)}{\partial \alpha_i} = 0$, pour i $= 1, ..., n$ donne

$$\alpha_{i(k)} = \frac{\sum_{\text{matchs}} bm_{i(k)} \phi(t - t_k))}{\sum_{\text{domicile}} \gamma \beta_{j(k)} + \sum_{\text{extérieur}} \beta_{j(k)}},$$

avec $bm_{i(k)}$ le nombre de buts marqués par l'équipe i lors de la journée de date t_k.

De même, $\frac{\partial(lnF)}{\partial \beta_i} = 0$, pour i $= 1, ..., n$ donne

$$\beta_{i(k)} = \frac{\sum_{\text{matchs}} be_{i(k)} \phi(t - t_k))}{\sum_{\text{domicile}} \alpha_{j(k)} + \sum_{\text{extérieur}} \gamma \alpha_{j(k)}},$$

avec $be_{i(k)}$ le nombre de buts encaissés par l'équipe i lors de la journée de date t.

Dans les deux expressions précédentes, les termes "matchs","domicile" et "extérieur" qui interviennent dans les sommes se réfèrent respectivement à l'ensemble des matchs disputés par l'équipe i, à tous tous les matchs qu'elle a joués à domicile et à tous ceux qu'elle a joués à l'extérieur.

Par ailleurs, l'équation $\frac{\partial(lnF)}{\partial \gamma} = 0$ nous permet d'obtenir γ :

$$\gamma = \frac{\sum_{\text{matchs}} \phi(t - t_k) x_k}{\sum_{\text{matchs}} \alpha_{i(k)} \beta_{j(k)} \phi(t - t_k)}.$$

Ici, les sommes s'effectuent sur l'ensemble des matchs disputés depuis le début de la saison jusqu'à la date t_k. L'indice i concerne l'équipe évoluant à domicile et l'indice j se réfère à celle qui joue à l'extérieur.

On applique ensuite la méthode d'itération :

On initialise les α_i, les β_i et γ à 1. On souhaite calculer ξ à partir de ces valeurs, recalculer les coefficient à partir de ce ξ, recalculer ξ et ainsi de suite.

Le problème rencontré avec ce modèle est qu'il est trop compliqué de calculer le coefficient ξ de manière exacte(pour la n ième journée on trouve un polynôme d'ordre n en exp(-ξ)). Nous avons alors tenté de contourner cette difficulté par le biais d'une approximation.

Dans l'article de Dixon et Coles sur lequel nous nous sommes appuyés, le coefficient ξ calculé était assez faible (de l'ordre de $6.5*10^{-3}$). Nous avons fait un développement limité à l'ordre 1 en $\xi*(t-t(k))$ autour de 0 du terme exponentiel intervenant dans l'expression de la dérivée $\frac{\partial(lnF)}{\partial\gamma}$. En maximisant la vraisemblance pondérée lors de la 33ème journée (t = 33), nous trouvons un ξ de valeur 0.035. Dès lors, pour t(k)=1 (ce qui revient à considérer un match s'étant déroulé à la première journée), on trouve $\xi * (t - t(k)) = 1.12$. Le développement limité que nous avions effectué devient ainsi injustifié.

Nous avons alors eu l'idée d'opter pour un nouveau modèle, que nous présentons dans la section suivante.

2.3.3 Pondération polynômiale

Nous choisissons désormais :

$$\phi(t) = 1 - \xi t + \frac{\xi^2 t^2}{2}.$$

Cette fonction polynômiale d'ordre 2 en ξ correspond au développement limité à l'ordre 2 en 0 de la fonction exponentielle utilisée précédemment.

Ce modèle se rapproche du modèle précédent, sauf que nous n'avons pas besoin de faire un développement limité pour calculer le ξ, qui s'obtient en résolvant une équation du second degré.

La fonction de vraisemblance mène à :

$$\ln(F(\{\alpha_1,...,\alpha_n\},\{\beta_1,...,\beta_n\},\gamma)) = \frac{\sum_{A_t}(1-\xi(t-t_k)+\frac{\xi^2(t-+t_k)^2}{2})\ln(\frac{\lambda_k^{x_k}exp(-\lambda_k)}{x_k!}\frac{\mu_k^{y_k}exp(-\mu_k)}{y_k!})}{\sum_{A_t}(1-\xi(t-t_k)+\frac{\xi^2(t-t_k)^2}{2})}$$

Les expressions des α, β et γ restent identiques à celles qui ont été présentées plus haut. En revanche, le calcul de ξ est désormais réalisable de manière exacte. Ce paramètre est obtenu à chaque itération en résolvant l'équation du second degré suivante :

$$10t - \sum(t-t_k)p(k) + 5t(t+1)\sum p(k) + \xi(\sum(t-t_k)^2 10t - \frac{10}{6}t(t+1)(2t+1)\sum p(k))$$

$$+\xi^2(-\frac{5}{2}t(t+1)\sum(t-t_k)^2p(k)) + \frac{5}{6}t(t+1)(2t+1)\sum(t-t_k)p(k) = 0$$

Nous avons testé ce calcul sur la saison actuelle du championnat de France après la 33ème journée. Cela nous a fourni $\xi = 0.041$ et des valeurs intéressantes pour les α_i, β_i et γ (les résultats complets sont détaillés dans la troisième partie du rapport). Ce modèle nous permet donc d'obtenir tous les coefficients que nous cherchons. C'est celui-ci que nous avons retenu pour notre calcul des différentes probabilités de résultats des matchs de la prochaine matchs journée de championnat.

Nous avons pu calculer les coefficients de toutes les équipes, donc leurs probabilités de victoire, nul et défaite pour leur prochain match à jouer. Ceci nous a ensuite permis de calculer nos propres cotes, à partir de ces probabilités.

Chapitre 3

Résultats de l'étude et commentaires

Nous disposons maintenant de modèles du marché qui expliqueront comment le prix des paris agit sur les quantités de ventes, c'est-à-dire comment la valeur des cotes agit sur les sommes misées. Cette transaction dépend cependant d'un paramètre incontournable qui est la probabilité de victoire du pari. On peut maintenant calculer ce prix, ou cette cote. Il nous faut maintenant déterminer des valeurs numériques.

3.1 Résultats numériques de la modélisation des matchs de football

3.1.1 Calcul des coefficients

On s'intéresse désormais uniquement au modèle où les paramètres α,β et γ varient au cours de la saison. Comme expliqué auparavant, la fonction de poids utilisée est la fonction

$$\phi(t) = 1 - \xi t + \frac{\xi^2 t^2}{2}.$$

Tous les résultats sont obtenus par la méthode d'itérations.

Nous avons appliqué cette méthode plus complexe et plus intéressante aux résultats du championnat de France de ligue 1 de cette saison, en prenant en compte tous les résultats jusqu'à la 33 ème journée. Nous avons déterminé les valeurs des paramètres α,β,γ et ξ à cet instant de la saison. Les résultats apparaissent dans ce tableau :

Equipe	α	β
Lyon	1.399209	0.5388168
Bordeaux	1.867123	0.6447525
Marseille	1.872845	0.5749719
Nancy	1.146854	0.8368348
Saint-Etienne	1.073224	1.039977
Rennes	1.190825	0.6578363
Lille	1.492541	0.6971205
Nice	1.304901	0.7684826
Le Mans	1.276565	1
Lorient	1.353568	0.8082447
Caen	1.211238	0.9264272
Monaco	1.365388	0.8102413
Valenciennes	0.9704916	0.6905502
Sochaux	1.165126	0.8758345
Auxerre	0.9806398	0.6543674
Paris	1.548529	0.6790789
Toulouse	1.299670	0.5009974
Grenoble	0.8167448	0.6829007
Nantes	0.9512097	0.9079237
Le Havre	1	1.256788

Les autres paramètres prennent les valeurs $\gamma = 1.305162$ et $\xi = 0.041$.

A partir de ces données, il nous a été possible de calculer les probabilités de résultats pour tous les matchs devant se jouer lors de la 34ème journée de championnat. On a réuni les résultats dans le tableau ci-dessous. Le classement des équipes avant le match est indiqué entre parenthèses.

Match	1	X	2
Valenciennes (15) - Lyon (3)	0.249279	0.341607	0.409114
Lorient (12) - Lille (6)	0.3690499	0.2740484	0.3569017
Grenoble (11) - Nice (8)	0.5329882	0.2659640	0.2010478
Monaco (9) - Auxerre (10)	0.4450869	0.3046957	0.2502175
Saint- Etienne (18) - Nancy (13)	0.3554773	0.2790275	0.3654952
Caen (19) - Le Mans (14)	0.4657449	0.2489923	0.2852628
Marseille (1) - Toulouse (5)	0.4749066	0.2988531	0.2262403
Paris (4) - Rennes (7)	0.4894855	0.2830069	0.2275076
Nantes (17) - Le Havre (20)	0.5471717	0.2381761	0.2146522
Bordeaux (2) - Sochaux (16)	0.6292742	0.2104882	0.1602376

En supposant les paramètres α,β et γ constants pour les 5 dernières journées restant à jouer, il nous est possible de simuler la fin du championnat. En effet, ces différents coefficients nous donnent accès aux probabilités de réalisation de tous les prochains matchs à jouer.

3.1.2 Simulation de la fin du championnat

Grâce à nos modèles, nous avons réussi à simuler le classement final du championnat de France de football. En effet, on calcule tous nos paramètres à l'aide des résultats de toutes les journées avant la 34^{me} journée. On obtient alors tous nos α et β pour toutes nos équipes. A partir

de la 34^{me} journée, pour faire nos calculs, on va considérer nos paramètres comme constants. Pour chaque journée, notre modèle nous fournit alors nos probabilités sur chacun des événements suivants : l'équipe considérée gagne, il y a match nul ou l'équipe considérée perd. A partir de ces probabilités, on en déduit alors une espérence de points qui sera gagné par une équipe donnée en considérant le fait que la victoire rapporte 3 points, le match nul, 1 point et la défaite, 0 point.

On en déduit que l'espérence de point gagnée par une équipe B sur une journée sera : $E = p(\text{victoire}) * 3 + p(\text{nul}) * 1 + p(\text{defaite}) * 0$.

Voici le classement que l'on trouve :

Equipe	Espérence de points
1.Marseille	76.1
2.Bordeaux	75.1
3.Lyon	69.0
4.Paris S.G	68.6
5.Toulouse	64.5
6.Lille	64.1
7.Rennes	57.7
8.Nice	53.9
9.Auxerre	49.8
10.Monaco	49.3
11.Grenoble	47.2
12.Lorient	46.9
13.Nancy	44.7
14.Le Mans	43.2
15.Valenciennes	42.7
16.Sochaux	39.1
17.Nantes	38.8
18.Saint-Etienne	37.8
19.Caen	36.7
20.Le Havre	25.2

Arrivé à ce stade, nous nous devons de faire quelques remarques quant à ce résultat :

Marseille champion mais...

Après la 33ème journée, Marseille est premier avec 67 points et 2 points d'avance sur Bordeaux, deuxième. Notre modèle prévoit que Marseille va être champion et va gagner 9 points sur 5 matchs, mais vu le calendrier beaucoup plus difficile de Marseille comparé à celui de Bordeaux, il est prévu que l'écart entre les deux équipes va se réduire à 1 point, ce qui semble tout à fait logique.

Lyon en difficulté

Malgré son calendrier pas très difficile pour les cinq dernières journées (Valenciennes, Nantes et Caen) qui sont des équipes en bas de classement, notre modèle ne prévoit que 8 points sur 15 possibles pour Lyon. On peut expliquer cela au fait que notre modèle acccorde plus d'importance aux matchs récents et que Lyon vit une période difficile accumulant les contre-performances, même contre des équipes de bas de classement.

Saint-Etienne en ligue 2 !

A la 33^{me} journée, Nantes devance Saint-Etienne de seulement un point. Selon notre modèle, cet écart de un point sera conservé à la fin de la saison. Cependant, on peut se montrer assez interrogatif sur ce résultat. En effet, le calendrier de Nantes semble beaucoup plus difficile que celui de Saint-Etienne avec notamment une rencontre contre Lyon. Le modèle se base sûrement sur la très grande forme des équipes qui seront opposées à Saint-Etienne durant ces trois prochaines semaines, à savoir Toulouse, Auxerre et Valenciennes, même si ces dernières semblent à première vue de niveau modeste.

Remarques de dernières minutes

Marseille vient de faire un match nul contre Toulouse et ne marque donc qu'un seul point sur cette journée, ce qui est inférieur à l'espérance de point que notre modèle lui attribuait pour ce match. Dans le même temps, Bordeaux a gagné. Si on calculait de nouveau les coefficients actualisés après la 34ème journée, on trouvera sûrement que Bordeaux serait champion.

Lyon perd encore, et cette fois-ci contre Valenciennes, ce qui met en évidence la période très difficile pour cette équipe et affaiblit encore ses coefficients actualisés à la 34ème journée. Lyon sera donc sûrement quatrième derrière le Paris Saint Germain, si le club de la capitale venait à s'imposer face à Rennes.

3.2 Calcul numérique des cotes

3.2.1 Résultats du sondage

Le sondage a été réalisé auprès de 100 étudiants de l'école. L'échantillon est assez important en nombre, ce qui devrait nous fournir des résultats précis. Remarquons également que la population ciblée est très spécifique (âge et études similaires), et que les personnes interrogées n'ont pas forcément pour habitude de parier de l'argent. Pour une étude plus globale du comportement des parieurs, la méthode à adopter resterait identique, mais il conviendrait donc de considérer une population plus diversifiée. C'est pourquoi notre sondage a plutôt pour ambition de donner une méthode générale pour calculer les coefficients, que de donner des résultats ayant une valeur générale.

Nous utilisons le logiciel de calcul R afin de déterminer les deux paramètres α et β, par une méthode des doubles moindres carrés(voir annexe 2).

Le calcul dans R nous donne $\alpha = 0.11$ et $\beta = 0.07$ pour l'utilité u_1 et $\alpha = 0.07$ et $\beta = 0.06$ pour l'utilité u_2. On tient à préciser que les valeurs obtenues ne sont pas des maximums globaux. En effet, on se devait de prendre un couple (α, β) qui répondait à la contrainte $\alpha > \beta$. Or, pour la plupart des valeurs initiales de (α, β) que l'on rentrait, le programme nous donnait des valeurs de β trop élevées par rapport à ce qu'on attendait. Il est évident que ces observations ne permettaient pas de considérer ces valeurs trouvées comme des valeurs pertinentes.

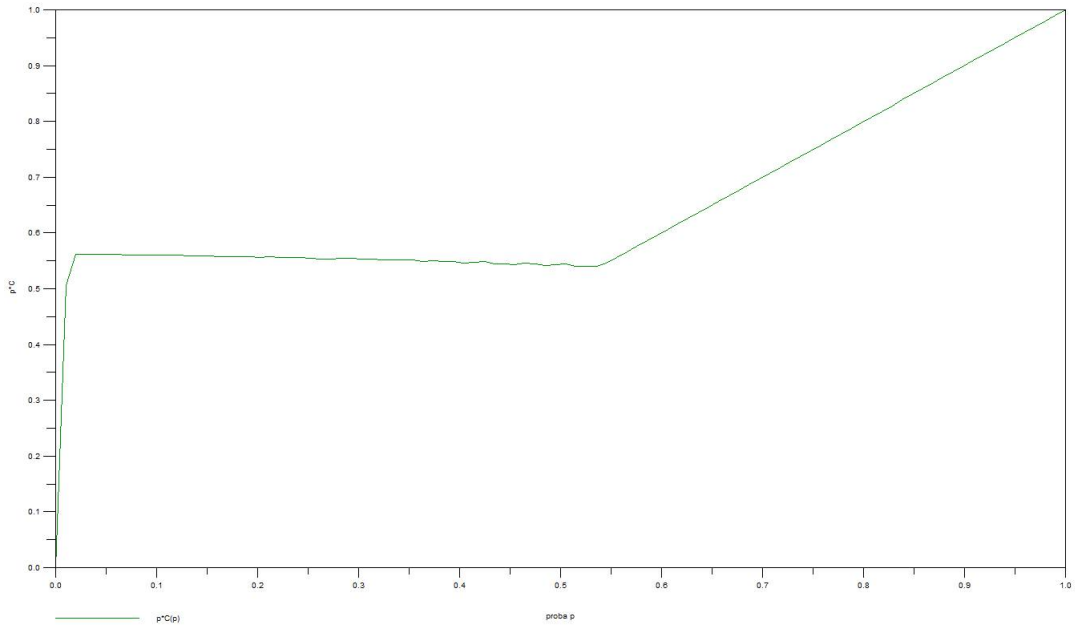
Essayons de comprendre pourquoi ce sondage ne nous donne pas des valeurs de α et de β utilisables. En questionnant les étudiants qui ont bien voulu répondre à notre questionnaire, nous avons dû admettre que notre sondage n'avait pas été très bien réalisé. En effet, pour les étudiants, le fait de donner les probabilités minimales n'avait pas été du tout intuitif laissant alors cours à des réponses plus ou moins aléatoires. De plus, nous avions précisé que notre sondage devait nous

aider à calculer l'excitation suscitée par le fait de miser sur un pari donné. Cette précision a forcé les personnes répondant au sondage à exagérer leurs probabilités donnant ainsi des réponses plus ou moins tronquées et exagérées.

Toutes ces remarques nous ont donc poussé à ne pas donner aux résultats de ce sondage l'importance qu'on imaginait pouvoir leur donner initialement. Cependant, les valeurs trouvées nous permettent de déduire une fourchette pour notre couple (α, β).

3.2.2 Etude d'un pari à une issue

A l'aide de scilab, et en suivant les indications données en première partie de 1.3.3, nous obtenons pour l'utilité u_1 aux paramètres $\alpha = 0.11$ et $\beta = 0.07$ la courbe pC en fonction de p suivante.



La courbe peut se décomposer en quatre phases. D'abord, pC est assez faible mais croît très vite. Ensuite pC semble se stabiliser. Puis, on observe d'étranges oscillations. Enfin, on rejoint la première bissectrice.

Expliquons les deux irrégularités correspondant à la première et la troisième phases. Le programme cherche en fait la cote de $1 + 0.1 * [|0, 500|] \subset [1, 50]$ qui maximise l'espérance de gain. Or, si le produit pC est constant, et que p est bien inférieur à $1/50 = 0.02$, alors C sature à 50. C'est pourquoi, en prenant C dans $[|1, 500|]$, le problème initial est moins observable. Par ailleurs, quand p devient assez grand, et pour un produit pC constant, la cote varie faiblement. Or ses valeurs étant discrètes, il lui faut faire des paliers avant de changer de valeur. Ce comportement en escalier conduit aux oscillations observées, quand C est multiplié par p.

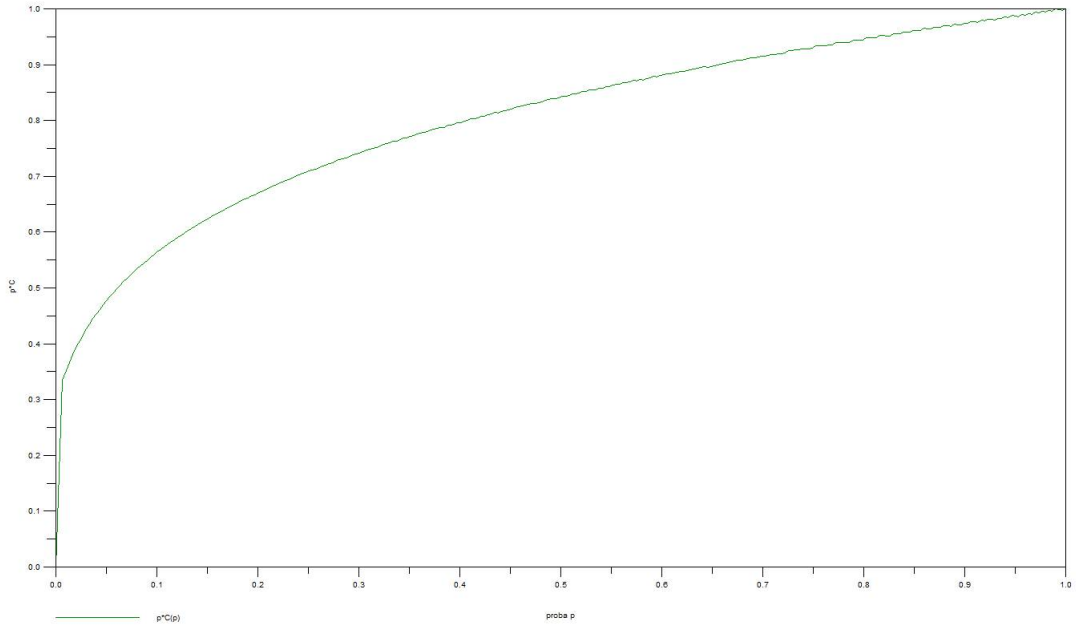
La dernière partie de la courbe est quant à elle aberrante. En effet, elle correspond à une cote $C = 1$, et signifie que pour des probabilités assez grandes (ici, supérieures à 0.55), la cote optimale pour le bookmaker est celle pour laquelle les joueurs sont sûrs de perdre. Cette aberration

est due au défaut du modèle relevé précédemment et selon laquelle il est possible que le joueur préfère jouer en étant sûr de perdre plutôt que ne pas jouer.

Il est possible d'estimer la valeur de la partie constante de la courbe. En effet, on se rend compte que la nullité de la dérivée revient à peu près à dire que le logarithme est de l'ordre de $1-pC$ qui est assez petit (ici, de l'ordre de 0.55. Le terme dont on prend le logarithme, $\frac{p}{q}\frac{\alpha R+\beta}{\alpha-\beta} = \lambda$, est donc proche de 1. Ainsi, on a $pC = p(R+1) = \lambda q(1-\beta/\alpha) - p\beta/\alpha + p = (p+\lambda q)(1-\beta/\alpha)$. Ainsi, pC est une fonction décroissante de p à cet endroit et $pC \approx (p+q)(1-\beta/\alpha) \approx 1-\beta/\alpha$. En fait, pC est légèrement supérieur à $1-\beta/\alpha$, surtout pour p petit et q grand, car λ est supérieur à 1. Ces estimations correspondent à la courbe dessinée.

Ainsi le rapport β/α s'interprête comme le taux de marge du bookmaker. Plus β/α est grand, plus l'effet de mise est proche de l'effet de gain, et c'est naturellement que le taux de marge du bookmaker augmente, celui-ci ayant intérêt à tirer parti de l'envie de jouer des parieurs.

Une étude similaire peut être réalisée pour la fonction d'utilité u_2 avec les paramètres $\alpha = 0.07$ et $\beta = 0.06$. On obtient la courbe suivante



Cette fois, la courbe est beaucoup plus régulière et on n'a pas de cas pathologique $C = 1$. On remarque que pC est croissant par rapport à p. Cela signifie que pour des probabilités plus grandes, la marge du bookmaker diminue, et celui-ci tente plus d'appâter les joueurs. A l'inverse pour des probabilités plus petites, le bookmaker se permet une grande marge, car il sait que les joueurs parieront quand même. C'est le *favorite longshot bias*!

3.2.3 Comparaison des cotes obtenues en utilisant u_1 et $u2$

Nous appliquons maintenant les résultats et l'algorithme décrit dans la seconde partie de 1.3.3. Le programme scilab dont le code se trouve en annexe 3, permet de calculer les cotes étant

données les probabilités des issues d'un match de football. Il est intéressant d'étudier le rapport entre les probabilités données et les probabilités déduites des cotes, c'est-à-dire les rapports entre p et λ.

Pour visualiser cet effet, observons l'annexe 4 qui donne une liste de matchs, les probabilités p_i et les λ_i calculés par nos modèles qui correspondent à ces probabilités. On remarque que le premier modèle a tendance à accentuer les écarts entre les probabilités, tandis que le second modèle les atténue.

Il s'agit d'une très belle visualisation du phénomène du *favorite longshot bias*. En effet, en atténuant les écarts entre les probabilités, le deuxième modèle diminue les grosses cotes, et augmente sa marge sur ces cotes. Il profite du fait que les gens surjouent malgré toutes ces cotes pour gagner plus d'argent, tandis qu'en augmentant les faibles cotes, il incite les joueurs à miser dessus, alors qu'elles sont d'ordinaire sous-jouées. A l'inverse, le premier modèle, comme on l'a vu en 3.2.2, a une marge qui diminue légèrement en fonction de p avant que la cote n'atteigne 1. Il comptera alors sur les mises sur les petites cotes, ce qui va à l'encontre du *favorite longshot bias*.

Le deuxième modèle est donc plus en accord avec les observations faites par les études sur le PMU. Cependant, en ce qui concerne le football, il ne retranscrit pas tout à fait les cotes réelles. C'est ce que nous allons maintenant voir.

3.3 Les cotes et les probabilités de la 34ème journée

Le tableau de l'annexe 1 présente pour chaque match de la 34ème journée

1. les cotes des principaux sites de paris sportifs,
2. la cote fabriquée seulement avec les probabilités calculées, avec 10% de marges pour le bookmaker,
3. les cotes calculées avec les deux modèles d'utilité.

Ce tableau nous permettant d'illustrer nos résultats, nous allons en profiter pour en tirer des remarques et des hypothèses sur nos modèles, dans l'optique d'expliquer les différences entre les cotes.

Pour cela nous avons choisi d'analyser les matchs de Lyon contre Valenciennes, et de Marseille contre Toulouse, car Lyon et Marseille sont deux équipes de parcours opposés. Lyon est parti fort en début de saison mais est en perte de régime, et Marseille est au contraire sur une bonne pente.

3.3.1 Remarques particulières sur les résultats de l'annexe 1

Voici les différentes cotes trouvées concernant les matchs Valenciennes-Lyon et Marseille-Toulouse :

Pour Valenciennes-Lyon :

Bookmaker	Cote 1	Cote N	Cote 2
Bwin	4.50	3.45	1.70
Unibet	4.20	3.40	1.85
betway	4.50	3.20	1.80
BetClic	4.75	3.40	1.75
sportingbet	4.50	3.25	1.75
Cotes probabilistes	3.61	2.63	2.20
Team PSC : u_1	3.91	2.65	2.09
Team PSC : u_2	3.46	2.65	2.25

Pour Marseille-Toulouse :

Bookmaker	Cote 1	Cote N	Cote 2
Bwin	1.55	3.30	6.50
Unibet	1.50	3.65	7.50
betway	1.55	3.50	6.25
BetClic	1.55	3.50	7.00
sportingbet	1.55	3.45	6.00
Cotes probabilistes	1.90	3.01	3.98
Team PSC : u_1	1.76	3.21	4.29
Team PSC : u_2	1.96	3.00	3.75

A propos des probabilités calculées (point de vue strictement sportif)

La série de résultats de Lyon et de Valenciennes avant la 34ème journée, est à peine différente, avec cependant un léger avantage pour Lyon. Valenciennes étant 15ème, et Lyon 3éme, si nous n'avions pas mis la dépendance en temps, Lyon aurait été le grand favori pour ce match, au vu de l'ensemble de la saison. Bien sûr, il l'était, mais dans une moindre mesure.

Nous voyons que par rapport aux bookmakers, notre modèle a privilégié Valenciennes (toutes les cotes de Lyon sont plus élevées que celles des bookmakers). On en déduit qu'être une grande équipe peut influer les chances de victoire, car on s'attend à un sursaut, les joueurs ont plus de qualités individuelles, il y a un plus grand potentiel.

Lyon : NDNVVDDDNV
Valenciennes : NDDVNVNVNV
Marseille : VVVVVVNVNV
Toulouse : NNDVNVDVNN
Bilan des derniers matchs, du plus récent au plus ancien
V=victoire, N=nul, D=défaite

Marseille, au contraire de Lyon, était sur une pente ascendante. On se rend compte que notre modèle de probabilité n'a pas annoncé Marseille si forte que l'annonçait les bookmakers, qui ont fixé des cotes autour de 7. Notre modèle donne environ à Toulouse une probabilité de victoire de 0.25, les bookmakers ne lui donnant que 0.14. Le cas de Marseille est délicat : tant du point de

vue des bookmakers comme nous le verrons ci-après, que du point de vue sportif : comme l'équipe de Marseille joue le titre, ce que notre modèle ne prend pas en compte, on imagine qu'elle a un surcroît de motivation qui améliore ses performances sportives.

A propos des cotes calculées (point de vue du bookmaker)

Il est intéressant de se placer maintenant du côté du bookmaker. Concernant le match Valenciennes-Lyon, les cotes pour la victoire de Lyon sont plus faibles chez les bookmakers que celles que l'on trouve avec nos modèles. La cote pour la victoire de Valenciennes est d'ailleurs bien plus élevée. Si on regarde la forme des deux équipes, on remarque cependant que Lyon semble à la peine en ce moment tout comme Valenciennes d'ailleurs. En ne s'appuyant que sur ces observations, on ne comprend pas ces cotes élevées sur la victoire de Valenciennes sur son propre terrain. On en déduit alors que d'autres phénomènes rentrent en ligne de compte. En effet, Lyon, tenant du titre, reste sur une série décevante face aux équipes qui jouent le titre comme Bordeaux et le Paris Saint-Germain. Le bookmaker s'attend alors à ce que tout le monde parie sur une réaction de Lyon face à un adversaire de moindre qualité qu'est Valenciennes, d'où ces cotes proposées par les bookmakers. Il est clair que notre modèle ne prend pas en compte ces phénomènes. En effet, il ne prend juste en considération la forme des équipes du moment. Or, comme Lyon et Valenciennes ont à peu près la même forme, notre modèle détermine des cotes plus neutres.

Considérons maintenant le match de Marseille face à Toulouse. Pour cela, il nous faut replacer le match dans son contexte. Marseille est en tête du championnat depuis trois journées et reste sur une série de victoires impressionnante. Toulouse, quant à eux, n'a plus à rien à jouer dans ce championnat et est dans une forme, pour ainsi dire, assez moyenne. Les bookmakers vont alors favoriser la victoire de Marseille sur son terrain. En effet, Marseille jouant le titre sur sa pelouse, tout le monde va penser que la victoire ne peut leur échapper, d'où cette cote de 1.5 de moyenne pour la victoire de Marseille. Notre modèle, quant à lui, ne prend pas en compte le fait que Marseille est en train de jouer le titre et peut compter alors sur une effervescence hors du commun, chose que ne peut pas connaitre Toulouse. Ainsi, on trouve des cotes moins élevées pour la victoire de Toulouse.

3.3.2 Remarques générales sur les résultats de l'annexe 1

Les matchs de football sont déterminés par une infinité de paramètres, comme le moral, le classement, les blessures, et notre modèle est une étude statistique, ne les prenant pas en compte. Ainsi il est normal de trouver des probabilités en décalage avec les probabilités réelles, qui sont impossibles à connaître précisément.

Notre modèle ne permet pas de prendre en compte certains phénomènes qui, pourtant, se rencontrent souvent dans les championnats de football. En effet, il ne peut pas prendre en compte le fait qu'une équipe joue la victoire en championnat, le nombre de supporters qu'a une équipe et qui parient sur elle (En effet, plus une équipe a de supporters, plus le bookmaker sait que les fans vont tous miser sur leur équipe).

Notre modèle tient plutôt compte de la forme des équipes en donnant un plus grand poids à leurs résultats les plus récents, de l'excitation qu'ont les parieurs au moment de miser de l'argent sur un pari donné. Cela nous permet de déterminer des cotes certes pas identiques à celles trouvées sur les sites de pari en ligne, mais elles correspondent cependant à une logique parfaitement

établie, logique dont on peut dire qu'elle n'est pas celle employée par les bookmakers.

Les cotes des bookmakers ne correspondent pas à la réalité des probabilités. De plus, elles sont totalement truquées, comme le montre les cotes pour Grenoble, de la 34ème journée, toutes fixées à 2.20 (annexe 1). Les bookmakers cherchent ainsi à se protéger et éviter les surebets qui consistent à parier sur les trois résultats possibles chez plusieurs bookmakers ayant des cotes suffisamment différentes pour pouvoir gagner quelque soit le résultat du match.

Chapitre 4

Remarques et remerciements

Notre groupe PSC se constitue de sept membres formant un groupe hétérogène mais complémentaire. Tous réunis autour d'une passion commune, le football, nous avons réussi à créer une véritable équipe travaillant ensemble et efficacement afin d'atteindre l'objectif qui était le notre au départ de ce projet. Pour cela, au cours de ces huit mois de travail, nous avons tenté de respecter l'organisation que nous avions mise en place au début, chose qui s'est révélée être parfois difficile. Alors que le projet touche à sa fin, il est temps pour notre groupe de tirer les enseignements du travail en équipe qui a été le notre tout au long de ce PSC.

4.1 Organisation du travail

Dès le début, nous avions envisagé de mener notre projet en trois phases bien définies qui permettaient de souligner au mieux nos axes d'études. De septembre à mi-décembre, nous avons commencé par nous constituer une bibliographie afin de nous familiariser avec l'univers des paris sportifs et de découvrir les études qui ont déjà été menées dans ce domaine. Durant ces trois mois, qui se sont révélés être essentiels pour la suite de notre étude, nous avons été confrontés à beaucoup de problèmes, notamment en ce qui concerne la compréhension des différents modèles déjà étudiés. Avec l'aide de notre tuteur, Benoît Jottreau, nous avons pu cerner avec plus ou moins d'efficacité les enjeux majeurs qui devaient faire l'objet de nos recherches.

Une fois cette première phase effectuée, il était clair que nous étions dans de bien meilleures conditions afin d'envisager la suite de notre étude. En effet, à la mi-décembre, nous avions décidé quels seraient alors nos principaux axes d'étude. Pouvait alors commencer la deuxième phase que nous avions envisagé de poursuivre jusqu'au mois de mars. Cette deuxième phase a représenté la phase centrale de notre PSC puisqu'elle consistait à utiliser des modèles préexistants, étudiés au cours de la première phase, dans le but de les prolonger afin d'en proposer de nouveaux, se voulant plus réalistes. Nous nous sommes alors lancés dans un travail de modélisation permettant de mettre en forme notre modèle. Afin de réaliser cela, le groupe s'est alors divisé en deux afin d'optimiser les travaux de recherche sur les deux axes principaux de notre étude, à savoir la modélisation des matchs de football et la modélisation du comportement du bookmaker.

Le groupe chargé d'étudier le comportement du bookmaker a été composé de Aldar Dugarzhapov, Lê Nguyên Hoang, Edouard Meyer et Alexis Watine. Son but a été de modéliser plusieurs fonctions d'utilité caractérisant le comportement des parieurs. Nous avons voulu que ces fonctions dépendent de plusieurs paramètres pour rendre notre modèle plus crédible et apporter ainsi

une véritable plus-value. Nous n'avons pas rencontré de problèmes majeurs quant à l'étude des fonctions d'utilité et nous avons même pu étudier le comportement des parieurs vis-à-vis de l'excitation que procure le fait de miser sur un pari. Arrivés jusque-là, il nous fallait un moyen pour déterminer au mieux les paramètres inconnus de nos fonctions d'utilité. Pour cela, nous avons envisagé de réaliser un sondage que nous avons proposé à un certain nombre d'élèves. Au final, nous avons eu une centaine de réponses, ce qui nous a permis d'estimer nos paramètres.

Le deuxième groupe s'est chargé d'étudier la modélisation des matchs de football. Partant de modèles préexistants, nous avons comparé les données simulées avec les données réelles afin de pouvoir choisir le meilleur modèle. Ici, c'est le calcul numérique qui nous a permis de réaliser des estimations par maximum de vraisemblance. Il fallait aussi tenir compte de la maniabilité des modèles étudiés pour choisir lesquels seraient à conserver, puisque certains d'entre eux se révélaient être très durs à manipuler.

Au mois d'avril, nous avons entamé la troisième et dernière phase de notre projet. Il s'agissait pour nous de mêler ces deux modèles pour répondre au mieux à notre problématique et en particulier déterminer les cotes qui maximisent les gains du bookmaker. La modélisation nous permettait, en effet, d'évaluer les probabilités des différentes issues d'un match, alors que nos modèles sur le comportement des parieurs nous permettaient de déterminer, pour une probabilité donnée, la cote optimale.

4.2 Problèmes rencontrés

Pendant toute notre étude, nous avons rencontré quelques problèmes. Tout d'abord, nous avons remarqué que notre sondage n'avait pas été idéalement réalisé et ne fournissait pas exactement les résultats que nous souhaitions. En effet, il a été dit par ceux qui ont bien voulu remplir notre questionnaire, que les questions posées étaient quelque peu difficiles à répondre et peu intuitives. Cet élement a sûrement conduit à un sondage quelque peu tronqué. Ainsi, les valeurs trouvées par l'intermédiaire de notre sondage n'ont pas pu être considérées comme des valeurs pertinentes de notre part. Cela sera probablement un enseignement important à retenir de cet expérience. Ensuite, lors de l'étude sur la modélisation des matchs de football, nous nous sommes rendus compte que nos programmes se révélaient être assez complexes avec de nombreuses lignes de code, rendant leurs corrections délicates.

4.3 Bibliographie

La liste qui suit reprend tous les articles que nous avons utilisés pour réaliser notre étude.

1. *Empirical evidence on the preferences of racetrack bettors* de Julien et Salanié
2. *Why are gambling markets organised so differently from financial markets ?*de Steven D. Levitt
3. *Formation des prix d'options binaires* de Benoît Jottreau
4. *Modelling association football scores and inefficiencies in the football betting market* de Dixon et Coles
5. *A birth process model for association football matchs* de Dixon et Robinson

4.4 Remerciements

Au moment de conclure, nous tenons à remercier notre tuteur, Benoit Jottreau, de nous avoir accordé de son temps pour nous aider et nous guider dans ce projet, ainsi que tous ceux qui ont répondu à notre sondage.

Chapitre 5

Conclusion

Notre étude concernant le marché des paris sportifs a été enrichissante à plusieurs égards. En premier lieu, l'actualité de notre sujet, liée à la libéralisation prochaine des sites de paris sportifs en France et l'expansion inévitable de ce marché, a rendu cette expérience intéressante et concrète. De plus, nous avons eu l'impression d'aborder un domaine encore peu étudié, ce qui a pu rendre nos recherches dignes d'intérêt. Certes, plusieurs études ont déjà été réalisées à propos des jeux d'argent, mais leur nombre relativement restreint nous a permis de créer nous-mêmes de nouveaux modèles et d'en étudier la validité, ce qui a constitué le point-clé de notre PSC.

Notre sujet nous a permis de mettre en pratique les connaissances acquises à l'Ecole dans des disciplines variées, comme la microéconomie, les mathématiques appliquées et l'informatique. La mise au point d'expériences (sondage, simulations informatiques) a donné de la valeur à nos études théoriques. Il est vrai que l'impossibilité d'accéder aux données ou méthodes des bookmakers eux-mêmes limite notre étude du comportement du bookmaker à un cadre très théorique. En revanche, il nous semble que les modèles que nous avons mis au point concernant les parieurs, consolidés par le sondage, peuvent offrir une alternative crédible aux fonctions d'utilité utilisées jusqu'ici dans la théorie des préférences sous risque, en prenant en compte la spécificité de l'action du pari.

Le modèle créé afin d'évaluer de manière rationnelle les probabilités des matchs a pour l'instant donné des résultats très satisfaisants : en effet, même s'il n'a pu être exploité que pour un nombre limité de matchs (le modèle étant récent), la confrontation des cotes fournies avec celles des principaux sites de paris en ligne a fourni des résultats au-delà de nos attentes. Cependant, il est clair que tous les paramètres intervenant dans l'issue d'un match ne sauraient être pris en compte dans ces modèles, fort heureusement pour l'intérêt du sport. En réalité il semblerait que les bookmakers essaient de prendre en compte un grand nombre de ces paramètres, sans qu'un modèle permette de les mesurer véritablement.

Finalement, il est encore discutable que les modèles mathématiques soient l'outil privilégié pour étudier le marché des paris sportifs et optimiser les gains des bookmakers, mais il était en tout cas très intéressant pour nous tous d'explorer ce domaine inconnu et d'y obtenir des résultats satisfaisants à partir de modèles que nous avons créés.

Chapitre 6

Annexes

6.1 Annexe 1 : comparaison des cotes

Les données comparent les cotes prises sur différents sites concernant les matchs de la 34ème journée de ligue 1 en 2008-2009. Les cotes probabilistes valent $\frac{1-\gamma}{p_i}$ pour $\gamma = 0.1$ et p_i les probabilités calculées par notre modèle des matchs de football et de calcul des paramètres. Les cotes de TeamPSC 1 sont celles déterminées ensuite avec l'utilité u_1 et les cotes de TeamPSC 2 sont celles trouvées grâce à u_2.

6.2 Annexe 2 : exploitation du sondage

Ces lignes de code écrites sur le logiciel R permettent de déterminer les paramètres des utilités u_1 et u_2.

6.3 Annexe 3 : calcul des cotes

Ces lignes de codes écrites sur le logiciel scilab permettent, étant données des probabilités p_i et une marge γ de déterminer les λ_i optimaux pour le bookmaker.

6.4 Annexe 4 : comparaison des p et λ

Ce tableau souligne les effets de nos modèles sur les probabilités déduites des cotes déterminées par le bookmaker qui maximise son espérance de gain.