Mines Nancy - Université de Lorraine
Fédération Française de Badminton

Mémoire de fin d'études :

Développement d'un modèle de rating adapté aux joueurs de badminton

Rédigé par : **Blanchard Armand**

Encadrant : **Enzo Hollville**
Professeur référent : **Sandie Ferrigno**

INSEP, Mars – Août 2025

Table des matières

Remerciements

Je tiens tout d'abord à remercier chaleureusement la Fédération Française de Badminton (FFBaD) et l'Institut National du Sport, de l'Expertise et de la Performance (INSEP) de m'avoir permis de réaliser ce stage dans un cadre rêvé.

Je remercie tout particulièrement Enzo Hollville, pour son accompagnement constant, sa disponibilité, ses conseils et la confiance qu'il m'a accordée tout au long de ce stage. J'adresse également mes remerciements à Lucas Minet, avec qui il a formé un duo complémentaire. Leurs retours pertinents m'ont permis de progresser tant sur le plan personnel que technique ou méthodologique.

Je souhaite également exprimer ma gratitude à Sandie Ferrigno, mon professeur référent à l'école des Mines, pour son suivi académique, et son soutien bienveillant.

Merci à l'ensemble du staff du pôle olympique de la FFBaD pour les échanges enrichissants, ainsi qu'aux joueuses et joueurs de haut niveau qui nous motivent au quotidien pour également nous dépasser dans nos travaux.

Abstract

Français

L'article suivant aborde le développement et la validation d'un système de rating des joueurs évoluant sur le circuit professionnel de badminton. Le but étant de proposer une alternative viable et dynamique au classement international établi par la fédération internationale de badminton (BWF). Ce système de ratings doit prendre en compte chaque match joué sur le circuit international de badminton et distribuer des points en fonction de l'issue du match et du niveau relatif des joueurs s'affrontant. Pour mettre en place un tel système, on utilise l'ensemble des matchs du circuit international de badminton de 2008 à 2025. On utilisera et adaptera les méthodes de ratings usuelles, en l'occurrence ELO et Glicko pour développer trois différents systèmes de ratings. Pour ce faire, nous utiliserons les méthodes ELO et Glicko comme décrites dans la littérature, et un troisième système adapté de ELO avec certains paramètres que l'on connait impactant la performance en badminton. Une première étape de test et de sélection de paramètres précèdera une étape de comparaison et de validation inter-méthodes. Ces méthodes de ratings pourront ensuite être automatisées et calculées chaque semaine afin d'analyser en temps réel les performances des athlètes. Les performances seront alors plus représentatives du niveau des athlètes et on pourra catégoriser les performances selon les gains et pertes de ratings. A l'avenir, les évolutions de ratings associés à d'autres paramètres pourront être prises en compte dans un modèle plus complet afin de mettre davantage de contexte sur la performance des athlètes et faire des bilans de saison par exemple.

Anglais

This article addresses the development and validation of a player rating system for athletes competing on the professional badminton circuit. The goal is to propose a viable and dynamic alternative to the international ranking established by the Badminton World Federation (BWF). This rating system must take into account every match played on the international circuit and allocate points based on the match outcome and the relative level of the competing players. To build such a system, we use all matches from the international badminton circuit between 2008 and 2025. We will apply and adapt existing rating methods—specifically, Elo and Glicko—to develop three distinct rating systems. The first two follow standard Elo and Glicko implementations from the literature, while the third is a modified Elo system incorporating parameters known to influence badminton performance. An initial testing phase will focus on parameter selection, followed by a comparison and validation phase between methods. These rating methods can then be automated and calculated on a weekly basis to enable real-time analysis of athlete performance. The resulting ratings will offer a more representative measure of player level,

and performances can be categorized based on rating gains or losses. In the future, rating evolution combined with other parameters could be included in a more comprehensive model to provide greater context to athlete performance and facilitate seasonal reviews, for example.

Abréviations

BWF	Badminton World Federation
MS	Men Single
WS	Women Single
MD	Men Double
WD	Women Double
XD	Mixed Double
FIDE	Fédération internationale des échecs
USCF	United States Chess Federation

Introduction

Depuis quelques années, l'essor des technologies numériques et la généralisation de la collecte de données ont fait évoluer la vision de la performance et les pratiques dans ce domaine. Le sport de haut niveau qui est en constante recherche de performance a donc besoin de s'inscrire de plus en plus dans cette dynamique. Ce mouvement s'inscrit désormais au cœur de la stratégie de développement de la fédération française de badminton, qui intègre de plus en plus des outils de data science pour objectiver les décisions en matière de détection, de préparation, et de suivi des athlètes. Dans ce contexte où l'on peut s'appuyer sur des jeux de données de plus en plus importants, les systèmes de notation dynamiques représentent une avancée prometteuse pour modéliser la performance des athlètes de manière plus fine et plus contextualisée que les classements traditionnels.

Contrairement au classement mis en place par la fédération internationale de badminton (Badminton World Federation – BWF), fondé sur l'accumulation de points selon des règles fixes liées aux résultats et au niveau des tournois, le système proposé cherchera à refléter le niveau réel des joueurs à un instant donné, en tenant compte du contexte des matchs disputés. Le contexte d'un match sera défini en majorité par le niveau de l'adversaire mais nous savons que la performance des joueurs internationaux est aussi affectée par le niveau du tournoi, l'avancée dans le tournoi, l'état de forme des joueurs, la durée d'inactivité liée aux blessures ou à l'arrêt momentané de la compétition etc. Le sport et la performance sportive étant aussi grandement impactée par des facteurs liées à l'humain et à l'aléatoire, on essaiera dans ce mémoire d'utiliser une sélection de facteurs qui nous semble avoir un réel impact sur l'issue des matchs pour adapter nos méthodes de ratings.

Dans un premier temps, nous étudierons trois différentes méthodes qui s'appuieront sur des modèles de notation dynamiques inspirés d'algorithmes utilisés dans d'autres disciplines sportives (comme Elo ou Glicko). Un système de notation dynamique est un modèle probabiliste permettant d'estimer le niveau d'un joueur à partir des résultats passés en ajustant en continu sa valeur. Ces systèmes permettent non seulement de produire un classement, mais aussi d'estimer la probabilité de victoire face à un adversaire donné, ce qui en fait un outil d'aide à la décision particulièrement utile en contexte de sélection ou de planification stratégique. Dans ce mémoire on étudiera ce qui a été fait dans d'autres disciplines mais aussi la théorie de ces méthodes afin de pouvoir au mieux les comprendre et les adapter au badminton et à ses spécificités. Une démarche comparative sera ensuite mise en place afin d'évaluer les performances respectives de ces différentes méthodes sur le circuit international, pour ensuite valider une ou plusieurs méthodes et les intégrer directement dans les outils d'aide à la décision du pôle olympique : Badminton Manager.

Ce travail s'inscrit dans une logique d'amélioration continue de la compréhension de la performance en badminton, dans un environnement où les entraîneurs, analystes et déci-

deurs ont besoin d'outils réactifs, contextuels et interprétables pour guider leurs décisions. Le mémoire contribuera ainsi à la fois à une meilleure modélisation de la performance et à son intégration opérationnelle dans les processus décisionnels du haut niveau.

Face à l'ensemble de ces objectifs, ce mémoire s'attache à répondre à une question centrale : Comment concevoir un système de classement alternatif et innovant en badminton, permettant de mieux représenter les rapports de force entre joueurs et de contextualiser leurs performances face à la concurrence mondiale ?

Pour y répondre, plusieurs interrogations complémentaires guideront notre démarche :

— Quels sont les limites du classement mondial actuel dans l'évaluation des dynamiques de performance et des progressions des joueurs ?

— Comment adapter et/ou améliorer des modèles existants comme Elo, Glicko ou des métriques bayésiennes pour mieux refléter les performances spécifiques au badminton ?

— Quels critères et variables (niveau des adversaires, score des matchs, régularité, conditions de jeu, surface, etc.) doivent être intégrés pour un classement plus représentatif ?

— Comment valider et tester ce nouveau système de classement pour s'assurer de sa robustesse et de sa pertinence pour les joueurs français et internationaux ?

Pour répondre à ces problématiques, le mémoire s'articulera en plusieurs parties : une revue de littérature pour contextualiser les systèmes de notation existants dans le sport, une présentation détaillée des modèles dynamiques choisis, leur implémentation et leur évaluation comparative, avant de discuter des résultats obtenus et des perspectives d'intégration dans un outil d'aide à la décision destiné au staff fédéral.

Cadre théorique et état de l'art

1.1 Théorie des systèmes de ratings : Elo et Glicko-2

Les systèmes de ratings ont été développés afin de mesurer la compétence relative de différents protagonistes dans un domaine précis. Il est possible d'imaginer cela comme un rapport de force entre les différents partis qui s'exprime par une différence de rating. Cette différence de rating se retranscrit par une probabilité de victoire.
Selon les applications, les systèmes de ratings peuvent rencontrer des défis comme l'incertitude des rencontres entre les joueurs, l'inactivité de certains qui impacte différemment le rating selon la discipline, ou encore l'impact du temps sur l'évolution des ratings.
Le système de ratings ELO développé en 1978[1] est de loin le plus connu et le plus démocratisé. C'est le pilier fondateur des systèmes de ratings. D'autres systèmes ont émergé par la suite, complexifiant ainsi le principe initial d'ELO. On notera en particulier le modèle Glicko[2] en 1999 et sa seconde version en 2001[3] ou encore le modèle TrueSkill en 2007.

1.1.1 Le système Elo : fondements et limites

La méthode de classement ELO conçue initialement pour les échecs est devenue un standard dans de nombreux domaines. L'imagination et la conception de ce classement sont associées à Arpad Elo qui décrit son approche statistique dans le livre *The Rating of Chess Players, Past and Present* écrit en 1978[1]. L'idée fondatrice de la méthode ELO est d'utiliser la courbe logistique comme base de travail pour développer une approche statistique du classement qui associerait des probabilités de victoire aux joueurs s'affrontant dans une rencontre. Par la suite, il démontrera empiriquement la pertinence de cette méthode à travers les données de la Fédération Internationale des Échecs (FIDE).

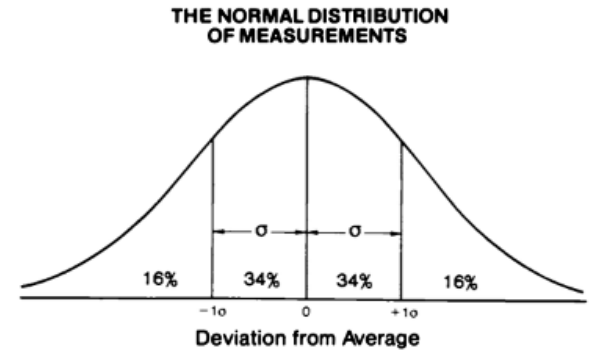


FIGURE 1.1 – Représentation de la distribution normale standard utilisée par A. Elo

L'hypothèse fondamentale de M. Elo est que dans le sport, l'expérience veut que le meilleur joueur ne batte pas toujours le plus faible. La performance d'un joueur est variable, et même à très haut niveau, un joueur peut connaître des hauts et des bas. L'hypothèse traduite est donc la suivante : « Lorsque les performances d'un individu sont étudiées sur une échelle appropriée, celles-ci seront distribuées selon une loi normale. » [1]

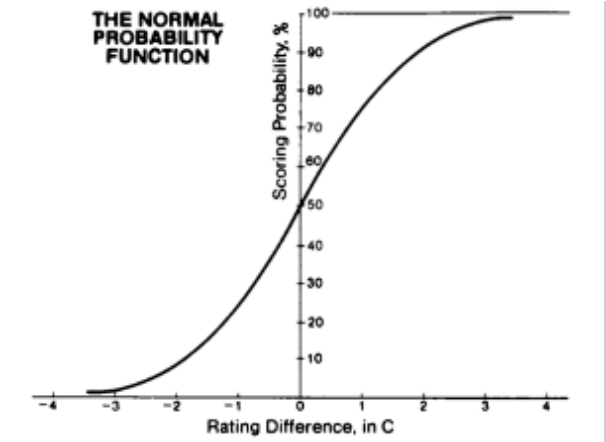


FIGURE 1.2 – Représentation de la régression logistique utilisée dans la méthode ELO

La distribution des points a été imaginée par Arpad Elo de la manière suivante :

$$R_n = R_o + K(W - W_e) \tag{1.1}$$

où :

— R_n est le nouveau rating du joueur après l'évènement ;

— R_o est le rating du joueur avant l'évènement ;

— K est un facteur (appelé facteur de développement) qui régit le nombre de points maximum pouvant être distribués lors d'un match ;

— W est l'issue du match (1 pour une victoire, 0 pour une défaite, et $\frac{1}{2}$ pour un match nul — ce qui n'existe pas en badminton) ;

— W_e est la probabilité de victoire du joueur calculée à partir de son rating R_o.

À la suite des travaux d'Arpad Elo, la probabilité de victoire basée sur la régression logistique a été confirmée et explicitée par l'USCF et la FIDE grâce au modèle de Bradley & Terry [4]

$$\mathbf{F}_L(v) = \frac{1}{1 + 10^{-v/\sigma}}, \tag{1.2}$$

Ou v est l'écart de rating entre les deux joueurs et sigma un facteur d'échelle pouvant être modifié. Ce facteur a été fixé à 400 initialement par Arpad Elo par des considérations empiriques qui permettent d'offrir une échelle lisible, où des tranches de 200 à 400 points correspondent à des différences nettes de niveau, mais pas excessives.

Le facteur K permet de donner plus ou moins d'importance au match en question, la variation de rating sur le match en sera dépendante. Un facteur K faible n'impactera que très peu le rating du joueur, et aura un impact conservatif sur le classement ; tandis qu'un

facteur K élevé entrainera une importante variation de rating ce qui rend le classement plus dynamique et plus réactif à chacun des matchs. Ce facteur K vaut 40 jusqu'à la 30e partie du joueur, sinon 20 pour un classement Elo en dessous de 2 400 Elo, sinon, 10 pour un classement Elo au-dessus de 2 400 (cela reste 10 même si le joueur repasse en dessous de 2400). Une exception est faite pour les joueurs en dessous de 18 ans qui ont un K valant 40 tant qu'ils ont moins de 2300 Elo.[1]

La méthode Elo malgré sa popularité et sa grande utilisation à travers le monde est basée sur des méthodes empiriques et sur l'hypothèse de Arpad Elo que la distribution des performances d'un individu suit une loi normale. Certains mathématiciens comme Sam Olesker-Taylor et Luca Zanetti ont donc utilisé des théories mathématiques comme les chaines de Markov pour donner un cadre théorique rigoureux. Le système Elo repose sur une mise à jour temporelle match après match. Il est interprété ici comme une descente de gradient stochastique sur la log-vraisemblance du modèle de Bradley & Terry vu précédemment. Les auteurs proposent ensuite une formulation en chaîne de Markov non réversible sur $\mathbb{R}^n$, ce qui permet une analyse probabiliste. Les résultats de cette étude sont tels que le modèle converge vers les paramètres du modèle de Bradley & Terry, ce qui confirme une convergence vers le niveau des joueurs. Enfin ils ont aussi montré expérimentalement que le système ELO restait stable en valeur absolue, ce qui confirme une convergence du modèle sans explosion des scores.

Sonas, a réalisé de nombreuses études sur les limites et les biais possibles de la méthode ELO. Dans une de ses récentes études[5] en 2023, il relève plusieurs limites du classement ELO aux échecs par exemple depuis que le système Elo existe les notations ont subi un mouvement de baisse globale qui fait les prédictions du système Elo actuelles sont moins fiables qu'avant. Il relève aussi que les jeunes joueurs qui réalisent une série de victoire peuvent voir leur classement gonflé artificiellement alors qu'à contrario les classements des joueurs expérimentés ou ayant été très forts mettent du temps à s'ajuster. Selon lui, c'est donc un biais lié au facteur K qui est constant et arbitraire. Celui-ci n'est donc pas adapté à tous les cas de figure, ce qui ajoute un biais supplémentaire à ce système de notation.

1.1.2 Glicko-2 : une version évoluée de Elo

Dr. Mark E. Glickman a développé en 1995 le système Glicko qui est une adaptation probabiliste et bayésienne de l'Elo. Ce système prend en compte une incertitude sur la compétence réelle d'un joueur. L'objectif de Glickman était donc d'améliorer la fiabilité et la réactivité du classement en fonction de la fréquence de jeu. Selon lui, le classement Elo deviendrait donc un cas particulier de Glicko où tous les joueurs auraient la même fréquence de jeu.

Dr. Mark E. Glickman part d'un constat qui est que deux joueurs ayant des historiques de compétition très différents devraient avoir des ratings plus ou moins variables. En effet, si deux joueurs ayant un même rating s'affrontent mais que l'un reprend la compétition après plusieurs années d'inactivité et que l'autre joue chaque semaine, ces deux joueurs devraient selon Glickman avoir des variations de rating différentes. Le système Elo entraîne une mise à jour symétrique des ratings quand Glicko voudrait mettre en jeu plus de points pour celui n'ayant pas joué depuis longtemps, car son rating est incertain. À

l'inverse, son adversaire jouant toutes les semaines à ce même niveau a probablement un rating très stable, donc devrait avoir moins de points en jeu sur ce match.

Pour ce faire, Glickman introduit deux fonctions représentant le niveau de jeu de chaque joueur. Le Rating (R) étant, comme pour ELO, une estimation du niveau de jeu et une *Rating Deviation* (RD) qui est un écart-type statistique correspondant à l'incertitude sur le niveau de jeu. Le rating étant donc une estimation avec un écart-type, Glickman propose donc de voir le niveau de jeu de l'individu comme un *intervalle de confiance*. Par exemple, une personne ayant un rating de 1850 ± 50 (RD) a donc 95% de chance d'avoir un niveau réel situé dans l'intervalle :

$$\mathbf{I} = [R - 2 \times RD, R + 2 \times RD] = [1750, 1950], \tag{1.3}$$

Par la suite, Dr. Mark E. Glickman a continué de développer son système de ratings et a publié une nouvelle version de son papier dans les années 2000 avec une version stabilisée en 2012. Il introduit alors une nouvelle variable *volatility* (σ) qui représente le degré de fluctuation des performances. Une volatilité élevée signifie que le joueur a des performances très irrégulières.

Le calcul des ratings selon la méthode Glicko-2 est explicité par Dr. Mark E. Glickman dans son papier. Tout d'abord, chaque joueur est initialisé avec des valeurs de R, RD et σ conseillées respectivement autour de 1500, 200 et 0.06 (modulables dans une certaine mesure selon l'utilisation). De même, une constante globale τ régissant l'évolution de la volatilité au fil des matchs est définie aux alentours de 0.5 (une valeur entre 0.3 et 1.2 est conseillée dans la littérature).

Ensuite viennent huit étapes de calcul : La première étant une conversion à « l'échelle Glicko-2 » qui est une échelle sans dimension centrée en zéro.

$$\mu = \frac{r - 1500}{173.7178} \quad \text{et} \quad \phi = \frac{\text{RD}}{173.7178}, \tag{1.4}$$

Ensuite, on calcule la quantité v qui est la quantité estimée de la variance uniquement basée sur le ou les matchs joués.

$$v = \left[\sum_{j=1}^{m} g(\phi_j)^2 \cdot E(\mu, \mu_j, \phi_j) \cdot (1 - E(\mu, \mu_j, \phi_j)) \right]^{-1}, \tag{1.5}$$

avec les fonctions auxiliaires suivantes :

$$g(\phi) = \frac{1}{\sqrt{1 + \frac{3\phi^2}{\pi^2}}} \tag{1.6}$$

$$E(\mu, \mu_j, \phi_j) = \frac{1}{1 + \exp\left(-g(\phi_j)(\mu - \mu_j)\right)} \tag{1.7}$$

La fonction g est un facteur de pondération de la fiabilité du rating adverse. Elle provient de la forme de la variance du logit. L'idée de Glickman est de prendre en compte l'incertitude du rating adverse en tant que pondération de la variance estimée autour de 0.

La fonction E quant à elle est la probabilité attendue que le joueur batte l'adversaire j. C'est une fonction logistique qui peut être inspirée du modèle de Bradley–Terry (comme Elo), mais cette fois ajustée par rapport à l'incertitude de l'adversaire.

La quatrième étape consiste à mettre à jour la volatilité σ en résolvant une équation non linéaire $f(\sigma') = 0$ définie comme suit :

$$f(x) = \frac{e^x \left(\Delta^2 - \phi^2 - v - e^x\right)}{2\left(\phi^2 + v + e^x\right)^2} - \frac{(x-a)}{\tau^2} \tag{1.8}$$

Pour trouver cette valeur, l'auteur utilise une méthode itérative basée sur celle de l'Illinois.

L'étape suivante consiste à mettre à jour la déviation standard (ϕ) ainsi que le rating du joueur. La déviation standard diminue si les nouveaux matchs apportent de l'information supplémentaire sur le niveau réel du joueur. Le nouveau rating est ensuite calculé, de manière analogue au classement Elo, mais en le pondérant par $g(\phi_j)$ (qui réduit l'influence des adversaires avec une incertitude élevée) et par ϕ'^2 (qui atténue les mises à jour si l'estimation du niveau est déjà très précise).

$$\phi^* = \sqrt{\phi^2 + \sigma^2} \tag{1.9}$$

$$\phi' = \left(\frac{1}{\phi^{*2}} + \frac{1}{v}\right)^{-\frac{1}{2}} \tag{1.10}$$

$$\mu' = \mu + \phi'^2 \sum_{j=1}^{m} g(\phi_j)\left(s_j - \mathrm{E}(\mu, \mu_j, \phi_j)\right) \tag{1.11}$$

Enfin, les joueurs sont remis à l'échelle classique de ratings :

$$r' = 173.7178\,\mu' + 1500 \tag{1.12}$$

$$\mathrm{RD}' = 173.7178\,\phi' \tag{1.13}$$

Glicko-2 permet également de traiter les périodes d'inactivité des joueurs. En cas d'absence de compétition sur une période, leur incertitude est augmentée de la manière suivante :

$$\phi' = \phi^* = \sqrt{\phi^2 + \sigma^2} \tag{1.14}$$

1.1.3 Autres systèmes de ratings et limites dans notre étude

Stephenson a développé en 2012 un système dérivé de Glicko prenant en compte deux paramètres supplémentaire : un bonus de participation et un paramètre de voisinage prenant en compte les résultats des adversaires entre eux pendant la période d'étude.

Les systèmes ELO, Glicko et Stephenson sont des systèmes de rating permettant de noter et classifier des « un contre un ». Microsoft a développé TrueSkill pour ses jeux Xbox afin de répondre à la problématique d'affecter des notes individuelles sur des jeux qui se jouent en équipes. Le calcul de l'actualisation des ratings est basé sur une densité de probabilité gaussienne contrairement à logistique pour Elo et Glicko. Enfin les formules d'actualisations sont plus complexes que pour les systèmes précédents.[6]

Dans le cadre de notre étude nous nous cantonnerons aux ratings Glicko-2 et Elo qui ont des formules d'actualisation lisibles et adaptées à notre application. En effet, nous pouvons calculer les ratings match par match ce qui rend le paramètre de voisinage de Stephenson inutile. De plus, les joueurs étant contraints par le classement BWF à jouer au minimum dix tournois dans l'année cela rend le bonus de participation moins pertinent. Enfin, on aurait pu imaginer un système type TrueSkill pour la gestion des paires de doubles mais par soucis de cohérence avec le système de ratings de simple on choisira de ne pas continuer avec TrueSkill.

1.2 Différentes approches de classement dans le sport

Aujourd'hui les fédérations internationales s'intéressent de plus en plus aux systèmes de ratings afin de classer de manière plus juste leurs populations. Pour rendre compte de la réalité des différents sports, on peut tout d'abord les répartir entre les sports collectifs comme le football, le basketball ou encore le rugby ; les sports de confrontation comme le tennis, le judo ou le badminton, et les sports à métriques comme l'athlétisme ou le ski. Ces sports étant des pratiques très différentes avec leurs propres systèmes de scores et de tournois ne sont pas adaptables de la même manière.[7]
Si l'on s'intéresse plus précisément aux sports cités, seul le football depuis 2018 a développé un système de classement type ELO. La FIFA distribue un nombre de points aux nations selon les probabilités de victoires de chaque match sur la base de leur classement. Chaque match est également pondéré en fonction de l'enjeu qui va du match amical à une coupe du monde.[8] La fédération internationale de basketball quant à elle utilise un système ajusté et moyenné sur 8 ans.[9] Enfin le rugby utilise un système de classement ajusté type probit (une sorte de cumul coefficienté en fonction du niveau de l'adversaire).[10] Pour les sports d'opposition comme le badminton, les systèmes de type ELO ne se sont pas encore développés au niveau des fédérations internationales. Les classements sont actuellement faits à partir d'un cumul de points qui varie selon les disciplines. Par exemple au tennis, le classement ATP est fait à partir d'un cumul de points sur une fenêtre glissante de 52 semaines durant lesquelles les joueurs accumulent des points sur 18 tournois. Chaque joueur accumule des points en fonction de ses performances sur les différents niveaux de tournois. Par exemple, une victoire en grand chelem (plus haut niveau de tournoi) rapporte 2000 points ATP quand un premier tour rapporte 10 points. Enfin l'athlétisme qui est un sport un peu particulier avec des métriques et un grand nombre d'épreuves différentes, met en place deux types de classements. L'un par épreuves ou groupe d'épreuves similaires

et le second toutes épreuves confondues. On fait alors un cumul de points sur trois ou cinq performances qui prennent en compte le classement sur l'épreuve, le résultat (chronomètre, longueur etc. . .) selon une grille de points. Un bonus de dix ou vingt points peut être ajouté en cas d'égalisation ou de nouveau record du monde.[11] Finalement, au niveau professionnel les systèmes de ratings ajustés de type Elo ne sont que très peu démocratisés en particulier dans les sports d'opposition comme le Badminton. Une proposition de papier sur le sujet a été faite en juin 2024 mais celui-ci n'a pas été publié et aucun rating officiel n'est disponible à l'heure actuelle.

1.3 Le classement BWF et ses limites

Le classement BWF est un classement cumulatif sur 52 semaines glissantes basé sur les performances des joueurs ou des paires de doubles. Il concerne toutes les disciplines du badminton international, c'est-à-dire aussi bien le simple dame (WS), que le simple homme (MS), que le double dame (WD), que le double homme (MD) ou encore le double mixte (XD). Le classement BWF est un système de points basé sur la performance des joueurs lors de tournois reconnus par la BWF. Ce classement est mis à jour chaque mardi afin de valider les inscriptions en tournois et de déterminer les têtes de série des tournois. Chaque tournoi rapporte un nombre de points selon le niveau du tournoi et le round atteint. Par exemple si l'on prend le grade le plus haut des tournois (hors championnat du monde et jeux olympiques) : un Super 1000, le nombre de points varie entre 3000 et 12000. Voici la répartition globale des points rapportés selon le niveau de tournoi et le round atteint par le joueur.

	Winner	Runner Up	* 3-4	5-8	9-16	17-32	33-64	65-128	129-256	257-512	513-1024
Grade 1 - BWF Tournaments (BWF World Championships and Olympic Games)	13000	11000	9200	7200	5200	3200	1300	650	260	130	65
Grade 2 – BWF World Tour – Level 1 (Finals) and Level 2	12000	10200	8400	6600	4800	3000	1200	600	240	120	60
Grade 2 – BWF World Tour, Level 3	11000	9350	7700	6050	4320	2660	1060	520	210	100	50
Grade 2 – BWF World Tour, Level 4	9200	7800	6420	5040	3600	2220	880	430	170	80	40
Grade 2 – BWF World Tour, Level 5	7000	5950	4900	3850	2750	1670	660	320	130	60	30
Grade 2 – BWF World Tour, Level 6	5500	4680	3850	3030	2110	1290	510	240	100	45	30
Grade 3 – International Challenge	4000	3400	2800	2200	1520	920	360	170	70	30	20
Grade 3 – International Series	2500	2130	1750	1370	920	550	210	100	40	20	10
Grade 3 – Future Series	1700	1420	1170	920	600	350	130	60	20	10	5

FIGURE 1.3 – Tableau de répartition des points en fonction de la place atteinte dans le tournoi

Le même barème de points est appliqué à toutes les catégories (WS, MS, WD, MD, XD).

Lors des 52 semaines glissantes sur lesquelles sont calculées le classement, seuls les 10 meilleurs résultats obtenus au cours des 52 dernières semaines sont comptabilisés. Certaines spécificités sont tout de même à prendre en compte. Par exemple, si un joueur a moins de 10 résultats, seuls ceux disponibles comptent. Seconde spécificité, lorsque la date d'un tournoi est déplacée d'une année sur l'autre, dès que celui-ci est rejoué les points de l'édition précédente ne sont plus comptabilisés.
Enfin, certaines particularités sont associées à la pratique du double. Le classement BWF pour une paire est associée à la paire et non au joueur individuel. En cas de nouvelle paire, les deux joueurs commencent donc sans historique ensemble. On peut d'ailleurs aussi noter qu'il est possible pour un joueur de jouer dans plusieurs catégories la même semaine. Il est d'usage chez certains joueurs de jouer en simple et double ou en double et double mixte. [12]

Voici ci-dessous un tableau récapitulatif et comparatif des différents systèmes de ratings que l'on pourrait mettre en place au badminton ainsi que le classement BWF. Selon ces critères :

TABLE 1.1 – Comparaison des systèmes de classement

Critère	**Elo**	**Glicko-2**	**Stephenson**	**BWF**
Type de modèle	Probabiliste logistique	Bayésien	Glicko+bonus	Cumulatif
Importance du match	Par K	Non	Non	Oui (points tournois)
Adapté au double	Non	Non	Non	Oui (par paire)
Inactivité prise en compte	Non	Oui	Oui	Non
Complexité d'implémentation	Faible	Moyenne	Haute	–

1.4 Les spécificités du badminton

Afin de contextualiser l'utilisation des systèmes de ratings dans le cadre du badminton, il est essentiel de détailler les spécificités de la discipline. Le badminton est un sport de confrontations directes. C'est un sport individuel ou en binôme qui se joue donc en un contre un pour le simple ou en deux contre deux pour le double. Le badminton permet donc une attribution claire du résultat à l'individu ou à la paire victorieuse.

Au niveau des catégories, le badminton comporte cinq disciplines distinctes : le simple homme (MS), le simple dame (WS), le double homme (MD), le double dame (WD) et le double mixte (XD). Les styles de jeux et dynamiques compétitives sont très différentes selon les disciplines, le double étant beaucoup plus dynamique que le simple par exemple.

Le format de jeu du badminton est de deux sets gagnants de 21 points, un troisième set peut donc être joué si chacun des deux camps gagne un set. Si les deux camps atteignent 20 points, le set est alors prolongé. Le camp menant par ma suite de deux points d'écarts sera ensuite annoncé victorieux. Si l'égalité persiste jusqu'à 29 points alors le camp mar-

quant le 30e point gagne le set. Ce format de rencontre entraine une certaine variabilité des scores ce qui marque une différence de niveau réelle ou simplement des moments clés basculants. Une des informations à noter est que le format du badminton admet forcément un vainqueur. Le rating n'aura donc que deux issues possibles la victoire ou la défaite et non le match nul.

Les disciplines de double introduisent une complexité supplémentaire dans la modélisation d'un système de rating. Comme énoncé dans la partie précédente, le classement est attribué à la paire et non aux joueurs individuellement. Lorsqu'une nouvelle paire se forme elle commence donc sans classement ni points historiques malgré les performances individuelles respectives. Cependant, même avec cette spécificité, de nombreux changements sont régulièrement opérés sur les paires de double. Il est courant de voir un même joueur jouer avec différents partenaires lors d'une même saison. A l'inverse, une paire peut jouer pendant plusieurs années ensemble sans changement. Il faut aussi noter que la performance d'une paire n'est pas un cumul de niveaux individuels et que la synergie entre les joueurs est très importante. La gestion des paires va donc constituer un enjeu majeur de notre étude.

Le circuit junior est assez développé avec des compétitions nationales et internationales en individuel et par équipe. La France est structurée avec des clubs avenir, des pôles espoirs et relève ayant pour but le développement de ces jeunes. Cependant, ce circuit présente de nombreuses catégories d'âges et est une étape de transition avant l'émergence des athlètes sur le circuit professionnel ce qui peut freiner le développement d'un système de ratings global (jeunes et séniors).

1.5 Contextualisation de la performance et positionnement de l'étude

L'étude qui va suivre s'articule donc autour de la contextualisation de la performance des athlètes et l'objectif de refléter un niveau réel de manière dynamique. Un système alternatif au classement BWF va donc être proposé pour comparer les classements, identifier si c'est un système qui pourrait permettre aux joueurs de mieux se situer, mais surtout pour aider à la décision les protagonistes de la performance de haut niveau.

Les systèmes de ratings proposés seront donc probabilistes et non cumulatifs comme le classement BWF. Il permettront donc de mieux prédire et détecter les progressions réelles ou creux de performances des athlètes.

Dans ce mémoire, l'objectif est de développer un système de rating interprétable et transparent qui devra s'adapter aux différentes disciplines du badminton (simple comme double). Ce système devra représenter le plus fidèlement les probabilités de victoire des joueurs et leur niveau relatif. Le but est donc de prendre en compte les différentes caractéristiques du sport et de son organisation pour rendre le système de rating le plus fidèle possible.

Trois différents modes de ratings seront mis en place dans un but comparatif. Le système Elo et le système Glicko, et un système adapté d'Elo sur lequel on essaiera de jouer

sur les paramètres pouvant influencer la performance des joueurs de badminton pour se rapprocher au mieux du niveau réel et donc des probabilités de victoires des différents joueurs.

Méthodologie

La construction et le développement de systèmes de rating adaptés au badminton nécessite une méthodologie claire de la collecte des données à l'évaluation des méthodes. On décrira par la suite le cheminement méthodologique commun aux trois méthodes en faisant un focus particulier sur le développement et l'implémentation de celles-ci.

2.1 Données utilisées et infrastructure technique

Les données utilisées comme base de travail pour la réalisation de ratings hebdomadaires sont les données officielles de matchs issues de la Fédération Internationale de Badminton (BWF). Dans l'étude qui va suivre, l'historique des matchs s'étend de janvier 2009 à début juillet 2025, ce qui permet une profondeur temporelle suffisante pour étudier l'évolution des performances des joueurs au fil du temps.

L'ensemble des données de matchs est stocké dans une base de données relationnelle nommée **Birdie**, spécifiquement conçue pour répondre aux besoins de suivi, d'analyse et de modélisation du haut niveau en badminton. Avant l'introduction des données de rating, la base de données contient 16 tables couvrant notamment les matchs de simple, de double, les différents joueurs, les paires de double, les pays, les confédérations continentales, ou encore les tournois et catégories d'événement. Cette base de données est mise à jour chaque mardi avec les résultats des tournois et les classements de la semaine précédente.

Toutes les tables sont mises en relation via un ensemble de clés primaires ou étrangères, telles que `MatchId`, `PlayerId` ou `PairId`, assurant l'intégrité référentielle et facilitant les jointures dans les requêtes analytiques. L'ensemble des relations entre les différentes tables de la base **Birdie** est disponible en annexe.

En termes de volume de données, non moins de 484 188 matchs couverts par la BWF ont été joués depuis 2009. Cela représente 218 173 matchs de doubles pour 266 015 matchs de simples. Le nombre de joueurs unique est de 48 585 pour 81 761 paires de double. Le fait d'avoir deux fois plus de paires de doubles ne correspond pas à un manque de joueurs dans la base mais au fait qu'un grand nombre de joueurs joue au moins un match en double lors de sa carrière et qu'un même joueur peut être dans plusieurs paires. Comme énoncé auparavant, il est très courant de changer de partenaire de double au Badminton. Enfin, 3 892 tournois sont répertoriés et répartis tout autour du monde.

La composante temporelle associée aux données est importante pour l'étude qui va suivre

car elle permet de valider le fait que le jeu de données est suffisamment conséquent et homogène entre 2009 et 2025 pour valider un modèle. Sur le graphique ci-dessous, on remarque que le circuit international et les données disponibles sur les matchs n'ont fait que se croitre entre 2009 et 2019, puis un creux dû à la pandémie de Covid-19 et donc à l'arrêt puis la reprise progressive des matchs, pour enfin reprendre un circuit plus conventionnel en 2022. Avec un minimum de plus de 12 000 matchs par année (hors période covid), et toutes les compétitions majeures le dataset est donc assez exhaustif pour notre étude.

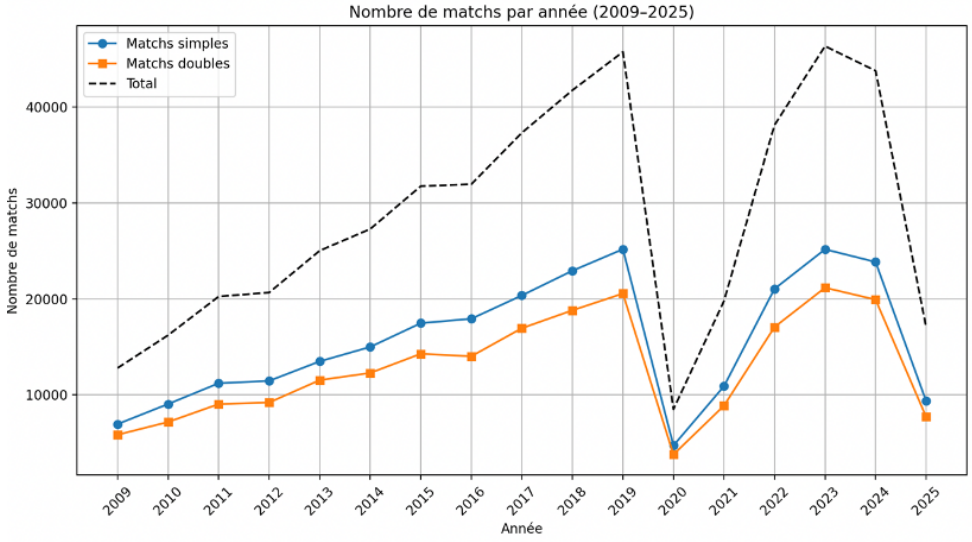


FIGURE 2.1 – Nombre de matchs par année de 2009 à 2025

Sur le plan technique, la consultation des données est faite via SQL sur Azure Data Studio. Le traitement et l'analyse des données ont par contre été réalisés à l'aide d'un environnement Python connecté à Birdie via *SQLAlchemy* et *pyodbc*. Les principaux outils utilisés seront alors la bibliothèque *pandas* couplée à *numpy* pour la manipulation des données, la bibliothèque matplotlib pour réaliser de première visualisation d'analyse. Enfin l'intégration et l'application réelle des ratings se fera via différentes visualisations *Plotly* dynamiques sur la plateforme *Dash* de la Fédération : *Badminton Manager*.

2.2 Prétraitement et qualité des données

Dans un premier temps il est important d'analyser la qualité des données. Pour anticiper tout biais potentiel liée à des erreurs de données, on vérifie l'ensemble des données de matchs. Le rating que l'on va développer se calcule à partir des données de matchs principalement contenues dans `MatchSingles` et `MatchDoubles`.
Ces deux tables comportent que très peu de valeurs manquantes. Toutes les colonnes indispensables à notre étude comme les joueurs de chaque rencontre, la date, le vainqueur, le tournois, la catégorie de jeu, les informations concernant les abandons et forfaits sont des données fiables à 100%. Toutefois, certaines informations secondaires telles que le round, la durée du match ou encore les scores par set présentent un taux de données manquantes relativement faible, avec respectivement 0,1 %, 3,5 % et 4 % de valeurs nulles. Pour contextualiser ces valeurs manquantes, seuls quelques tournois sont affectés par celles-ci

et principalement des tournois jeunes sur lesquels certaines informations ne sont pas traitées de la même manière. Par exemple, l'ensemble des rounds manquants provient des championnats du monde junior par équipe de 2023.

Les tables `MatchSingles`, `MatchDoubles`, `Player`, `Event`, `Country`, `Week`, `Tournament`, `TournamentCategory`, `RankingSingles` et `RankingDoubles` présentes dans le schéma de la base relationnelle Birdie en annexe seront les tables utilisées pour la mise en place du système de ratings.

Ces tables ont été importées dans leur intégralité dans notre environnement de travail Python grâce aux bibliothèques `SQLAlchemy` et `pyodbc`.En effet, nous aurons besoin de l'intégralité des données afin de pouvoir définir une logique et une validité temporelle des données.

Avant de procéder au calcul des ratings, un prétraitement est appliqué aux données des matchs afin de garantir leur cohérence, leur validité, et leur pertinence pour le développement d'un système adapté aux particularités du badminton.

D'abord nous allons comme pour le classement BWF développer un rating par type d'épreuve. En effet, il n'est pas pertinent de développer un rating global entre des individus qui ne peuvent pas se rencontrer en tournois. Donc pour chaque type de rating les données sont filtrées selon la discipline (MS, WS, MD, WD, XD). Un paramètre sera donné en entrée de la fonction pour sélectionner le mode de jeu sur lequel effectuer les calculs.

Ensuite, chaque match est enrichi d'informations sur le tournoi auquel il appartient. En l'occurrence : la catégorie du tournoi pour être en mesure d'utiliser le poids relatif des tournois ; et l'identifiant de semaine (WeekId) auquel est relié le tournoi pour regrouper les matchs par semaine et assurer une cohérence entre les identifiants des différentes tables.

Les matchs sont ensuite filtrés selon la catégorie de tournoi pour ne garder uniquement les matchs séniors et enlever certains tournois de test ou d'exhibition non pertinents pour le rating voulu. Le choix de retirer les tournois Jeunes du rating a été fait pour ne pas biaiser le modèle avec des résultats peu représentatifs du niveau réel des jeunes. En effet, par exemple un jeune champion d'Europe junior aura très probablement un très bon rating qui peut être basé uniquement sur des matchs Jeunes mais se rating ne reflète pas son niveau sur le circuit sénior. Certaines catégories d'âge sont plus denses que d'autres, c'est pourquoi un jeune ne peut pas démarrer le circuit sénior avec un rating basé sur ces matchs juniors. A l'inverse un talent générationnel jouera très rapidement sur le circuit sénior et engrangera vite les points pour refléter son niveau réel.

Aucun match actuellement n'a pas de date dans les tables `MatchSingles` et `MatchDoubles`, mais comme le rating est une donnée basée sur la temporalité, une jointure de sécurité est tout de même faite avec la table Week. Ce qui impliquerait pour les matchs ne disposant pas de date explicite (`Date`), une valeur de remplacement est attribuée en utilisant la date de fin de la semaine (`EndDate`) associée. Cela garantit qu'aucun match ne reste sans date de référence.

Enfin, le tableau des matchs est trié par date croissante pour certifier un traitement dans l'ordre chronologique des matchs essentiel au rating.

2.3 Modélisation des systèmes de ratings

Trois systèmes de rating vont être mis en place. Les deux premiers étant des système Elo et Glicko avec de très légères modifications pour qu'ils soient adaptés au badminton. Le dernier est une méthode Elo à laquelle on apportera quelques modifications pour tenter de modéliser au mieux la variation de performance des athlètes et donc leur niveau réel.

Le fonctionnement global du rating est le même pour les trois modes différents. Une fois le tableau des matchs correctement nettoyé et mis en forme, chaque identifiant de joueur unique est récupéré afin de créer une initialisation de leur rating. Pour les trois systèmes, le rating initial est fixé à 1500. Cette note évoluera ensuite positivement ou négativement en fonction des performances des athlètes.

Le rating de chaque joueur est voué à être calculé et mis à jour chaque semaine. On a donc décider de créer une logique de traitement des matchs par semaine.

Pour suivre la logique de la BWF, le choix d'un rating unique par paire de double a été fait. Chaque paire a donc un rating commun et non individuel et sera traité comme un joueur de simple. Certaines spécificités lors de l'initialisation propre à chaque type de rating seront décrites dans les parties respectives.

Enfin, en termes de structure chaque méthode sera développée pour le simple et pour le double dans deux codes pythons différents. On aura donc six scripts qui prendront en entrée les semaines à traiter et les paramètres des modèles.

2.3.1 Méthode Elo

La méthode Elo est la plus simple et la plus classique des méthodes à implémenter. Elle prend en arguments trois paramètres : K, L et k que nous définirons par la suite. Cette méthode reprend les bases d'Elo avec un K fixe et une fonction de probabilité de victoire similaire.

Pour que cette méthode soit tout de même adaptée au badminton, l'objectif est d'identifier les différences avec les échecs pour modifier uniquement les paramètres essentiels. En l'occurrence l'issue du match ne peut pas être nulle mais un abandon ou un forfait sur blessure est possible. De même, une blessure ou une inactivité peut entrainer une baisse non négligeable de niveau (physique et technique) qu'il peut être judicieux de prendre en compte. Comme au badminton le classement BWF prend en compte 10 résultats de tournois sur 52 semaines glissants, très peu de joueurs sont inactifs sans être blessé car leur classement leur permet de participer aux plus grandes compétitions. Dans la suite, une proposition de gestion de l'inactivité sera faite puis validée ou non dans la partie résultats.

Comme décrit précédemment, la manière de distribuer les points du modèle Elo est basée sur la différence entre l'issue du match (1 en cas de victoire, 0.5 pour le match nul et 0 pour la défaite) et la probabilité de victoire du joueur avant le match.

$$\Delta R = K \cdot (S - E) \quad \text{où} \quad E = \frac{1}{1 + 10^{\frac{R_{\text{opponent}} - R_{\text{player}}}{400}}} \tag{2.1}$$

Pour faire correspondre cette distribution des points aux issues possibles d'un match de Badminton, un score S de 1 sera attribué lors d'une victoire, et 0 lors d'une défaite. Un pourcentage du nombre de points distribués dans le cas classique sera alors attribué pour les abandons et les forfaits. Lorsqu'un des deux joueurs déclare forfait avant le match, seulement 8% des points seront attribués aux deux joueurs. Dans le cas d'un abandon pendant le match, 50% des points seront distribués. Par exemple, si deux joueurs A et B ont des ratings assez éloignés pour qu'il y ait 50 points en jeu sur le match alors :

— Si A gagne sans abandon ni forfait, alors A gagne 50 points et B en perd 50 ;
— Si A gagne par forfait, alors A gagne 4 points et B en perd 4 ;
— Si a gagne par abandon, alors A gagne 25 points et B en perd 25 ;

L'inactivité est une variable plus compliquée à identifier. Afin de refléter la baisse potentielle de performance liée à une période prolongée sans compétition, on ajoute au système elo une pénalité d'inactivité hebdomadaire. Cette pénalité est calculée pour chaque joueur n'ayant pas joué depuis un certain nombre de semaines, et s'applique de manière progressive selon un modèle logistique. Deux coefficients, L et k permettront d'ajuster la décroissance en fonction des résultats de tests.

On a ensuite défini un seuil de semaines d'inactivités au-delà duquel les joueurs commenceront à subir la pénalité logistique. Ce seuil est défini en fonction du nombre de compétition et donc de matchs par années qui est croissant au fil du temps. Entre 2009 et 2011 ce seuil est fixé à 13 semaines, de 2011 à 2017 le seuil est à 11 semaine puis après 2017 le seuil est fixé à 10 semaines. Il est tout de même important de prendre en compte la période Covid pendant laquelle les joueurs étaient contraints à l'inactivité. Les ratings sont alors gelés pendant l'arrêt complet des tournois puis une reprise progressive avec un coefficient multiplicateur entre 50% et 90% est appliqué jusqu'à la semaine du 3 mars 2022 où le circuit est revenu à la normale.
De ce fait le nombre de semaines d'inactivité dépasse le seuil autorisé, la pénalité de rating est appliquée à l'aide de la formule suivante :

$$\text{penalty} = \frac{L}{1 + 2e^{-k \cdot \Delta t}} * MultiplicateurCovid \tag{2.2}$$

— Δw : le nombre de semaines d'inactivité au-delà du seuil autorisé,
— L : l'amplitude maximale de la pénalité,
— k : un paramètre contrôlant la vitesse de décroissance de la pénalité,
— `covid_multiplier` : un multiplicateur correctif permettant d'atténuer temporairement l'effet de la pandémie sur les ratings.

Le fonctionnement est similaire pour les simples et les doubles. Cependant comme chaque paire a un rating unique et que les paires peuvent régulièrement changer, un système d'initialisation de rating pour le double a été pensé différemment du simple.

En effet, comme énoncé ci-dessus, pour le simple chaque joueur est initialisé à 1500. En double, si la paire n'a pas de rating enregistré, l'historique de chacun des deux joueurs est étudié. La paire sera initialisée à la moyenne des maxima de ratings atteints par chacun des joueurs dans la discipline dans leur historique récent. C'est-à-dire que pour un double dame par exemple, si une nouvelle paire se crée entre une joueuse A et une joueuse B, alors on prend le rating maximum des deux joueuses sur l'année précédente en double dame et on fait la moyenne de ces ratings. Dans le cas, où l'un des joueurs n'a pas de rating historique des deux dernières années alors il est initialisé à 1500. Le schéma ci-dessous illustre un exemple de cas de figure entre deux joueuses fictives de double dame.

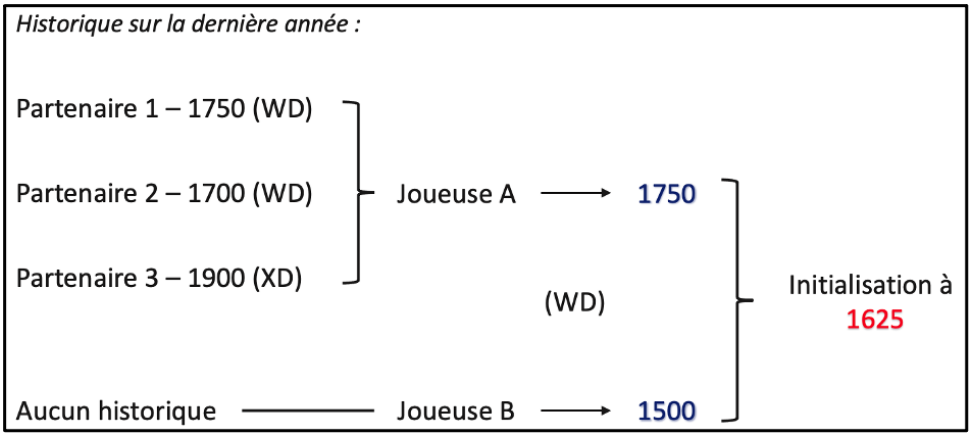


FIGURE 2.2 – Initialisation des paires sans historique commun

Enfin, concernant le traitement des ratings, celui-ci se réalise donc semaine par semaine. Pour chaque semaine, le système Elo met à jour les ratings des joueurs en fonction des matchs disputés. La première étape consiste à extraire l'ensemble des matchs joués pendant la semaine considérée et à récupérer les ratings initiaux de l'ensemble des joueurs. Ensuite, le système applique l'éventuelle pénalité d'inactivité aux joueurs n'ayant pas participé à une rencontre depuis plusieurs semaines.
Comme le veut la méthode Elo, pour chaque match les ratings actuels des deux adversaires sont utilisés pour calculer la probabilité de victoire à l'aide de la formule classique du modèle Elo. Puis les ratings sont mis à jour en fonction du résultat réel du match via la formule 2.4. Les ajustements spécifiques à la méthode sont appliqués dans les cas de forfait ou d'abandon, afin de limiter l'impact sur les ratings.
Les changements de rating sont stockés pour chaque match, permettant une analyse détaillée des dynamiques de performance. À l'issue de la semaine, un classement hebdomadaire est également généré, en attribuant un rang aux joueurs ayant été actifs sur les 52 dernières semaines. Enfin, les nouveaux ratings sont intégrés à l'historique global pour chaque joueur. Ce mécanisme hebdomadaire garantit une mise à jour progressive, sensible au contexte, et permet de suivre l'évolution de la performance des athlètes dans le temps de façon fine et réactive.

2.3.2 Méthode Glicko-2

La méthode Glicko-2 est pensée pour être une version améliorée d'Elo en prenant en compte une incertitude sur le rating des joueurs et une mise à jour des variables d'un

joueur inactif. De plus la méthode est plus complexe et les paramètres sont moins adaptables à notre contexte. C'est pourquoi bien que la méthode globale reste similaire, celle-ci va être traitée différemment du système elo étudié ci-dessus.

La méthode Glicko-2 implémentée repose sur trois variables fondamentales associées à chaque joueur : le rating (R), la déviation du rating (RD) et la volatilité (σ). En parallèle, comme pour la méthode elo précédente, trois paramètres additionnels gèrent l'effet de l'inactivité : L, k, et `K_didnotcompete`.

Comme Elo, Glicko-2 est un modèle initialement pour les échecs. Il est aussi possible de l'adapter au contexte badminton bien que certains aspects soient moins facilement contrôlables que pour Elo. La méthode proposée se voudra comme pour la version ci-dessus assez proche de la version originale de Glicko-2.

Pour garder une échelle similaire à Elo et rester adapté à l'échelle Glicko-2 chacun des joueurs ou des paires n'ayant pas d'historique de matchs est initialisé à un rating de 1500. Plus précisément, chaque joueur se voit attribuer un vecteur contenant son historique de ratings hebdomadaires (s'il en contient un), sa déviation (RD), et sa volatilité (σ). De la même manière que pour elo la mise à jour des scores se fait semaine par semaine. Les mises à jour de ratings, de déviation et de volatilité se font en utilisant la bibliothèque glicko2, qui applique l'algorithme itératif de recalcul proposé par Mark Glickman. À chaque match, la probabilité de victoire des joueurs est calculée selon la formule adaptée de Glicko.

L'algorithme Glicko prévoit que la déviation RD augmente progressivement quand un joueur ne joue pas. Lors de l'implémentation, on a choisi de pouvoir contrôler cette augmentation grâce au paramètre `K_didnotcompete`, qui simule plusieurs itérations de la mise à jour pour inactivité de la RD du joueur correspondant à l'équation (numéro de l'équation). En effet, comme le traitement des matchs est réalisé par semaine et que Glickman ne documente pas la temporalité à laquelle est mise à jour la RD suite à l'inactivité d'un joueur, l'introduction d'une variable contrôlant cet effet semblait judicieux.

En complément, pour refléter la baisse de forme potentielle due à l'inactivité, le système intègre une pénalité logistique similaire à celle utilisée dans le modèle Elo :

$$\text{penalty} = \frac{L}{1 + 2e^{-k \cdot \Delta t}} * MultiplicateurCovid \tag{2.3}$$

L'application conjointe d'une pénalité sur le rating et d'une augmentation de l'incertitude (RD) peut engendrer une instabilité dans la notation au moment du retour d'un joueur, ce qui peut biaiser l'interprétation de ses performances initiales post-reprise. C'est pourquoi cette pénalité est soumise à une phase de test et de validation dans la suite de l'étude. Il faut aussi garder à l'esprit que la version originale de Glicko-2 n'est qu'un cas particulier avec L=0 et `K_didnotcompete`=1.

La gestion des paires de double a par contre été pensée différemment pour Glicko-2. Comme les ratings initialisés sont accompagnés d'une rating deviation (RD) élevée, ils sont très volatils en début d'exercice. Ils mettront ensuite moins de temps à atteindre un niveau proche de leur niveau réel qu'avec le classement elo. On initialise alors les paires de

doubles comme les joueurs de simples pour ne pas cumuler les effets de la rating deviation et d'une initialisation potentiellement trop élevée.

Le traitement hebdomadaire des ratings se fait de la même manière que pour elo sauf que la mise à jour des attributs des joueurs (R, RD, vol) se fait via la classe glicko2. Pour chaque semaine traitée, les matchs sont récupérés, les ratings et la RD sont mis à jour en cas d'inactivité grâce à la fonction de glicko-2 et à la fonction de pénalité logistique mises en place. Ensuite les matchs sont traités individuellement pour mettre à jours les valeurs de rating et de paramètres associés pour chaque joueur.
Chaque mise à jour hebdomadaire stocke les valeurs recalculées dans un historique (PlayerRating, RD, σ) et alimente une table hebdomadaire pour le suivi longitudinal des performances. Enfin, à l'issue de chaque semaine, les joueurs actifs sur les 52 dernières semaines sont classés selon leur rating.

2.3.3 Méthode Elo adaptée

La troisième et dernière méthode implémentée dans cette étude repose sur une adaptation du système elo classique, conçue pour intégrer l'importance relative des matchs dans la distribution des points. Cette méthode, reprend l'intégralité des principes fondamentaux du système Elo exposé précédemment : prise en compte de l'inactivité, ajustement selon les abandons et forfaits, et mise à jour hebdomadaire des ratings.

Cette méthode a été imaginée pour mesurer l'importance relative des rencontres sans pour autant réduire la stabilité des classements. Comme vu précédemment, il est assez simple de modifier l'importance d'un match grâce au facteur K de la formule de distribution de points d'Elo (rappelée ci-dessous)

$$\Delta R = K_{eff} \cdot (S - E) \quad \text{où} \quad E = \frac{1}{1 + 10^{\frac{R_{\text{opponent}} - R_{\text{player}}}{400}}} \tag{2.4}$$

Où K_{eff} est la valeur effective de K adaptée au match en question.

En effet, faire varier le facteur K permet de distribuer plus ou moins de points lors d'une rencontre. C'est d'ailleurs ce qui a été imaginé lors de la gestion des forfaits et des abandons avec un pourcentage du nombre de points distribués. L'objectif de cette variation du facteur K est d'accorder plus d'importance ou moins d'importance aux matchs comme on peut le retrouver dans le classement BWF actuel. En effet, on peut imaginer qu'une finale de Jeux Olympiques a plus de poids qu'un premier tour de championnat national.

On a donc construit une table de facteurs en fonction de l'importance des tournois et une table en fonction de l'importance des rounds. Les coefficients appliqués ont été définis sur la base de la densité compétitive de chaque type de tournoi, estimée à partir du niveau moyen des joueurs engagés (selon leur classement), et de la structure de points du classement BWF, qui reflète l'importance attribuée par la fédération à chaque tour et niveau de tournoi. Un coefficient entre 0 et 1 a donc été attribué pour les différents niveau de tournois. (A noter qu'un coefficient de 0 a été attribué aux tournois Jeunes qui ne sont pas pris en compte dans le rating.) De la même manière, une table avec un coefficient

entre 0,9 et 1 a été mise en place en fonction du round dans le tournoi.

TABLE 2.1 – Facteurs par niveau de tournoi

Facteur	**Niveau de tournoi**
1.00	JO / Championnats du monde
0.94	Finals de fin de saison
0.92	Open de très haut niveau (Super 1000)
0.88	Open de haut niveau (Super 750)
0.80	Open intermédiaire (Super 300 / 100)
0.40	Open de faible niveau (IS, IC)
0.18	Autres tournois
0.00	Jeunes / Juniors / Universitaires

TABLE 2.2 – Facteurs par round atteint

Round	**Facteur**
Tour de qualification	0.90
32^{e} de finale	0.94
16^{e} de finale	0.94
8^{e} de finale	0.94
Quart de finale	0.96
Demi-finale	0.98
Finale	1.00

À partir de ces éléments, un coefficient f[0,1] a été attribué à chaque combinaison Tournoi x Round, reflétant de manière continue l'enjeu relatif de chaque rencontre. Par exemple, une demi-finale de Super 750 (facteurs 0.88 et 0.98) entraînera un facteur K effectif de :

$$K_{eff} = K * 0.88 * 0.98 = 0.86K \tag{2.5}$$

Cela permet d'attribuer plus de points qu'un match de premier tour dans un tournoi mineur, tout en gardant la même formule de base.

Bien que cette méthode paraisse tout à fait cohérente, il est possible qu'elle freine la progression des joueurs dans la zone basse du classement et qu'à l'inverse les meilleurs joueurs mondiaux creusent des écarts démesurés avec le niveau moyen des joueurs internationaux. En effet, les joueurs participant aux tours avancés des plus grands tournois sont ceux ayant un rating élevé et à l'inverse les joueurs participants aux tournois plus modestes ont un rating faible et plus de mal à marquer des points.

C'est pourquoi, afin de rendre le système encore plus réaliste, un facteur de pondération dépendant du rating actuel a été introduit dans le calcul. Ce facteur permet de réduire progressivement l'impact des matchs sur les joueurs très bien classés, et à l'inverse de

maintenir une dynamique d'évolution plus rapide pour les joueurs en progression. La fonction utilisée est la suivante :

$$f(R) = 1 + \frac{6}{1 + 2^{\frac{R-1500}{63}}} \tag{2.6}$$

- R est le rating actuel du joueur,
- $f(R)$ est un facteur multiplicatif appliqué à la variation de points,
- la constante 1500 représente le rating de référence neutre,
- le dénominateur $2^{(R-1500)/63}$ permet une transition rapide autour de 1500, avec un lissage logarithmique.

L'effet de la fonction sera alors tel qu'un joueur autour de R=1500, aura un coefficient de 4, ce qui rend les résultats impactant. C'est aussi une manière d'encourager la progression des nouveaux joueurs ou nouvelles paires sur le circuit. Pour un joueur très bien classé (ex. R $>$ 1800), f(R)$<$1, ceci limite la montée incontrôlée des ratings. Enfin pour les joueurs moins bien classés (ex. R $<$ 1300), f(R) tendra vers 7, cela encourage les progressions rapides.

Ce mécanisme couplé aux ajustements faits sur le facteur K permet de freiner artificiellement la progression des très hauts niveaux, évitant une inflation incontrôlée du rating tout en conservant une réactivité pour les joueurs moins expérimentés ou en pleine ascension. La gestion des paires se passe de la même manière que pour la méthode Elo avec une initialisation comme décrite f2.6.
Finalement, la mise à jour se fait de manière hebdomadaire comme pour les autres méthodes. Le processus est identique à celui d'Elo hormis un ajustement des ratings opéré via 2.6 avant chaque match et que le K utilisé est un K effectif 2.5 dépendant de la rencontre.

2.4 Méthodes d'évaluation et de validation

Afin de comparer objectivement les différents systèmes de rating développés (Elo, Glicko-2 et Elo adaptée), une stratégie d'évaluation rigoureuse a été mise en place, combinant validation temporelle, métriques quantitatives classiques, analyses de calibration et méthodes visuelles.

2.4.1 Validation temporelle

Une validation dite temporelle a été utilisée, en accord avec la nature évolutive et chronologique du rating. Pour chaque méthode, une période d'entraînement a d'abord permis de stabiliser les ratings des joueurs à partir d'un historique suffisamment riche. Cette période initiale a été définie sur plusieurs années entre 2009 et 2014. A partir de 2014, on considèrera les ratings comme stabilisés. Cette période d'initialisation sera suivie d'une période de test, sur laquelle les probabilités de victoire prédites ont été confrontées aux résultats réels des matchs. Cela permet une évaluation réaliste du pouvoir prédictif des modèles sur des données non vues au moment du calcul des ratings.

2.4.2 Métriques quantitatives

Pour évaluer la qualité des modèles implémentés, trois métriques issues de la littérature sur la prévision probabiliste et en particulier la modélisation binaire ont été retenues : le Brier score, l'Accuracy et la Logarithmic loss (log loss). Ces trois mesures ont un objectif commun mais une façon de mesurer la précision du modèle différentes. Elles sont donc complémentaires, offrant un regard différent sur la performance du modèle.

Accuracy

L'accuracy est sûrement la métrique la plus intuitive des trois. Elle mesure simplement le pourcentage de prédictions correctes en considérant que l'on prédit un joueur vainqueur lorsque sa probabilité de victoire est supérieure à 0,5 et inversement pour les défaites avec des joueurs ayant une probabilité de victoire inférieure à 0,5.

Une formulation simple de l'accuracy peut être la suivante :

Accuracy $= \frac{\text{Vrais Positifs+Vrais Négatifs}}{\text{Ensemble de valeurs}}$

Cette métrique est simple est facile à interpréter, cependant pour un modèle probabiliste elle ne tient pas en compte du degré de confiance des probabilités prédites. L'accuracy reste néanmoins un indicateur robuste de la justesse générale du modèle.

Log Loss

La Log loss est une mesure quantifiant la qualité des probabilités en pénalisant sévèrement les prédictions fausses et très confiantes. Elle est définie par :

$$\text{Log loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{p}_i) + (1 - y_i) \cdot \log(1 - \hat{p}_i)]$$

Cette métrique est donc très sensible aux mauvaises prédictions très confiantes. Par exemple, si le joueur A a une probabilité de victoire de 0,94 et que l'inverse se produit alors cela pénalisera grandement la métrique. Cette métrique est largement utilisée pour tester et valider des modèles prédictifs de machine learning. Dans notre cas, la log loss est pertinente pour conserver un modèle stable qui ne prédit pas des issues de matchs toujours aux alentours de 0 ou 1. Cependant, il faut prendre en compte l'aspect humain et aléatoire de la performance sportive. Les contre-performances arrivent régulièrement dans le monde du sport et c'est aussi vrai au badminton, c'est pourquoi il faut utiliser intelligemment la Log loss et la combiner à d'autres indicateurs pour valider la précision et la pertinence des modèles.

Brier Score

Le Brier Score introduit par Brier en 1950[13], est une mesure de l'erreur quadratique moyenne entre les probabilités prédites (p_i) et les issues observées (y_i). Il est défini par la formule suivante :

Avec N le nombre d'observations.

Comme les deux métriques précédentes, le Brier Score renvoie une valeur entre 0 et 1. L'objectif étant d'avoir une modélisation renvoyant un Brier Score proche 0, ce qui définirait un modèle discriminant mais juste. C'est une métrique qui n'est pas trop pénalisante. En effet, contrairement à l'Accuracy, qui ne considère que la classe prédite, le Brier Score évalue la qualité de calibration des prédictions probabilistes mais de manière moins pénalisante que la Log loss. C'est un aspect essentiel dans un contexte de compétition sportive où l'incertitude fait partie intégrante de la performance. Son usage est courant dans l'évaluation de modèles de rating probabilistes linéaires, en particulier dans le sport où il est utilisé par exemple dans le papier de Giovanni Angelini al.

Les méthodes visuelles sont un bon moyen d'illustrer les métriques précédentes, de valider visuellement la calibration de nos modèles et de comprendre plus en détails leurs fonctionnements. Ces représentations ne visent pas à présenter les résultats des méthodes mais à s'assurer qu'elles se comportent conformément aux hypothèses théoriques.

2.4.3 Méthodes visuelles

Les méthodes visuelles sont un bon moyen d'illustrer les métriques précédentes, de valider visuellement la calibration de nos modèles et de comprendre plus en détails leurs fonctionnements. Ces représentations ne visent pas à présenter les résultats des méthodes mais à s'assurer qu'elles se comportent conformément aux hypothèses théoriques.

Courbe de Calibration

Dans un premier temps, des analyses visuelles de calibration peuvent être réalisées. En effet, une première vérification porte sur la calibration des probabilités prédites. Une courbe de calibration, obtenue en comparant les probabilités de victoire prédites avec les résultats effectivement observés, permet d'évaluer la qualité de ces prédictions. L'ajout d'une droite de calibration idéale ($y = x$) et la pondération des points selon le nombre de matchs dans chaque intervalle permettent une interprétation plus rigoureuse de cette relation.

Matrice de confusion

La matrice de confusion permet de visualiser la capacité du modèle à prédire correctement les issues de matchs (victoire/défaite) en définissant un seuil de probabilité à partir duquel les prédictions sont prises en compte. Elle permet de visualiser le taux de vrais positifs, faux positifs etc... C'est en quelques sorte une visualisation de l'Accuracy avec un seuil de probabilité choisi.

Courbe d'évlution du rating

Une visualisation de l'évolution du rating de certains joueurs au fil du temps avec la connaissance de leurs résultats permet de détecter les effets attendus liés à l'activité compétitive, aux périodes d'inactivité ou aux performances en tournoi.

Histogramme de distribution des ratings

Pour vérifier la convergence des ratings tout en illustrant la répartition de ceux-ci, on trace la distribution des ratings. Cette visualisation peut aussi permettre de comparer plusieurs modèles en termes de discrimination.

Autres visualisations descriptives

Enfin, d'autres visualisations plus descriptives comme l'impact de l'écart de rating sur la variation de rating ou l'écart de points remportés dans le match par les deux joueurs en fonction du rating peuvent permettre de comprendre l'utilité opérationnelle du rating et d'imaginer l'avenir opérationnel en termes de contextualisation de la performance.

2.5 Sélection des constantes

Suite à l'implémentation des trois différentes méthodes de ratings, un certain nombre de constantes sont à évaluer. Pour chacune des constantes faisant partie des modèle elo et glicko-2, des valeurs approximatives ou des intervalles sont conseillés dans la littérature pour que les modèles soient les plus fiables possible.

Dans le cadre de notre étude, le système de rating doit correspondre au jeu de données et être fiable dans le contexte du badminton professionnel. C'est pourquoi le choix de tester empiriquement les constantes a été fait.

Pour chacun des modèles entre 40 et 254 combinaisons de constantes ont été testées. Pour chaque combinaison de constantes, les ratings étaient calculés de 2009 à 2025. Comme décrit précédemment un système d'entraînement et de test a été mis en place avec une période d'adaptation du rating entre 2009 et 2014, puis les métriques (Brier Score, Accuracy et Log Loss) étaient stockées avec la combinaison de constantes associées dans un fichier csv traitable en aval.

Le traitement de ses fichiers a donc été fait en comparant l'effet de chacune des constantes et en sélectionnant les combinaisons qui offrent le plus de fiabilité à chacun des modèles.

2.6 Utilisation future

Une fois la ou les méthodes développées et validées, les données de ratings seront implémentées dans la base de données Birdie. Une mise à jour hebdomadaire des ratings sera alors réalisée pour l'ensemble des joueurs.
Pour cela, une création de tables dédiées au stockage des ratings hebdomadaire, aux informations match par match et par type d'épreuve sera nécessaire. Dans un second temps une logique d'automatisation du calcul hebdomadaire via des scripts pythons sera développée.

Cette mise en base permettra une exploitation transversale des ratings, en particulier pour la création de modules analytiques dans Badminton Manager au service de la performance. Des visualisations dynamiques seront imaginées puis développées pour apporter une valeur ajoutée au support d'aide à la décision de la fédération.

Résultats

Cette section présente les résultats issus de la mise en œuvre des trois méthodes de calcul de rating détaillées précédemment : le système Elo standard, la méthode Glicko-2 et la méthode Elo adaptée.

L'évaluation des performances de ces trois méthodes repose à la fois sur l'analyse des métriques permettant de quantifier la précision et la fiabilité des méthodes mais aussi sur des méthodes visuelles permettant de comparer et d'interpréter les comportements dynamiques des systèmes.

De manière générale dans cette section, nous présentons les résultats pour les matchs de simples. Une évaluation similaire a été menée pour les matchs de doubles, dont certains des résultats importants pour étudier l'impact des changements de paires par exemple seront explicités au fil de la section. Par défaut les visualisations seront donc l'illustration de chacune des méthodes pour les matchs de simple, les visualisations similaires pour les doubles seront disponibles en annexes.

Comme précisé précédemment, les résultats présentés ci-dessous sont les résultats des trois algorithmes après 2014 pour que les modèles puissent se stabiliser sur les premières années.

3.1 Résultats par système et sélection des constantes

La sélection des constantes est une étape cruciale pour l'implémentation future des méthodes de ratings.

Dans un premier temps, nous allons analyser individuellement chaque méthode afin de choisir la meilleure combinaison de paramètres en fonction de leurs influence relative sur les métriques.

Pour les trois méthodes de ratings, nous voyons que logiquement les trois métriques (Brier Score, Accuracy et Log Loss) sont corrélées. Pour la sélection des constantes, une visualisation en trois dimensions (égales aux trois métriques) a été réalisée grâce à la bibliothèque Plotly de Python. Cette visualisation permettra de choisir le meilleur triplet de métriques et de récupérer les paramètres associés en récupérant les informations sur le point en question. Par soucis de lisibilité, dans ce mémoire, les représentations seront en deux dimensions avec une dimension supplémentaire en nuances de couleurs.

Les valeurs testées pour chaque paramètre de chaque méthode, sont basées sur les valeurs conseillées dans la littérature ainsi que l'expérience de l'opérateur ayant réalisé les tests.

3.1.1 Méthode Elo

Pour cette première méthode, les paramètres testés sont K, L, et k.
Les valeurs testées pour K sont : 26, 28, 30, 32, 34, 35, 38, 40, 55, 70, 85, 90, 95 et 100
Les valeurs testées pour L sont : 0, 1, 2, 5, 10
Les valeurs testées pour k sont : 0,05 et 0,1

Comme présenté précédemment, le graphique ci-dessous permet de sélectionner le meilleur compromis entre les métriques. Le point en haut à gauche représente la meilleure combinaison de paramètres valant : K=90, L=0 et k=0,05 ou k=0,1.

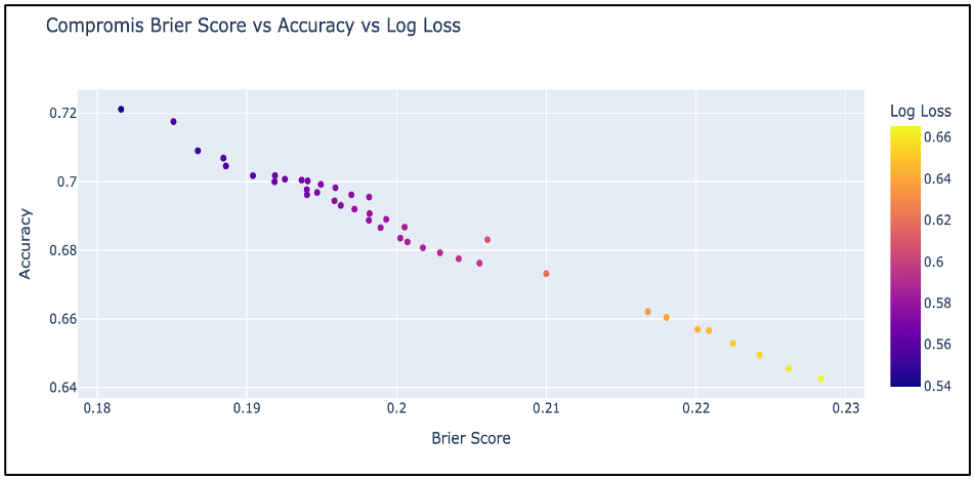


FIGURE 3.1 – Analyse multi-critères des hyperparamètres pour Elo : compromis entre Log Loss, Accuracy et Brier Score

Des valeurs supérieures à 90 pour K ont été testées manuellement mais aucune amélioration notable était présente au niveau des métriques. La calibration des probabilités de victoires par rapport aux fréquences observées devenait d'ailleurs moins bonne au-delà de K=90 à cause d'une volatilité des ratings trop importante, c'est pourquoi une bonne supérieure à 90 a été fixée. La même remarque est valable pour la méthode Elo Adaptée.

3.1.2 Méthode Glicko-2

Pour cette méthode, les paramètres testés sont RD, sigma, L, k et `K_didnotcompete`.
Les valeurs testées pour RD sont : 200, 250, 300, 350 et 400
Les valeurs testées pour sigma sont : 0.01, 0.03, 0.06, 0.09, 0.1, 0.12, 0.15, 0.16, 0.18, et 0.2
Les valeurs testées pour L sont : 0, 1, 2, 3 et 5
Les valeurs testées pour k sont : 0,05
Les valeurs testées pour `K_didnotcompete` sont : 1, 2, 4 et 10

Pour cette méthode, grâce à la visualisation graphique ci-dessous les valeurs retenues sont : RD=350, sigma=0.16, L=0, k=0.05 et `K_didnotcompete`=1.

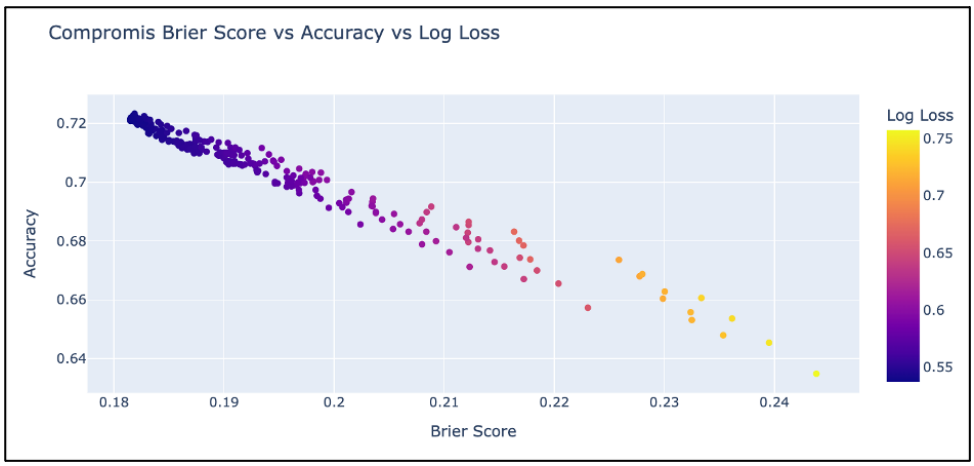


FIGURE 3.2 – Analyse multi-critères des hyperparamètres pour Glicko : compromis entre Log Loss, Accuracy et Brier Score

3.1.3 Méthode Elo adaptée

Pour cette dernière méthode, les paramètres testés sont K, L, et k.
Les valeurs testées pour K sont : 26, 27, 28, 30, 31, 32, 34, 38, 55, 70, 85, 90, 95 et 100
Les valeurs testées pour L sont : 0, 1, 2, 5, 10
Les valeurs testées pour k sont : 0.02, 0.05 et 0.1

Pour cette méthode, de la même manière graphique, les valeurs de K=90, L=0 et k=0,05 ont été retenues.

3.1.4 Récapitulatif global

Stefani, Ray (2011)[7] qui a réalisé une étude des différents systèmes de ratings adaptés au sport donne la valeur de 68% d'Accuracy comme valeur de référence d'un bon système de rating. Avec les choix de constantes ci-dessus chacun des systèmes de rating prédit le bon résultat d'une confrontation à plus de 68%. Les valeurs des métriques permettent donc de faire une première validation des méthodes. L'approche graphique complémentaire va permettre de comparer les méthodes et d'apporter une validation définitive sur la précision, la stabilité et la dynamique de celles-ci.

Tableau récapitulatif des paramètres pour chacune des méthodes en fonction de la catégorie :

ELO							
K	L	k	Brier score		Accuracy	Log-loss	
90	0	0.05	0.1816		0.7212	0.5396	
85	0	0.05	0.1871		0.7226	0.5540	
Glicko-2							
RD	Sigma	L	k	K_didnotcompete	Brier Score	Accuracy	Log-loss
350	0.16	0	0.05	1	0.1787	0.7267	0.5313
350	0.16	0	0.05	1	0.1983	0.6873	0.5820
ELO adaptée							
K	L	k	Brier score		Accuracy	Log-loss	
90	0	0.05	0.1894		0.7103	0.5595	
85	0	0.05	0.1878		0.7185	0.5551	

FIGURE 3.3 – Récapitulatif des combinaisons de paramètres optimaux pour chacun des modes

Une des choses importantes à noter est que l'ensemble des paramètres L est fixé à 0 et que son influence était négative sur la précision des ratings. Les systèmes implémentés ne possèderont donc pas de décroissance de rating due à l'inactivité. Il est aussi important que le paramètre k n'a pas d'influence lorsque L est fixé à 0.

3.2 Analyse graphique comparative

Une fois les constantes choisies à partir des métriques, les résultats des trois algorithmes sont passés en revue par des méthodes graphiques afin de valider la calibration des probabilités mais aussi d'analyser le comportement et la dynamique des méthodes sur divers points.

3.2.1 Calibration des probabilités

Calibration des probabilités

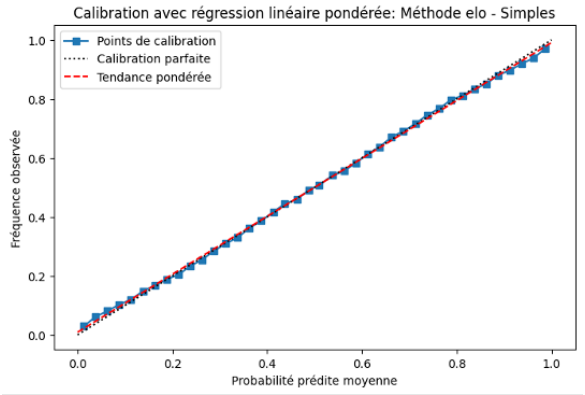
Afin d'évaluer la qualité des probabilités de victoire produites par les différents modèles, une calibration probabiliste a été réalisée. Celle-ci vise à déterminer si les prédictions du modèle reflètent correctement les fréquences observées de victoire. Autrement dit, un joueur à qui l'on attribue une probabilité de victoire de 70% devrait effectivement gagner environ 70% du temps sur l'ensemble des matchs similaires.

Les figures ci-dessous représentent les courbe de calibrations obtenues pour les matchs de simple. Les prédictions ont été regroupées en intervalles réguliers (bins) de 0.02, puis la fréquence de victoire réelle a été calculée pour chaque intervalle. La droite noire en pointillés correspond à une calibration parfaite ($y = x$) : les probabilités prédites sont égales aux fréquences observées. La courbe rouge est une régression linéaire pondérée des points bleus, utilisée pour estimer la tendance globale de nos modèles. Enfin les points bleus représentent la moyenne des probabilités sur chaque intervalle et les fréquences empiriques associées.

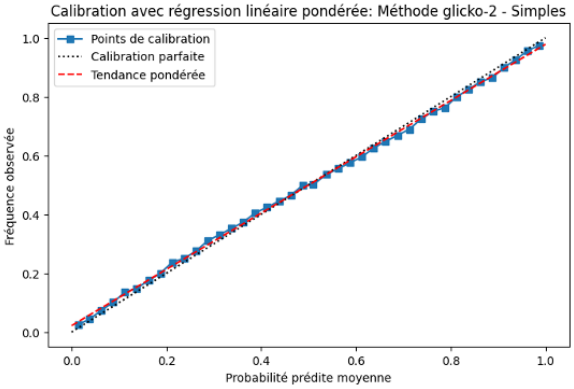
La calibration est l'un des éléments principaux de notre analyse, en effet la précision et la validité des probabilités de victoire avant match est un des enjeux majeurs de notre étude. C'est pourquoi il est important d'analyser les calibrations en simple et en double car le

changement régulier de partenaire et l'initialisation des paires pourrait avoir un impact non négligeable sur la calibration.

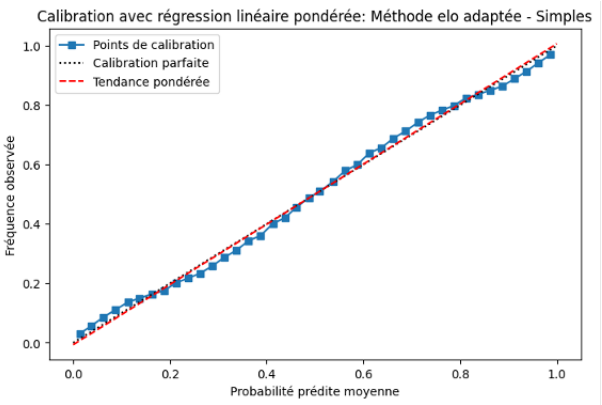
Dans un premier temps, voici les résultats graphiques pour le simple :



Méthode Elo



Méthode Glicko-2

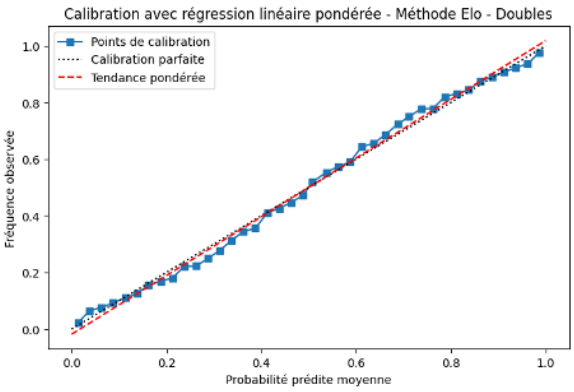


Méthode Elo adaptée

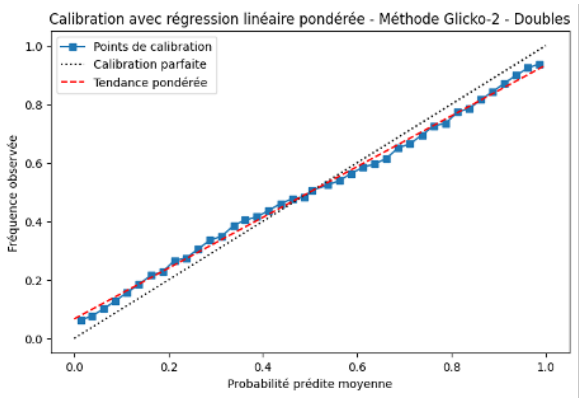
FIGURE 3.4 – Comparaison des calibrations selon les méthodes de rating – Simples

trois méthodes sont très bien calibrées car on remarque que la courbe bleu et la courbe rouge sur les trois méthodes suivent quasiment parfaitement la droite d'équation y=x en pointillés noirs. On remarquera tout de même un léger décalage pour la méthode Elo Adaptée.

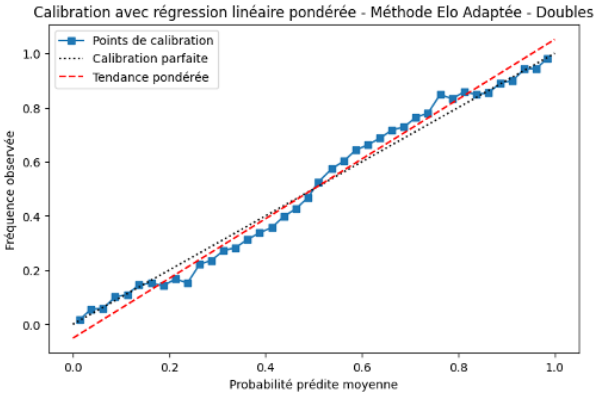
Le résultats pour le double sont les suivants :



Méthode Elo



Méthode Glicko-2



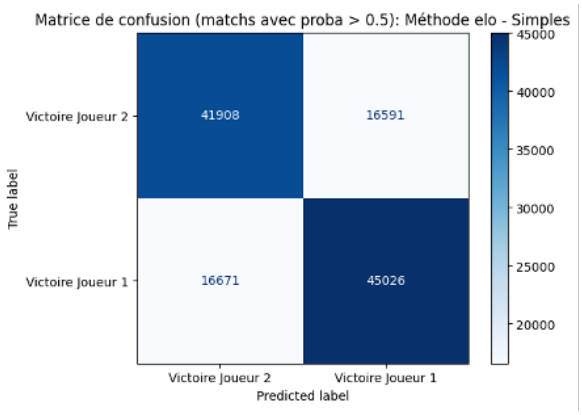
Méthode Elo adaptée

FIGURE 3.5 – Comparaison des calibrations selon les méthodes de rating – Doubles

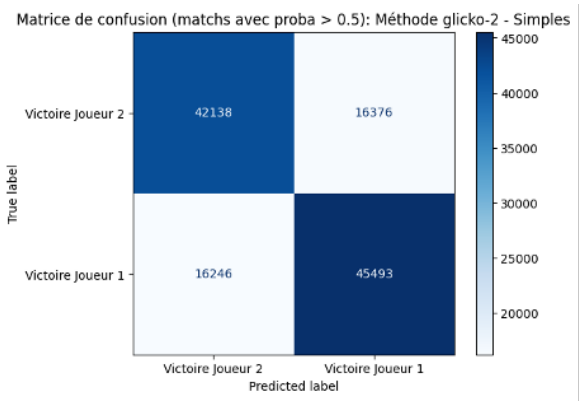
Les trois méthodes sont globalement bien calibrées car on remarque que les courbes bleues et rouges sur les trois méthodes suivent grossièrement la droite d'équation y=x en pointillés noirs. On remarquera tout de même de une légère sous estimation des probabilités réelles pour Glicko-2.

3.2.2 Matrice de confusion

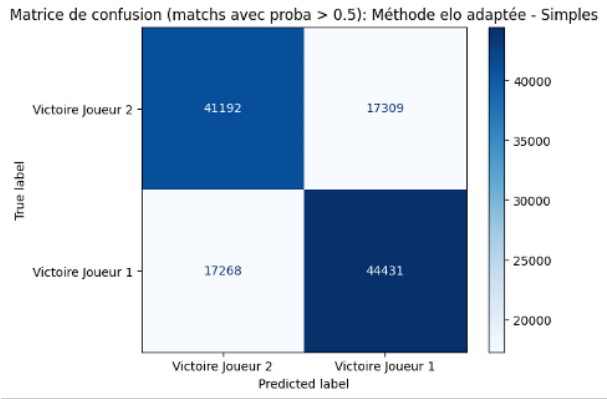
Afin d'évaluer les performances de classification binaire des trois modèles de ratings, des matrices de confusions ont été construites à partir des prédictions de probabilité, en appliquant un seuil de 0.5 pour déterminer le joueur favori. Ces matrices permettent de visualiser les victoires correctement prédites (visibles dans la diagonale), mais aussi les erreurs de classification (hors diagonale) lorsque le joueur pressenti comme gagnant a finalement perdu.



Méthode Elo



Méthode Glicko-2



Méthode Elo adaptée

FIGURE 3.6 – Comparaison des matrices de confusion selon les méthodes de rating – Simples

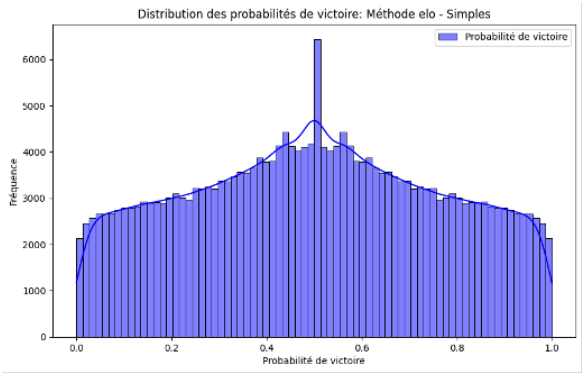
Les résultats de cette matrice permettent d'illustrer et de confirmer les valeurs d'accuracy énoncées précédemment. Par exemple pour le rating Elo, le nombre de vrais positifs sur le nombre total de matchs disputés vaut 0.72 comme l'accuracy de ce modèle.

On peut également noter que les systèmes de rating sont tous équilibrés de la même manière entre différentes prédictions. Il y a donc une cohérence inter-modèles qui est intéressante à relever.

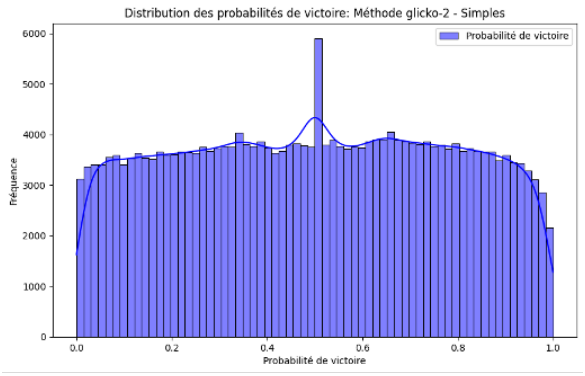
Les résultats et la logique sont similaires pour les doubles. Les visualisations correspondantes sont présentes en annexe.

3.2.3 Histogrammes de distribution

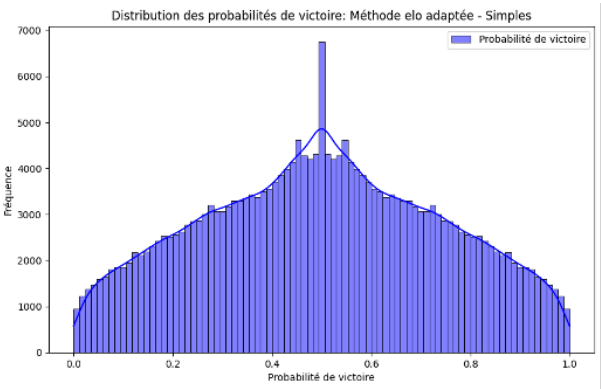
Afin de mieux comprendre la manière dont chaque algorithme attribue des probabilités de victoire, nous avons analysé la distribution des probabilités produites pour l'ensemble des matchs de simple. Les histogrammes ci-dessous présentent la fréquence des probabilités prédites, réparties entre 0 et 1, pour chacun des trois modèles pour les matchs de simple.



Méthode Elo



Méthode Glicko-2



Méthode Elo adaptée

FIGURE 3.7 – Comparaison des histogrammes de distribution de probabilités selon les méthodes de rating – Simples

Chacune de ces distributions présente un pic en 0,5 qui est dû à l'initialisation des ratings à une valeur commune (1500). Par ailleurs ce pic est centré ce qui confirme que les probabilités sont généralement bien calibrées. On observe que le modèle Elo Adapté produit des probabilités centrées autour de 0.5, tandis que Glicko-2 affiche une meilleure discrimination des niveaux et une plus grande confiance en s'éloignant davantage des extrêmes. Le modèle Elo propose un compromis entre les deux, traduisant une amélioration du pouvoir discriminant tout en restant relativement bien calibré.

Pour mieux comprendre la répartition des scores attribués par les systèmes de notation, nous avons examiné la distribution des ratings des joueurs actifs au cours de la semaine du 08 juillet 2025 pour les trois algorithmes.

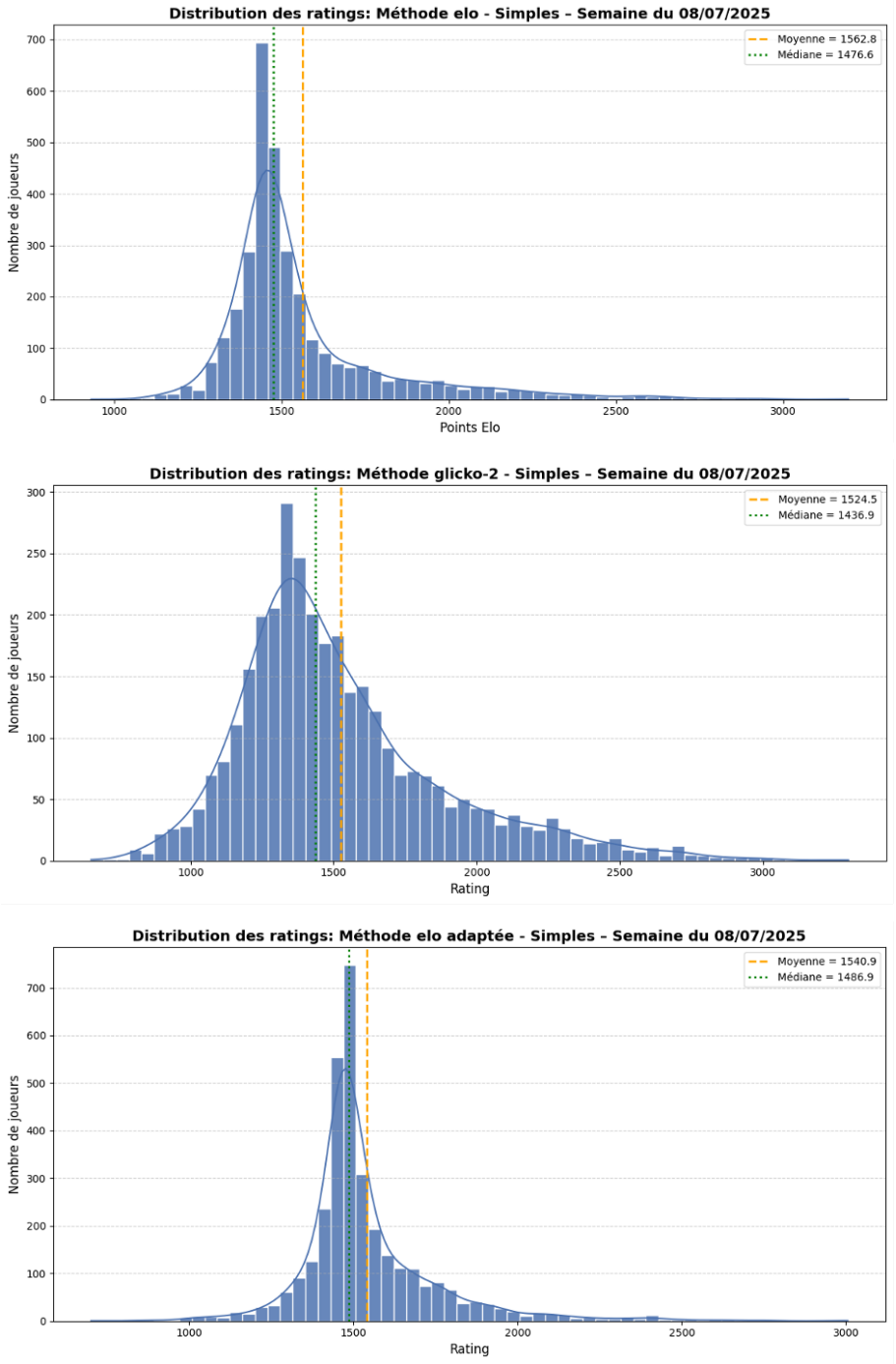


FIGURE 3.8 – Comparaison des distributions de ratings selon les méthodes – Simples

L'analyse de ces histogrammes permet une visualisation de la répartition des ratings des joueurs et la discrimination de niveau des joueurs par les algorithmes.

En termes de répartition, tous les algorithmes ont des ratings allant de 500/1000 à environ 2500/3000 avec une moyenne autour de 1550 et une médiane aux alentours de 1450. Ces visuels témoignent d'une répartition plutôt lisse des ratings. Aucun des algorithmes n'a de ratings allant vers des valeurs démesurément élevées ce qui confirme la convergence des différents algorithmes appliqués au badminton avec les paramètres choisis.

La discrimination des niveaux des joueurs par les algorithmes est la seconde information capitale fournie par ces graphs. Les visuels confirment la conclusion fournie par les histogrammes de distribution de probabilités : la méthode Glicko-2 est la plus discriminante, quand Elo et Elo Adaptée ont des ratings plus resserrés.

3.2.4 Autres visualisations descriptives

L'un des éléments clés d'un système de rating est la manière dont il ajuste les scores en fonction du résultat, conditionnellement à l'écart de niveau estimé entre les joueurs. Les figures suivantes analysent la relation entre l'écart de rating initial et la variation de points obtenue à l'issue d'un match. La couleur représentant la densité dans chaque zone des graphs.

graphicx subcaption float

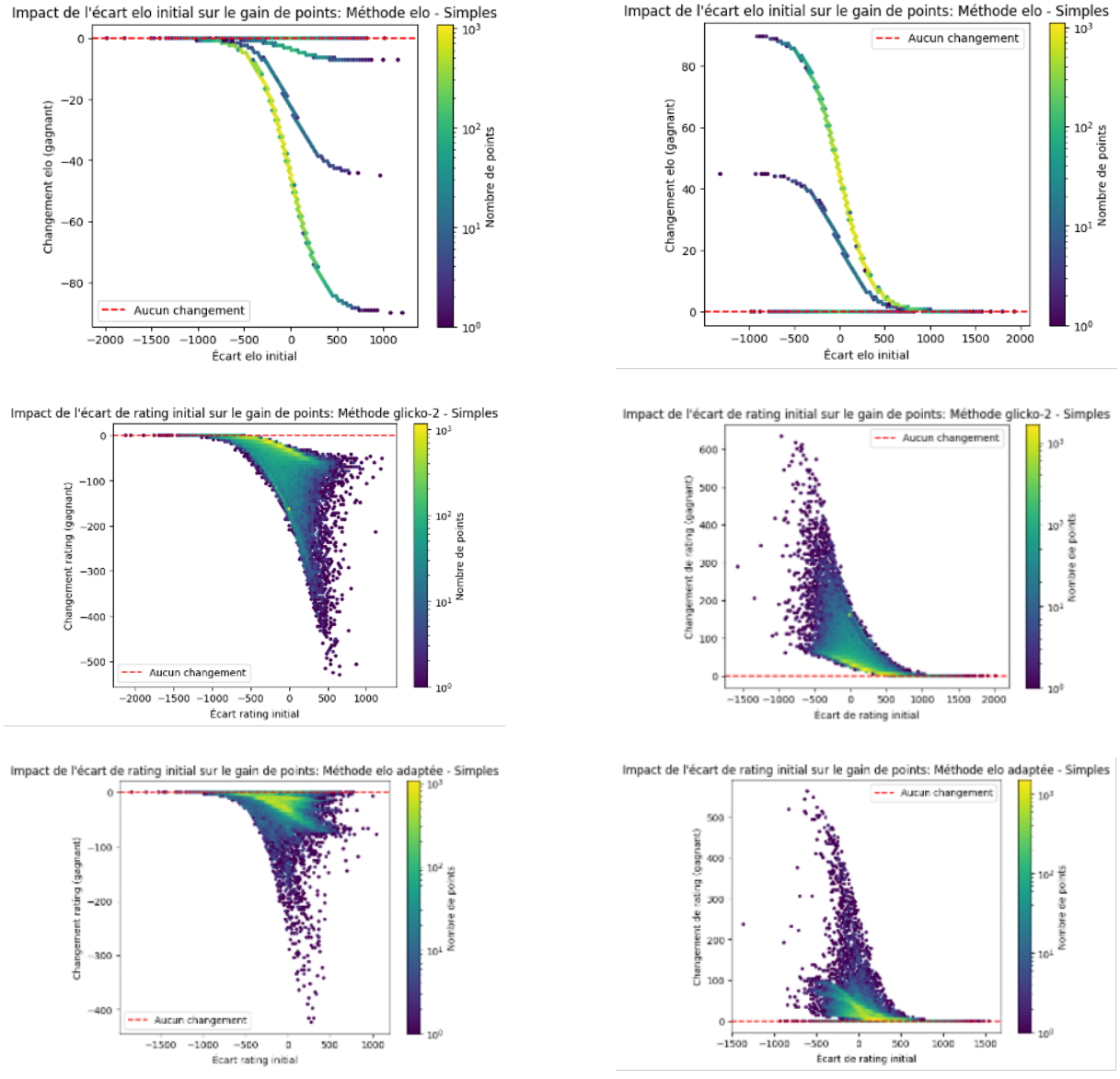


FIGURE 3.9 – Impact de l'écart de rating initial sur le gain de points selon la méthode et le résultat

Lors de l'analyse de ces visualisations, on retrouve bien la répartition de points et les méthodes préalablement décrites.

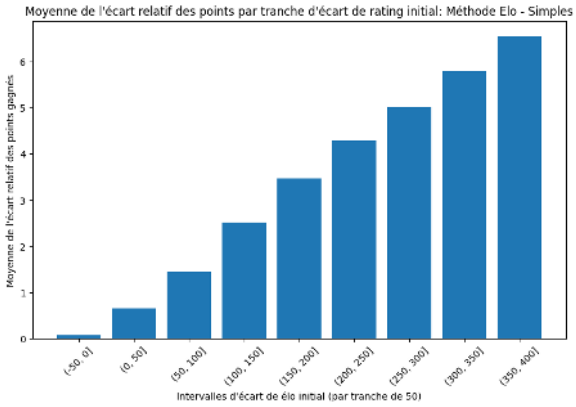
D'un côté, la méthode Elo attribue un nombre de points entre 0 et 90 au vainqueur (les différentes branches de la représentation décrivant les abandons et les forfaits coefficientées à un pourcentage du nombre de points dans le cadre d'une issue classique).

Les méthodes Glicko-2 et Elo Adaptée quant à elles offrent une répartition beaucoup plus variée du nombre de points distribués après chaque match. On remarque d'ailleurs que pour certains cas exceptionnels avec des différences de ratings de plus de 500, certains athlètes peuvent voir leurs ratings évoluer de 500 points. Pour la méthode Glicko-2 cela implique une grande RD, tandis que le coefficient K associé au match doit être élevé pour la méthode Elo Adapté.

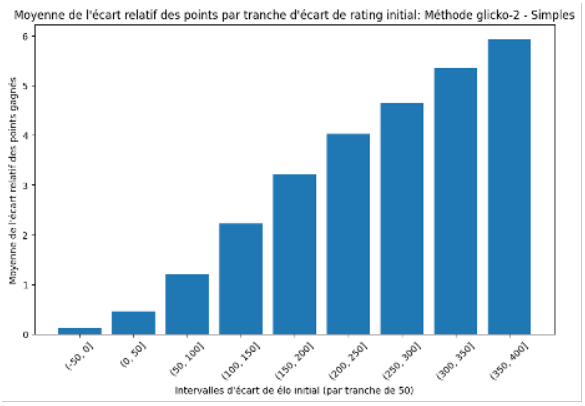
Cette variété du nombre de points distribués provient de l'incertitude RD et la volatilité sigma associés à chaque joueur pour la méthode Glicko-2. Dans le cas de la méthode Elo Adapté, chaque match distribue aussi des points uniques en fonction du tournoi, du tour et des ratings des deux joueurs.

Cette hétérogénéité dans la distribution des points permet ainsi aux systèmes Glicko-2 et Elo Adapté de mieux refléter la complexité des dynamiques compétitives. Elle offre une sensibilité accrue aux performances atypiques, tout en tenant compte du contexte du match (incertitude, importance du tournoi, statut des joueurs). En ce sens, ces systèmes offrent un outil d'évaluation plus nuancé que le modèle Elo standard.

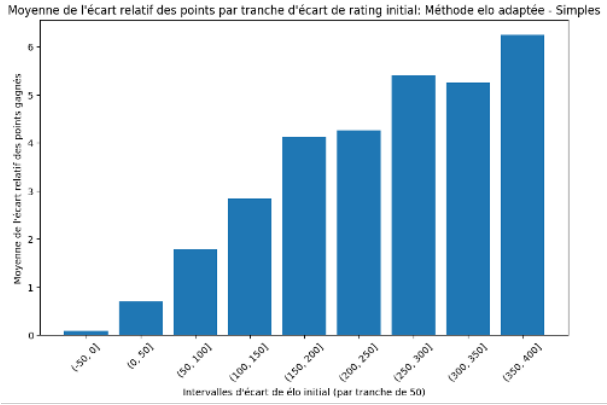
La dernière visualisation proposée est une première étape vers la contextualisation de la performance. Elles illustrent la relation entre l'écart de rating initial et l'écart relatif du nombre de points marqués par le vainqueur par rapport à son adversaire dans le match, pour différentes méthodes (Elo, Glicko-2, Elo Adaptée).



Méthode Elo



Méthode Glicko-2



Méthode Elo adaptée

FIGURE 3.10 – Comparaison de l'écart relatif de points par tranche initiale d'elo selon les méthodes de rating – Simples

De manière assez logique, plus l'écart de rating est important entre les deux joueurs, plus le score du match sera déséquilibré en moyenne. Par exemple pour la méthode Elo, un écart de rating de 300 entrainera un écart moyen de 6 points (21/18-21/18 par exemple). Les méthodes Elo et Glicko-2 ont un accroissement qui est plus linéaire que la méthode Elo Adapté. C'est un des facteurs pouvant être pris en compte pour décrire le fait que les méthodes Elo et Glicko-2 ont plus de facilité à évaluer l'écart de niveau relatif des joueurs.

En résumé, cette section a permis de confronter les trois systèmes de rating sur un ensemble d'indicateurs quantitatifs et visuels. Les résultats obtenus montrent que les trois méthodes atteignent un niveau de précision satisfaisant, avec une calibration globalement fiable. Néanmoins, des différences notables apparaissent dans leur comportement dynamique, leur capacité à discriminer les niveaux ou à réagir aux performances atypiques. Ces éléments soulignent les forces et limites propres à chaque système, et posent les bases d'une réflexion plus approfondie sur leur utilisation dans un cadre décisionnel.

Discussion

Les trois systèmes de ratings testés (Elo, Glicko-2 et Elo Adapté) ont montré de bonnes performances en termes de prédictions et de calibration, avec tout de même des différences notables sur leur comportement dynamique. L'objectif de cette partie est d'interpréter, de mettre en perspective, de discuter des implications, des limites, et des perspectives des résultats décrits précédemment dans le contexte du badminton international.

3.3 Interprétation approfondie des résultats et mise en contexte

3.3.1 Précision et calibration

La calibration des probabilités et la précision de celles-ci était un enjeu majeur pour l'étude menée. Elle signifie que le modèle est performant, qu'il prédit des probabilités de victoires cohérentes avec le niveau relatif des joueurs. Au-delà d'utiliser les ratings dans une logique de classement, cette précision au niveau des probabilités est essentielle pour pouvoir contextualiser les performances, et conseiller les acteurs principaux de la performance française (entraîneurs nationaux, directeurs de pôles ou de performances etc...)

Parmi les modèles testés, globalement la calibration est meilleure en simple qu'en double. En effet, le double représente un réel enjeu avec les changements de paires, les joueurs ayant plusieurs paires actives et les joueurs évoluant dans plusieurs disciplines. Autant de facteurs rendant l'issue des matchs plus difficile à prévoir. C'est dans cette logique qu'on a développé le système d'initialisation des nouvelles paires en fonction de leurs résultats antérieurs.

Concernant le simple, Glicko-2 affiche la meilleure calibration (confirmé par la proximité de la courbe de calibration à la courbe théorique $y = x$). Elo et Elo Adapté restent également très bien calibrés avec un léger avantage pour la méthode Elo. Sur le plan de la précision prédictive, les trois méthodes dépassent le seuil de 68% d'accuracy, considéré comme une référence minimale dans la littérature (Stefani, 2011)[7]. Cela suggère qu'elles sont toutes capables d'apporter une valeur ajoutée dans le suivi et l'évaluation du niveau des joueurs.

3.3.2 Discrimination des niveaux

Un autre enjeu majeur d'un système de ratings est sa capacité à discriminer le plus finement possible les niveaux relatifs. En effet, un système peut être très bien calibré, si

toutes les probabilités de victoire se situent autour de 0,5, cela ne permet pas de démontrer le réel niveau des joueurs. Cette capacité est bien illustrée par la distribution des ratings, mais aussi par la distribution des probabilités de victoire produites par chaque méthode.

En ce sens, que ce soit pour les matchs de simple ou de double, Glicko-2 se distingue clairement, en générant des probabilités plus tranchées, souvent plus proches de 0 ou 1, là où Elo Adaptée tend à produire des valeurs plus centrées. Cette plus grande discrimination est liée au fait que Glicko-2 intègre une mesure d'incertitude (RD) et une volatilité individuelle (σ) dans son calcul. Cela permet au modèle de réagir fortement lorsqu'un jeune joueur ou un joueur peu connu réalise une performance notable. Si cela permet une meilleure reconnaissance des talents émergents, cela peut aussi conduire à des surinterprétations ponctuelles du niveau de certains joueurs. Par exemple, un joueur irrégulier réalisant une victoire contre le numéro un mondial verra son rating exploser. À l'inverse, Elo Adaptée, bien que moins discriminant, offre une répartition plus conservatrice des scores, qui peut s'avérer plus robuste dans des contextes à forte variabilité.

3.3.3 Réactivités des ratings

Pour continuer l'argumentation de la partie précédente, la manière dont un système de rating réagit à un résultat (notamment lorsqu'il est inattendu) est un critère déterminant pour son adoption. Ici, la méthode Elo Adaptée se distingue par une grande adaptabilité à la discipline : la quantité de points attribués dépend du tournoi, du tour, de l'écart de niveau, et des coefficients spécifiques choisis. De plus, une réactivité accrue est conférée au système lors d'une contre-performance sur un match important. Cela lui confère une forte sensibilité contextuelle qui peut être un point fort de notre modèle.

A l'inverse, la méthode Glicko-2 associe les paramètres de variabilité au joueur avec les variables RD et sigma. Cela permet au système de réagir plus fortement sur des performances inattendues pour un joueur indépendamment du contexte.

Enfin la méthode Elo est en simple comme en double la plus conservatrice des trois systèmes de rating. La réactivité est constante et cela entraine une faible adaptation du système dans les contextes à forte variabilité. Dans le cas du badminton, ces contextes de fortes variabilités sont réservés à l'émergence de nouveaux joueurs ou à la création de nouvelles paires. Ces contextes étant restreints et l'incertitude sur les nouvelles paires étant en partie traitée via l'initialisation, la méthode Elo classique bien que peu réactive peut s'avérer intéressante.

3.3.4 Application aux doubles

Comme évoqués plusieurs fois dans ce document, la gestion des paires de doubles et l'application des méthodes disponibles aux doubles n'est pas évidente du fait des particularités de la discipline. La rotation fréquente des paires complexifie l'estimation du niveau de chacune des paires. Une paire nouvellement formée peut surprendre par ses performances, ou au contraire ne pas être représentative du niveau réel de chacun des membres.

Selon les méthodes la gestion des paires est différente, d'un côté aucune adaptation n'est faite pour Glicko-2 tandis qu'une initialisation de rating est proposée pour les méthodes basées sur le système Elo. On remarque par ailleurs qu'elle est bien meilleure pour les méthodes Elo et Elo Adaptée que pour la version Glicko-2. En effet, les métriques pour la méthode Glicko-2 en double sont en dessous des standards affichés par les autres méthodes. De plus, la calibration des probabilités est sous-estimée et la distribution des ratings et des probabilités de victoire est très resserrée autour respectivement de 1500 et de 0,5. Les métriques étant par ailleurs similaires au simple pour Elo et Elo Adapté, les probabilités étant plutôt bien calibrées ; cela confirme que notre méthode d'initialisation des paires pour les méthodes Elo est cohérente et correspond à la discipline.

3.3.5 Gestion de l'inactivité

Le choix de fixer le paramètre L = 0 annulant ainsi la perte de rating liée à l'inactivité se justifie par le fait que les résultats tant en termes de métrique que d'interprétation graphique n'étaient pas concluants. Le modèle se comporte mieux sans cette mise à jour. Dans un sport où l'enchaînement des compétitions peut être perturbé par des blessures ou des choix stratégiques qui n'altèrent pas forcément le niveau réel de l'individu cela peut être une manière de mieux refléter le niveau réel. Cela permet en l'occurrence d'éviter des diminutions artificielles du niveau de joueurs momentanément absents, mais peut aussi masquer des déclins réels de performance. Une autre manière de modéliser l'inactivité est peut-être à envisager.

3.4 Validité du modèle dans notre contexte

Comme énoncé dans la partie précédente, l'étude de Stefani, Ray (2011) [7] qui a réalisé un benchmark des différents systèmes de ratings adapté au sport donne la valeur de 68% d'accuracy comme valeur de référence d'un bon système de rating. Dans le cadre de cet article, en choisissant les constantes vues précédemment et en réalisant les ajustements proposés chacun de nos modèles vérifie cette condition.

Cependant, prendre l'accuracy comme unique métrique de référence est un peu fragile. Selon l'étude réalisée, l'ensemble des méthodes peuvent tout de même être validées car elles présentent toutes de points forts selon les situations.

3.5 Limites de notre étude

Bien que cette étude analyse en profondeur chacune des méthodes mises en place, certaines limites et pistes d'améliorations peuvent être relevée.

Tout d'abord, les méthodes de ratings mises en place partent de l'hypothèse que des matchs sont indépendant. Cela ignore les dynamiques de forme (fatigue, confiance, blessures), l'influence de la programmation (enchaînements de matchs), ou encore les effets tactiques ou psychologiques entre joueurs connus.

Le résultat d'un match est la seule information utilisée pour décrire l'issue du match. Le score détaillé, la durée, ou les circonstances particulières sont ignorés. C'est une des

choses qui peut conduire à des mises à jour de ratings qui ne reflètent pas réellement la réalité sportive bien que ces aléas fassent partie de cette réalité. Il pourrait être intéressant d'essayer d'intégrer des statistiques de performances à nos modèles.

De même, pour le développement de la méthode Elo Adaptée, un certain nombre de facteurs ont été pris en compte pour décrire le contexte du match mais les modèles n'intègrent pas les conditions de jeu, le lieu, ou les enjeux spécifiques non détectable par le round ou le tournoi (match contre un rival, attache particulière à un tournoi etc...) Cela peut donc biaiser les évaluations, en particulier dans les sports comme le badminton où les conditions varient selon les compétitions.
Un des facteurs limitant et aussi la non prise en compte de l'inactivité. L'idée initiale était d'intégrer ce facteur à nos méthodes, finalement le choix de ne pas l'intégrer a été fait pour privilégier la qualité des prédictions et du classement.
Enfin, les joueurs peu actifs sont difficiles à évaluer et rendent la notation des autres joueurs plus instable. Ce facteur est normalement limité par les règles BWF actuelles qui prennent en compte 10 tournois par an dans le classement mondial. Cependant, dans une logique de démocratisation d'un système de rating cela pourrait être un frein.

3.6 Perspectives et applications concrètes

Chacune des méthodes a été implémentée dans la base de données Birdie de la fédération. L'intégration de ces algorithmes dans le système fédéral permet d'avoir une vision différente des performances des athlètes. Certains exemples d'application et d'utilisation sont détaillés ci-dessous.

Tout d'abord, cela permet de faire un suivi longitudinal en temps réel de la performance des athlètes. Par exemple, en temps réel les responsables et l'athlète sont en mesure de détecter des phases de stagnation, de déclin ou un pic de forme. Cela n'est a priori pas utile pour un entraineur suivant son joueur au quotidien, mais pour le responsable de la performance ou autre acteur cela permet d'avoir une vision en temps réel de chacun des joueurs. Il peut d'ailleurs être intéressant de faire le lien avec les périodes de charge ou les choix d'entraînements. Enfin au terme de la saison, l'étude du rating associé au plan d'entrainement permet de décrire les phases d'évolution du joueur ce qui est intéressant pour la définition des plans de performance individuel de chacun des athlètes.

Ce système permet aussi de comparer le modèle français à l'international de manière plus objective. Des comparaisons entre les joueurs français et joueurs étranger à rating égal peuvent être réalisées. Un positionnement global du badminton français peut aussi être imaginé à partir d'un certain nombre de joueurs.

Dans la préparation de match d'un athlète, la dynamique de rating adverse peut être ajoutée aux facteurs étudiés.

Enfin, à plus long terme, le système de rating pourrait être un outil d'aide à la sélection si celui-ci est applicable aux catégories juniors.

Cette liste non exhaustive permet d'avoir une première vision des applications et utilisations possibles des systèmes de rating développés.

Conclusion

Ce mémoire a permis de développer et d'évaluer plusieurs modèles de rating appliqués au badminton international, en s'appuyant les données réelles des matchs du circuit BWF. Grâce à une approche rigoureuse de traitement de ces données puis de modélisation et d'analyse, nous avons pu mettre en lumière les limites des systèmes de classement traditionnels et proposer des alternatives dynamiques et contextualisées pour évaluer les joueurs.

L'intégration de certains facteurs de contextualisation a permis d'affiner les estimations de performance, tout en assurant une meilleure stabilité du système. Les modèles testés (Elo, Glicko-2, Elo modifié) ont été comparés à l'aide de métriques standard de calibration et de qualité de prédiction (log loss, Brier score, accuracy), mettant en avant la robustesse des méthodes adoptées.

Au-delà des résultats numériques, ce travail propose un cadre méthodologique réplicable pour d'autres disciplines, fondé sur une vision systémique et évolutive de la performance. Il offre également des pistes vers un outil d'aide à la décision potentiellement précieux pour les entraîneurs, analystes et responsables de la performance, en apportant un éclairage objectif et continu sur le niveau réel des joueurs, indépendamment des classements traditionnels.

Bibliographie

[1] Elo, A. The Rating of Chessplayers, Past and Present. **1978**, Base du système Elo classique.

[2] Glickman, P. M. E. The Glicko Rating System. 1999 ; Présentation du système Glicko et Glicko-2.

[3] Glickman, P. M. E. Example of the Glicko-2 system. 2022 ; Document explicatif de Glicko-2 avec un exemple pas à pas.

[4] Djebbi, A. Analyse des Dynamiques de Jeu en Badminton de Haut Niveau par Apprentissage Supervisé. *STAPS - Revue internationale des sciences du sport et de l'éducation physique* **2020**, *129*, 65–78.

[5] Sonas, J. Revisiting the K-factor in Chess Ratings. **2023**, Proposition d'un K optimal autour de 24 pour une meilleure réactivité.

[6] Microsoft Research TrueSkill Ranking System. `https://www.microsoft.com/en-us/research/project/trueskill-ranking-system/`, 2007 ; Accessed : 2025-07-14.

[7] Stefani, R. The Methodology of Officially Recognized International Sports Rating Systems. *Journal of Quantitative Analysis in Sports* **2011**, *7*, 1–15.

[8] FIFA Classement mondial de la FIFA. 2024 ; `https://inside.fifa.com/fr/fifa-world-ranking`, Consulté le 14 juillet 2025.

[9] FIBA Guide complet du classement FIBA 3x3 (version française). 2020 ; `https://www.ffbb.com/sites/default/files/homologation_tournois3x3/fiba_3x3_rankings_guide_complet_francais_2020.pdf`, Consulté le 14 juillet 2025.

[10] World Rugby Explication du classement mondial World Rugby. 2024 ; `https://www.world.rugby/rankings/explanation`, Consulté le 14 juillet 2025.

[11] World Athletics Règles du classement mondial d'athlétisme - principes de base. 2024 ; `https://worldathletics.org/world-ranking-rules/basics`, Consulté le 14 juillet 2025.

[12] Federation, B. W. World Ranking System. `https://bwfbadminton.com/rankings/`, 2021 ; Règlement officiel du classement mondial BWF.

[13] Brier, G. W. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review* **1950**, *78*, 1–3.

4.1 Annexes

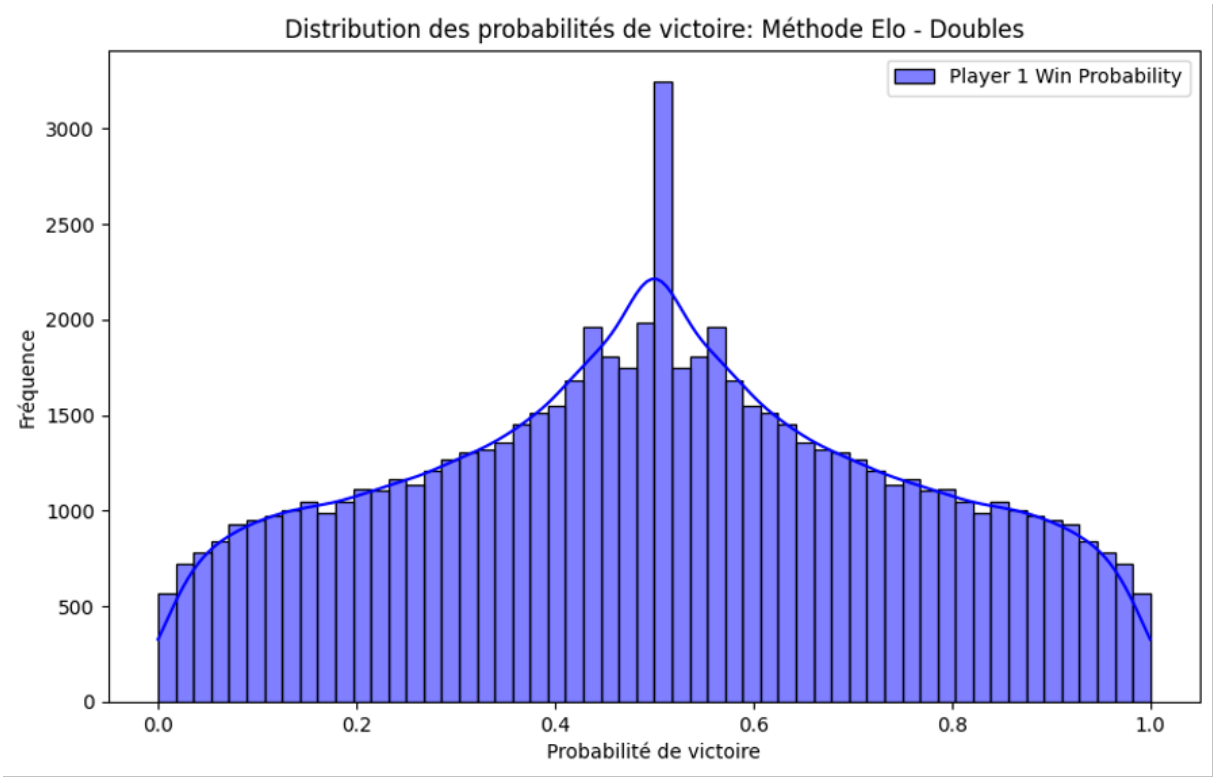


FIGURE 4.1 – Annexe 1

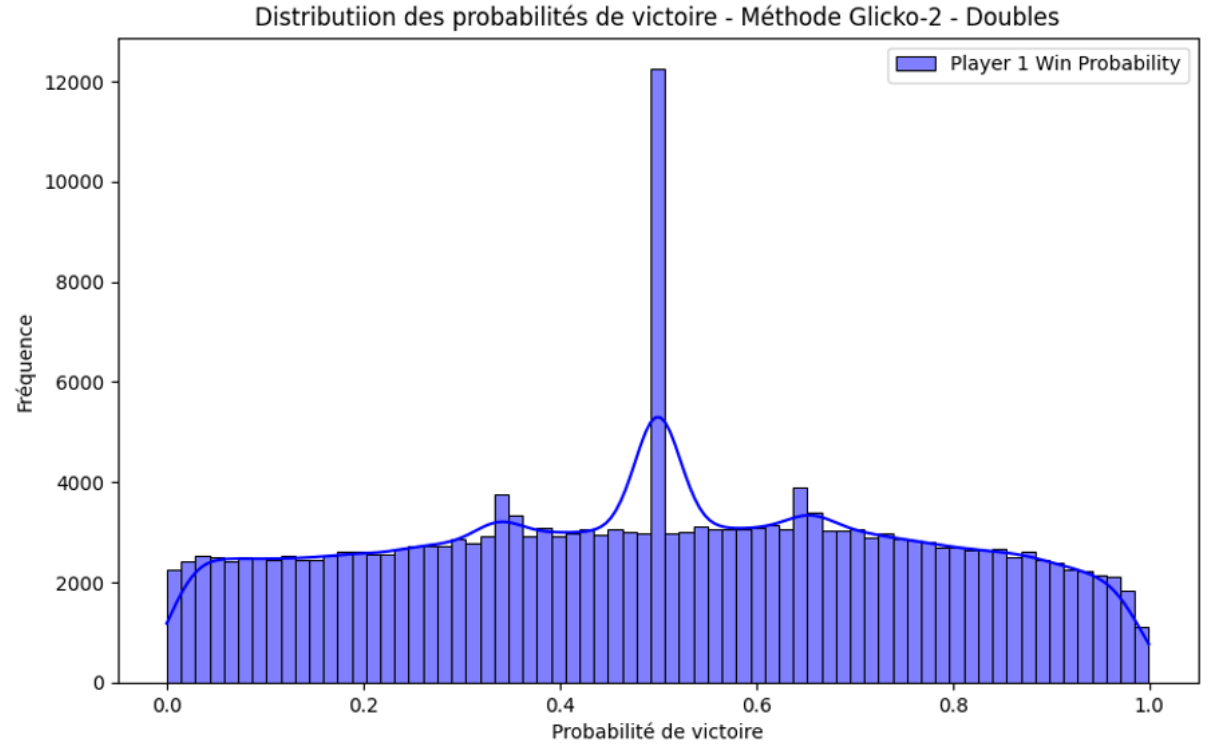


FIGURE 4.2 – Annexe 2

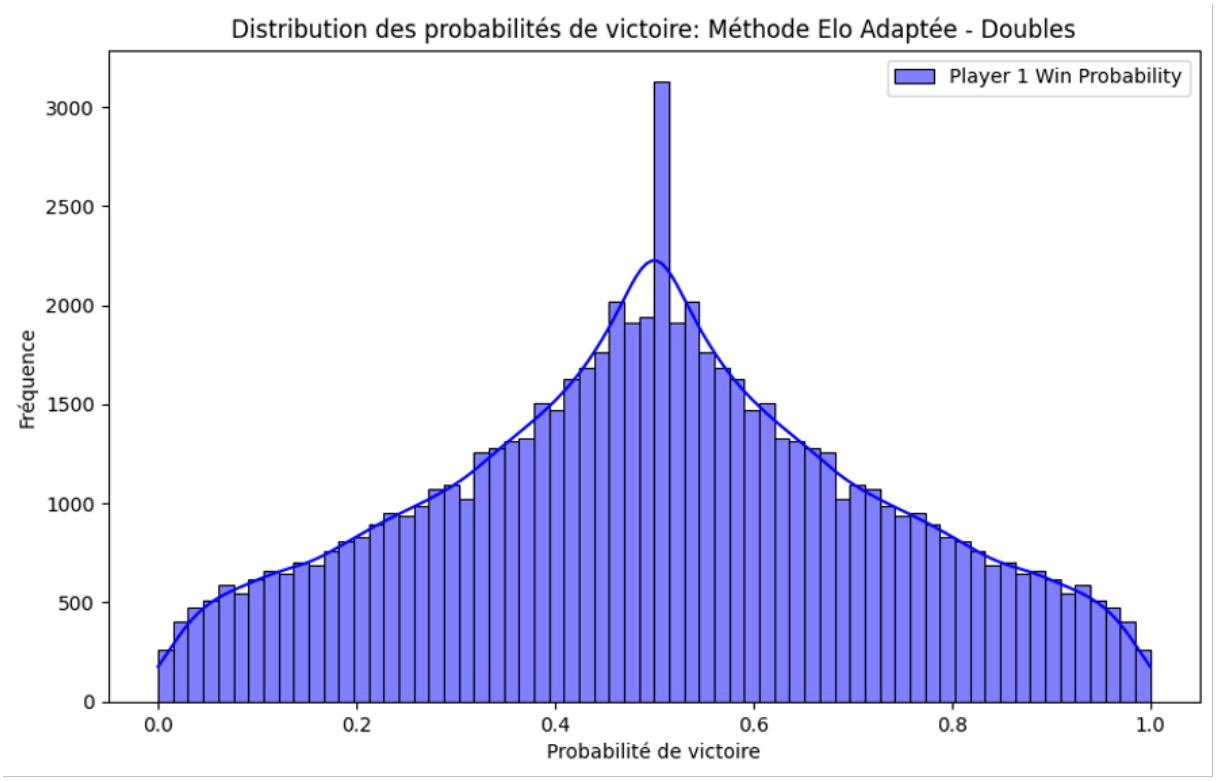


FIGURE 4.3 – Annexe 3

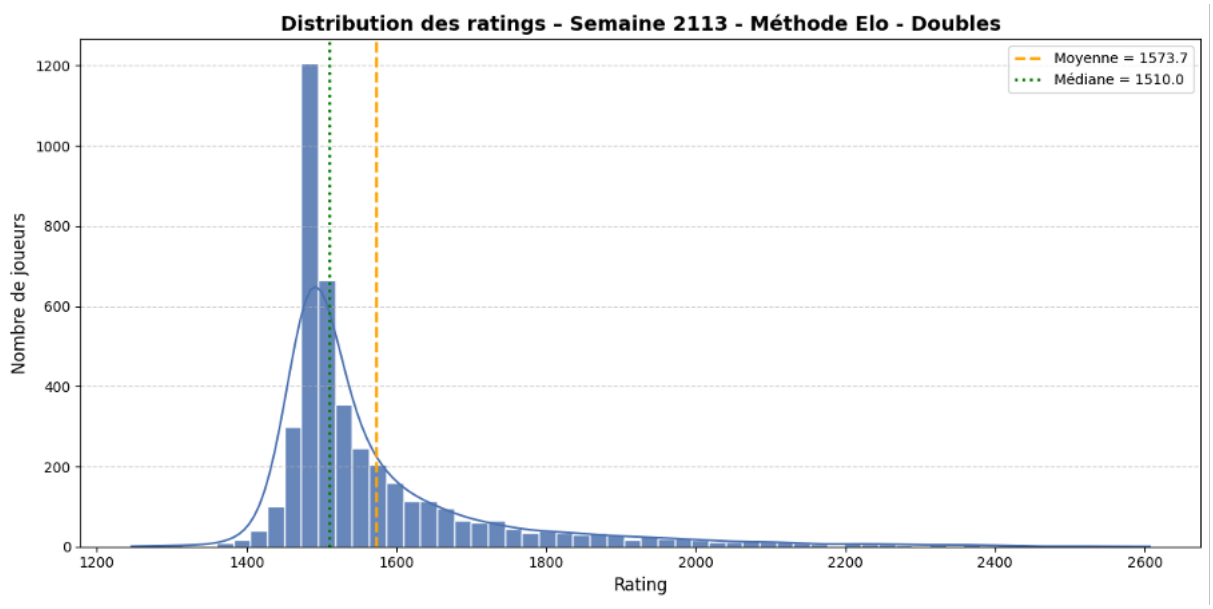


FIGURE 4.4 – Annexe 4

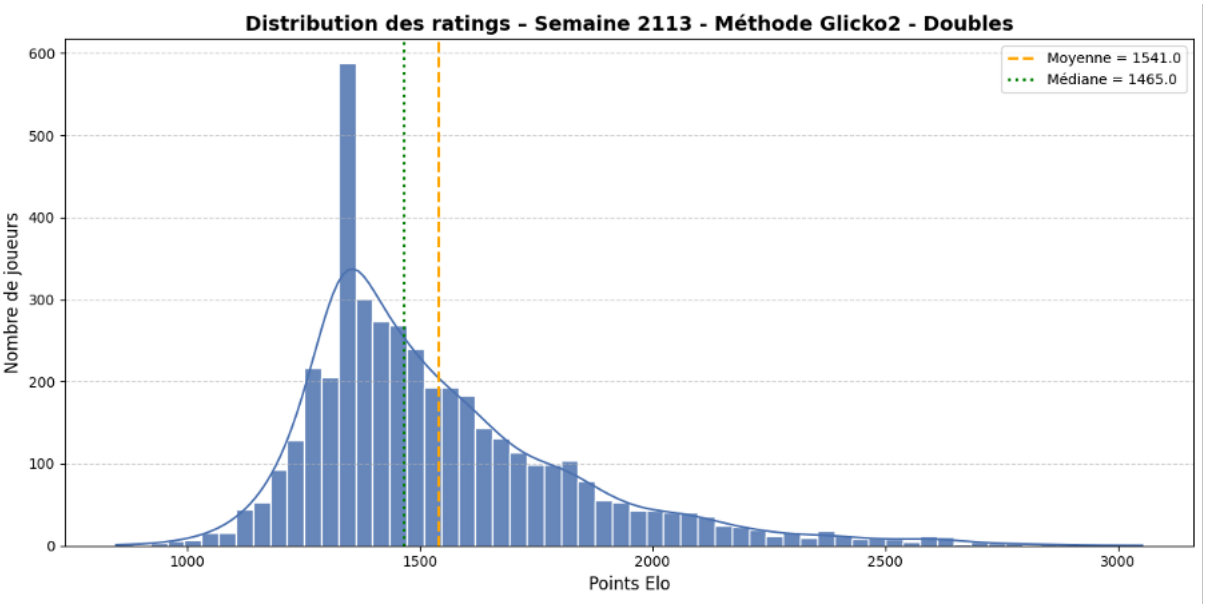


FIGURE 4.5 – Annexe 5

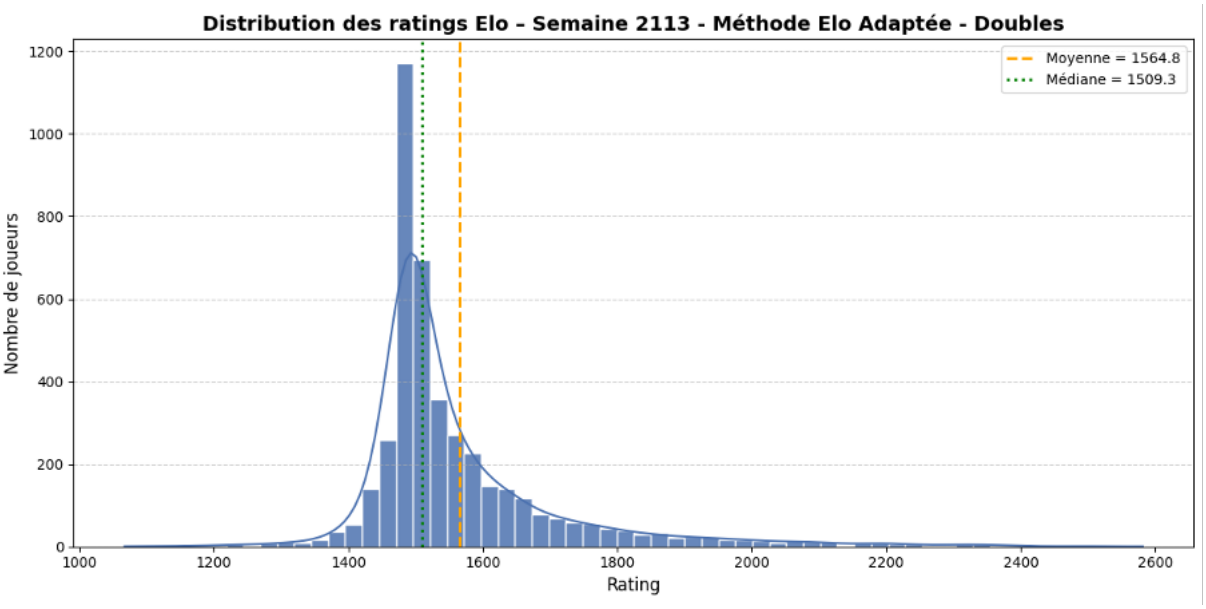


FIGURE 4.6 – Annexe 6

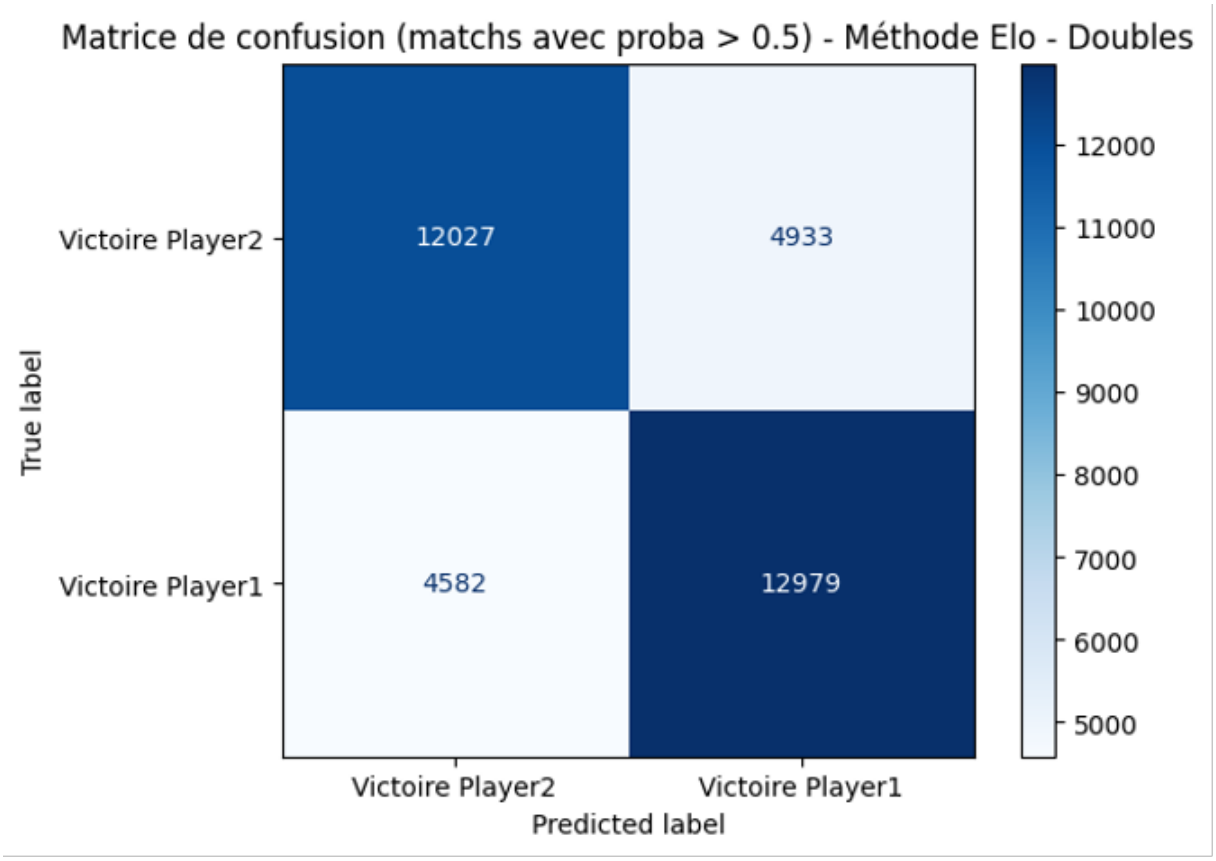


FIGURE 4.7 – Annexe 7

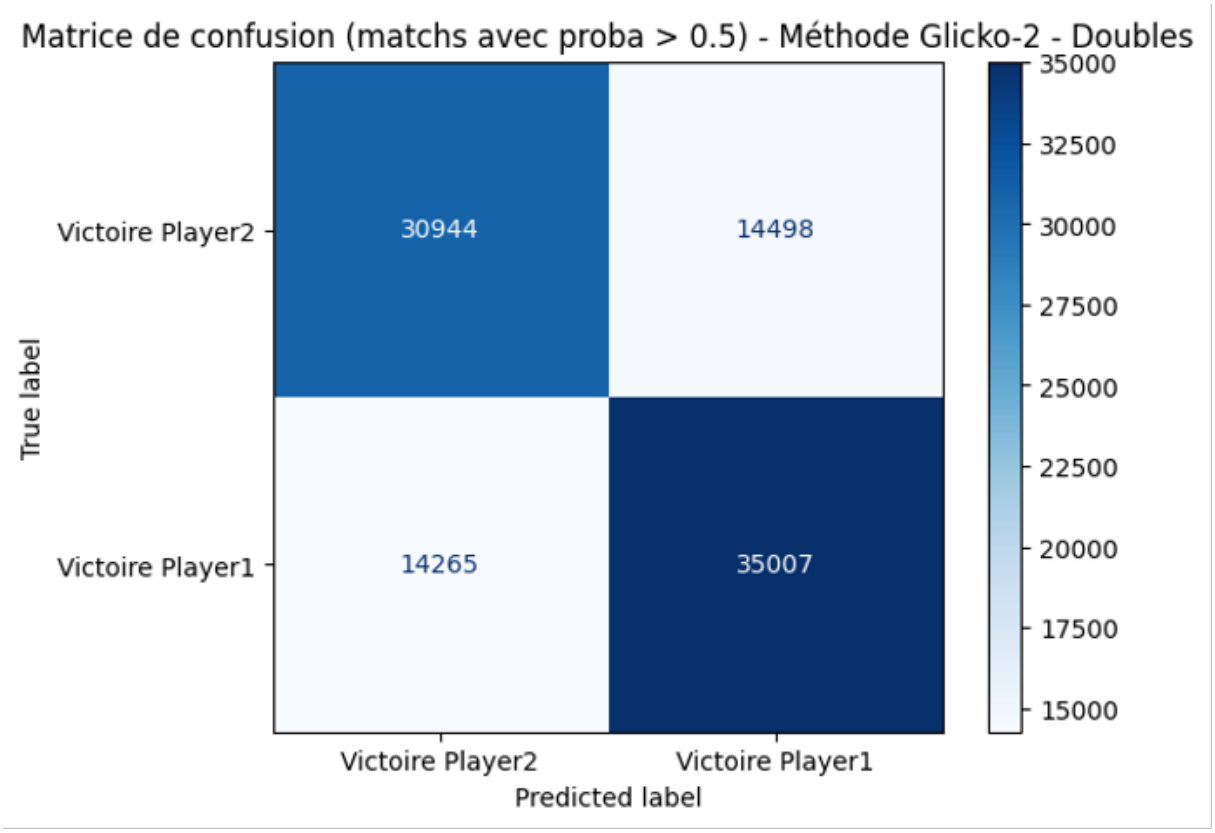


FIGURE 4.8 – Annexe 8

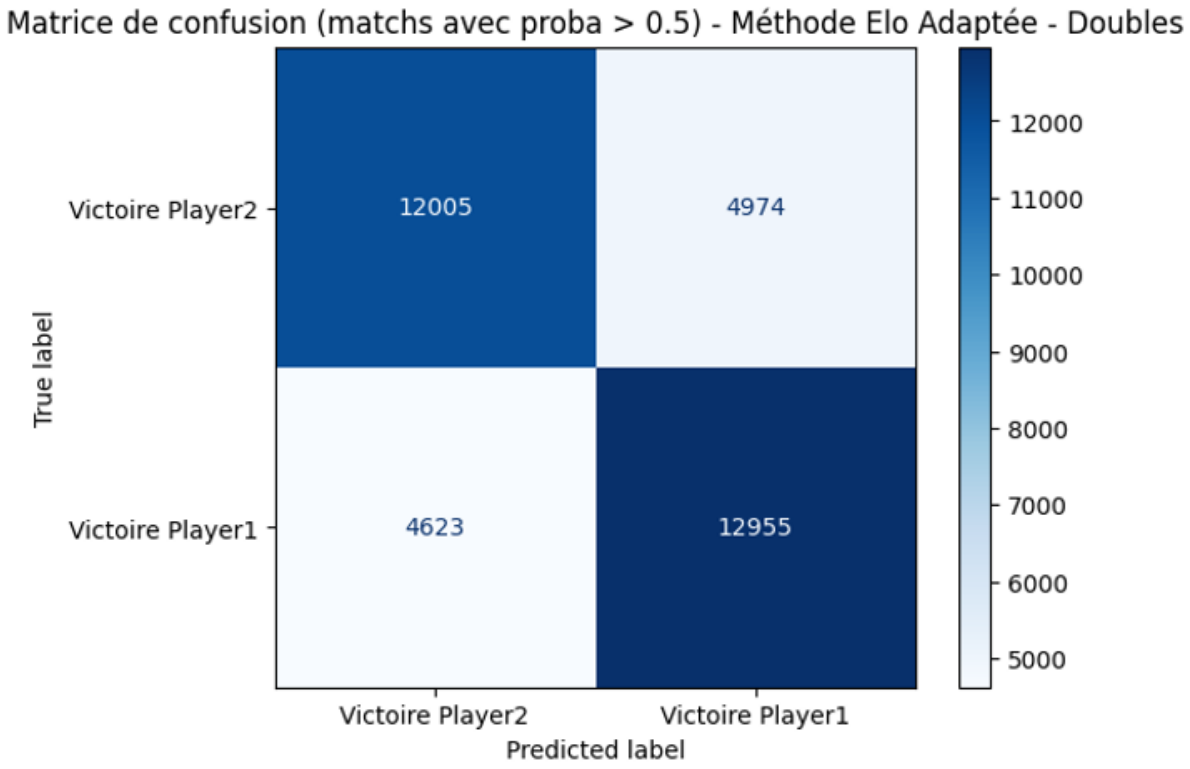


FIGURE 4.9 – Annexe 9

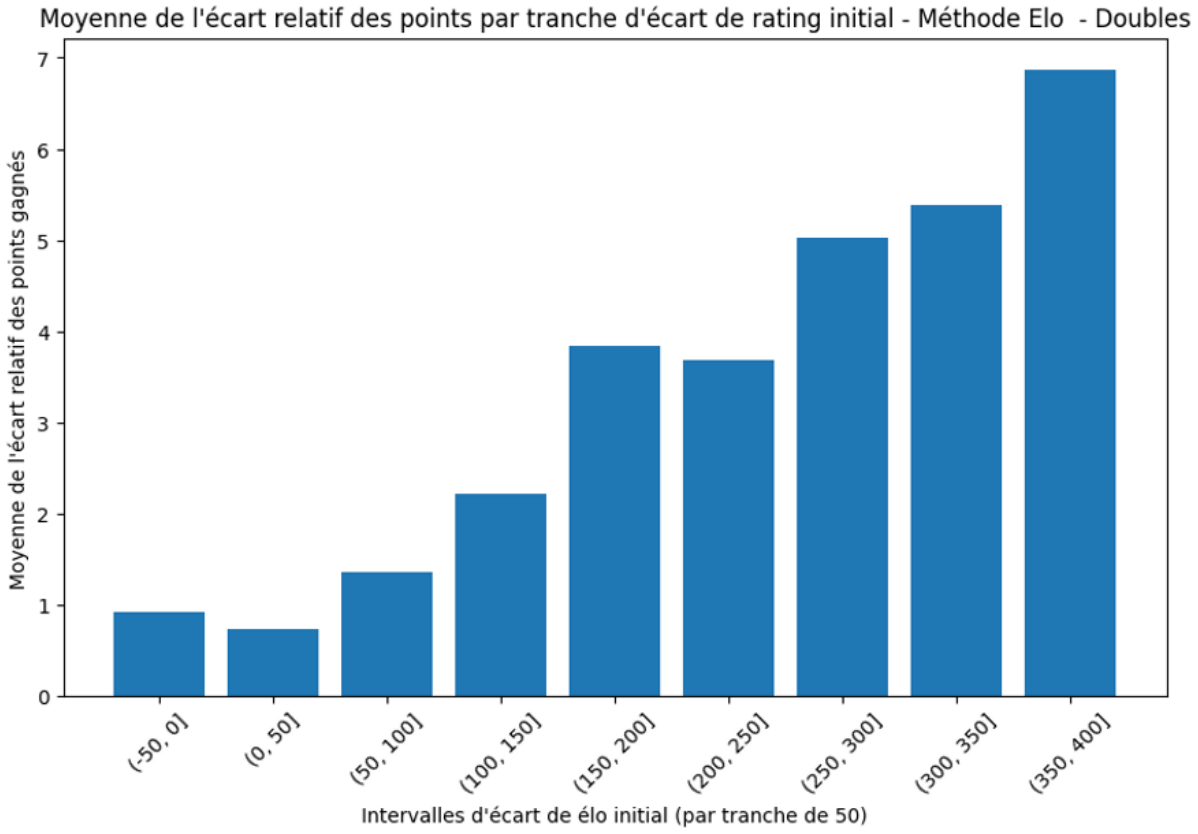


FIGURE 4.10 – Annexe 10

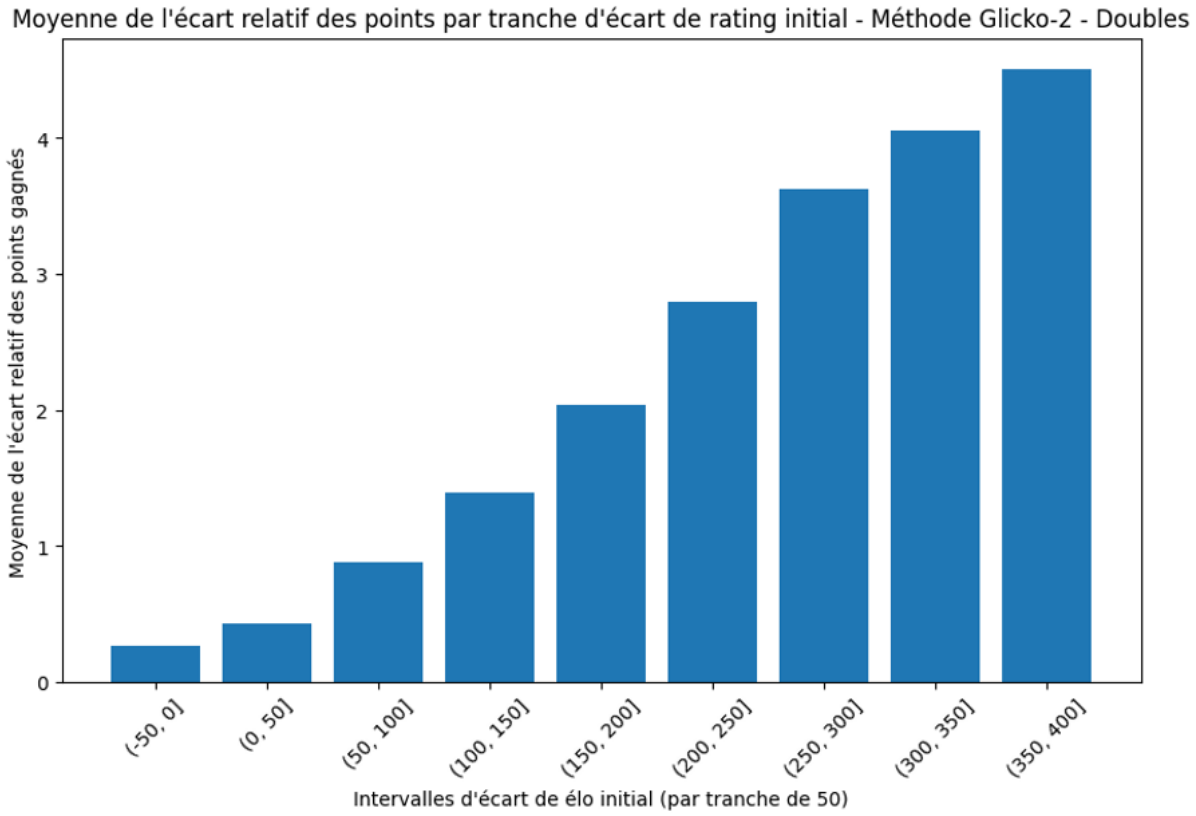


FIGURE 4.11 – Annexe 11

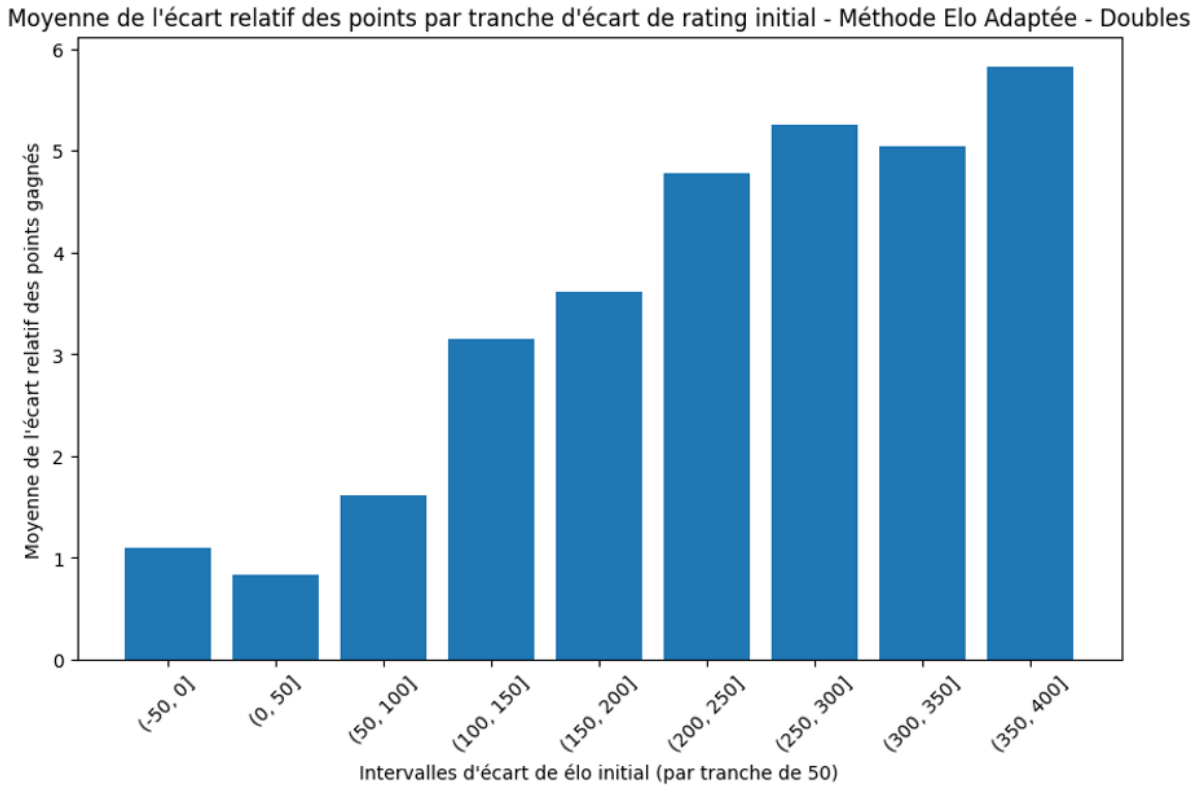


FIGURE 4.12 – Annexe 12

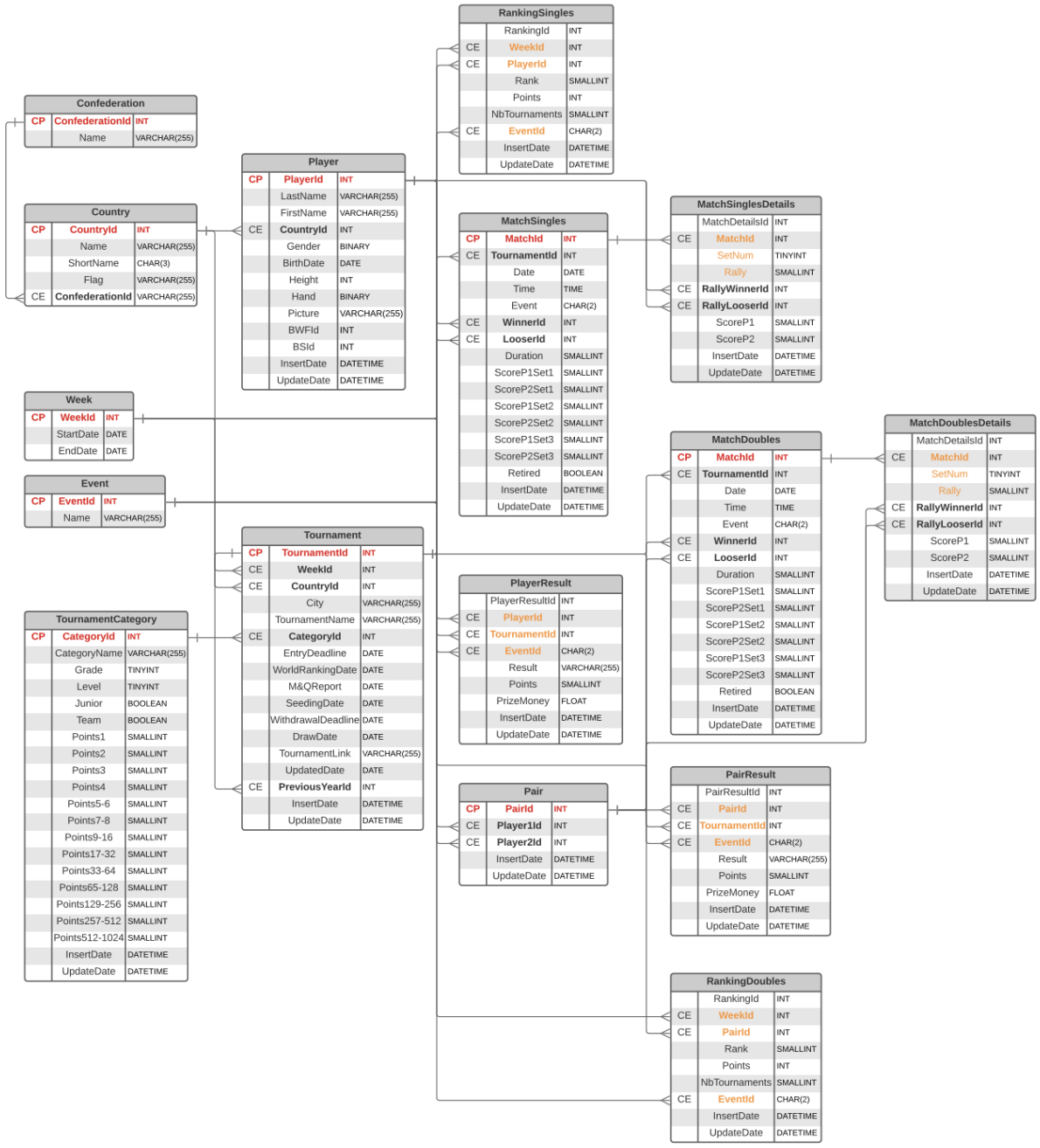


FIGURE 4.13 – Annexe 13 : Schéma de la base de données relationnelle Birdie