EKAP Master Économétrie Appliquée — IAE NANTES ÉCONOMIE & MANAGEMENT

Prévisions des résultats au tennis

Thibaud MEYNIER & Yasmina AMDJHADI

Résumé

This work aims to predict outcoume of tennis matchs, both the winner and the number of sets played with a biprobit model. We tried to predict the winner with two approachs : predict the favorite according to bookmakers odd's, and the best ranked player. According to our datas, this type of modelization is uncorrect, hence we made 2 simple probit regression to predict the outcome. We find interesting results to predict the winner of the match, but we were unable to predict correctly the number of set played with our regressors. Finaly, we have tested somes models out of sample to verify the robustness of our results, and we used it for betting. We obtain a return on investment of 11.3% with the best model.

Table des matières

1 Introduction

L'utilisation des statistiques est devenue de plus en plus populaire dans le sport. Les émissions télévisées nous informent en temps réel sur le pourcentage de possession de balle des équipes de foot, les chronos de références en athlétisme, ou encore le nombre de coup gagnants ou de fautes directes au tennis. Toutes ces statistiques fournissent un aperçu de quelle équipe ou joueur est particulièrement performant dans un match et est donc plus susceptible de gagner. Cependant, une estimation directe de la probabilité qu'un joueur ou une équipe remporte un match est rarement affichée. Depuis quelques années des applications se sont développés afin de permettre aux spectateurs de prédire qui sera le joueur gagnant. Parallèlement, avec l'augmentation de la publicité et du sponsoring des grandes entreprises de paris en ligne, le secteur est en plein essor. D'après le site statista, entre 2012 et 2017, le nombre de compte de parieurs étant actif en France, ont été multipliés par plus de 2,6 [1] .

Avec le développement conjoint des outils statistiques, de la diversité ainsi que de la disponibilité des données, de nouvelles perspectives s'ouvrent. Lisi et Ganella (2017), un modèle de régression logistique et obtiennent 77.2 % de bonnes prédictions sur un échantillon de validation out of sample. Ces derniers test leur modèle en sélectionnant, via des seuils de cote, les paris à prendre et obtiennent un retour sur investissement de 16 %.

L'objet de notre étude est donc de prédire le résultat des matchs de tennis. Plus précisément, nous cherchons à expliquer la probabilité que le joueur favori gagne le match, i.e le joueur ayant la cote la plus basse [2]. De plus nous allons chercher à déterminer le nombre de sets disputés durant le match. Ainsi, nous cherchons à répondre à la question suivante : à partir des cotes disponibles, qu'est-ce qui expliquerait que le favori des bookmakers remporte le match, et en combien de set ? Autrement dit, nous cherchons à prédire de manière plus fine le résultat du match, en prédisant à la fois le vainqueur et le nombre de set. Nous comparerons les résultats obtenus avec l'approche basée sur le classement, où l'on prédit non pas la victoire du favori à partir des cotes, mais la victoire du joueur

1. Statista.com
2. Les bookmakers sont les sites de paris en ligne qui propose les cotes sur les matchs. Dans le cadre du tennis 2 principales cotes sont disponible : la cote du joueur 1 et du joueur 2. Nous définissions le favori comme étant le joueur ayant la plus petite cote. La cote reflète directement la probabilité de victoire du joueur i. Cependant ces probabilités sont tronqués afin d'assurer une marge pour les bookmakers. En ce sens, la somme des probabilités n'est pas égale à 1.

le mieux classé comme ce qui est fait dans la littérature.

Pour ceci, nous utilisons un modèle de régression biprobit, en utilisant des variables explicatives déjà utilisées dans la littérature, mais également des variables disponibles sur différents sites gratuits comme le nombre de matchs disputés au cours du même mois de compétition ou les victoires récentes.

Dans la section suivante, nous décrivons les variables à modéliser ainsi que les variables explicatives. Dans la $3^{ème}$ partie nous faisons les statistiques descriptives de nos données. Dans la $4^{ème}$ section, nous présentons les modèles estimés. Dans la $5^{ème}$ section, nous testons nos modèles sur un échantillion test et nous utilisons 2 modèles pour la prise de paris. Enfin nous concluons et discutons de nos résultats.

2 Analyse des variables

2.1 Les variables dépendantes

Nous utilisons un modèle biprobit afin de prédire le résultat du match entre deux joueurs. Plus précisément, le favori, avec la cote la plus basse noté O_f contre l'outsider avec une cote supérieure noté O_o , ainsi que le nombre de set de ce match. Nous faisons de même entre le joueur le mieux classé noté F_R contre le joueur le moins bien classé noté O_R. Ainsi, nous avons deux variables dépendantes à prédire simultanément. Nous supposons que ces deux phénomènes ne sont pas indépendants l'un de l'autre, ce qui justifie l'utilisation d'un modèle biprobit [3]. Le paramètre ρ permet de rendre compte de l'indépendance ou non entre les deux phénomènes.

2.1.1 *Le vainqueur du match*

Notons Y_1 la première variable dichotomique à modéliser qui prend 2 valeurs :

$$Y_1 = \begin{cases} 1 \ si \ O_f \ ou \ F_R & \text{gagne} \\ 0 \ si \ O_o \ ou \ O_R & \text{gagne} \end{cases}$$

3. Nous n'avons pris que des matchs s'étant disputés au meilleur des 3 sets afin de n'avoir que 4 cas de figures possible.

2.1.2 *Le nombre de sets*

Notons Y_2 la seconde variable dichotomique à prédire :

$$Y_2 = \begin{cases} 1 \ si \ N_{set} = 2 \\ 0 \ si \ N_{set} = 3 \end{cases}$$

Ainsi, 4 cas de figures sont possibles :

$Y_1 = 1$ et $Y_2 = 1$, i.e le joueur favori s'impose facilement,

$Y_1 = 1$ et $Y_2 = 0$, i.e le joueur favori s'impose difficilement,

$Y_1 = 0$ et $Y_2 = 1$, i.e l'outsider s'impose facilement,

$Y_1 = 0$ et $Y_2 = 0$, i.e l'outsider s'impose difficilement.

La modélisation biprobit permet de modéliser les 4 cas de figures. Nous cherchons, à l'aide de variables explicatives à prédire ces 4 possibilités.

2.2 La force des joueurs

Le classement ATP est fonction du nombre de points gagnés par le joueur sur le circuit au cours des 52 dernières semaines. Cette variable prend donc en compte les résultats du joueur sur la dernière année. Magnus et Klaasen (2003) dans leur modèle logit, ont utilisé le classement ATP. Cependant, les auteurs ont transformé cette variable afin de prendre en compte la non linéarité des écarts entre joueurs. Autrement dit, le tennis est un sport dit pyramidale en terme de performance. L'idée sous jacente est que l'écart de niveau entre le numéro 1 mondial et le numéro 2 est plus grande que l'écart entre le 100ème mondial et le 101ème. Toutefois, comme nous regardons l'écart de niveau entre le favori et l'outsider, nous prenons les points ATP et non le classement pour avoir la relation dans le bon sens, i.e plus un joueur est fort, plus il a de points ; ceci afin d'avoir un écart positif entre le favori et l'outsider. En ce sens, nous allons prendre le logarithme des points ATP de chaque joueur, afin de pénalisé plus l'écart de points entre les mieux classés.

Notons $\Delta Lpts$ la variable mesurant l'écart des logarithme des points ATP entre le joueurs favori et l'outsider. Le coefficient associé à cette variable est supposé positif, car plus l'écart de points est grand entre les deux joueurs, plus le favori à de chance de remporter le match facilement.

Une deuxième utilisation possible de cette variable est de former des groupes par tranches de points, via des clusters. C'est ce qu'on fait Lisi et Ganella (2017), obtenant

ainsi 4 catégories résumé dans le tableau 1. Nous reprendrons ces catégories, permettant de comparé laquelle des deux est plus éfficace. Nous appellerons cette variable $\Delta Class$

TABLE 1 – Catégories de points pour la variable D_Class

Intervalle 1	Intervalle 2	Intervalle 3	Intervalle 4	Intervalle 5
0-560	560-920	920-1460	1460-2000	<2000

Source : Lisi et Zanella (2017)

2.3 La forme des joueurs

Le classement ATP, mesure les qualités globales de chaque joueur. Toutefois, il est possible qu'un joueur soit plus en forme que l'autre, i.e qu'un joueur "surf" sur une meilleure dynamique que son adversaire ou celui ci possède plus le rythme de la compétition ce qui favorise ses performances en match. Pour prendre en compte cette dimension, nous regardons 2 indicateurs : l'écart de victoire sur les 5 derniers matchs entre les deux joueurs, ainsi que le nombre de matchs disputés le même mois que le tournoi joué par les deux joueurs. Nous pensons que ces variables peuvent permettre d'expliquer le fait que des joueurs outsiders l'emporte, car la confiance est un élément clé dans le sport et surtout au tennis.

Nous noterons Dmatch la variable mesurant l'écart de matchs disputés sur le mois avant le tournoi comprenant les victoires et les défaites. Ainsi, plus un joueur a disputé de match sur le mois, plus celui ci à de repères quant à ses sensations et son niveau de jeu du moment. A contrario moins celui-ci joue, moins il a de rythme. Toutefois, il est possible que cette variable comporte des effets non linéaire, i.e que passé un certain niveau de match joués, il est possible que l'apport soit plus délétaire que bénéfique, notament à cause de la fatigue physique que cel peut engendrer.

Pour l'écart de victoire sur les 5 derniers matchs nous noterons cette variable Δ_W_L5. Le coefficient associé aux deux variables est supposé positif pour expliquer la victoire finale et le nombre de sets.

2.4 L'information des bookmakers

Les bookmakers faisant les cotes des matchs ont plus d'informations que les parieurs sur les données sous-jacentes aux matchs. C'est en ce sens que Lisi et Zanella (2017) ont créé

une variable dichotomique prenant en compte l'information délivrée par les bookmakers via les cotes. Nous notons cette variable Odd_S prenant pour valeur 1 si le joueur le mieux classé des deux a une cote supérieur à 1.9. Ainsi, selon les books, le joueur le mieux classé n'est pas en position favorable pour gagner le match. Nous estimons que la relation entre cette variable et la probabilité de victoire du joueur favori est négative. En ce qui concerne son impact sur la probabilité que le match se termine en 3 sets, nous pensons que la relation est positive.

3 Statistiques descriptives

La base de données à été construite à partir d'une base secondaire téléchargée sur tennis-data.co.uk, a laquelle nous avons rajouté des données trouvées sur flashscore.com et tennisexplorer.com. La base brute contient 192 observations regroupant 6 tournois ATP disputés entre le 6 septembe et le 12 Octobre 2020 [4].

3.1 Les variables dépendantes

TABLE 2 – Modalité des variables dépendantes de la base favori

	3 sets	2 sets	Total
Outsider Win	28	41	69
Favori Win	39	84	123
Total	67	125	192

TABLE 3 – Modalité des variables dépendantes de la base rank corrigée des outliers

	3 sets	2 sets	Total
Outsider Win	33	52	85
Favori Win	31	65	96
Total	64	117	181

4. Nous avons pris uniquement des matchs du circuit ATP qui est le circuit principal et "mainstream" dans le sens ou ce sont ces tournois qui sont diffusés à la télévisions. Ce sont également les tournois avec le plus de gains. Pour des raisons de réglementation en France, on ne peut parier que sur ces tournois. Le circuit challenger, que l'on peut comparé à la deuxième division au foot est parfois sujet à des trucages de match, en ce sens nous ne retenons que le circuit principal pour eviter les biais sur nos résultats

3.2 Les variables explicatives

3.2.1 *Base Favori*

FIGURE 1 – Boxplot des variables quantitatives

D'après la figure 1, 3 variables quantitatives peuvent comporter des points atypiques dans leur distribution. Toutefois, d'après le test de Rosner, ces points ne sont pas jugés atypiques, en ce sens aucunes observations n'est retirées de cette base (voir annexe 3).

TABLE 4 – T.test entre la variable Odd S et les variables quantitatives

	Mean Gp = 0	Mean Gp = 1	P-value
Dmatch	-0,22	1,476	0***
DLpts	0,9	-0,6	0***
D_Class	1,57	-1,07	0***
D_W_L5	0,032	0,18	0***

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE 5 – Statistiques descriptives des variables explicatives

	Moyenne	Ecart-type	Médiane	Min	Max	Q1	Q3
Dmatch	0.1458	2.31	0	-6	7	-1	1
DLpts	0.578	0.883	0.556	-2.439	2.848	0.097	1.108
D_Class	0.995	1.56	1	-3	4	0	2
D_W_L5	0.0645	0.255	0	-0.6	0.8	0	0.2
Odd_S	$152 = 0\,;\, 40 = 1$	-	-	-	-	-	-

TABLE 6 – Matrice des corrélations entre les variables quantitatives

	D_Class	Dmatch	D_W_L5	DLpts
D_Class	1			
Dmatch	-0.37	1		
D_W_L5	-0.287	0.49	1	
DLpts	0.837	-0.386	-0.253	1

D'après le tableau 4, la variable explicative Odd_S est corrélée avec nos variables quantitatives. Nous devrons faire attention au VIF des modèles que nous estimerons.

D'après le tableau 6, nos variables explicatives ne sont pas très corrélées entre elles sauf pour D_Class et DLpts, qui représentent la même information (L'écart de niveau entre les joueurs, induit par le classement ATP), présentée différement. En ce sens, le risque de multicolinéarité dans les modèles est faible.

3.2.2 *Base Rank*

FIGURE 2 – Box plot des variables quantitatives

D'après la figure 2, nos variables explicatives de la base rank semblent présenter des point atypiques dans leur distribution. Cette fois ci, La variables D_Class comporte comme valeur abérante 4 (voir tableau 7). En ce sens nous la retirons de la base. Nous perdons ainsi 11 observations (voir annexe 4).

TABLE 7 – T.test entre Odd_S et les variables quantitatives

	Mean Gp =0	Mean gp = 1	p-value
Dmatch	-0.177	-1.3705	0.0069***
DLpts	0.9	0.579	0.00***
D_Class	1.39	1.02	0.0285**
D_W_L5	0.034	-0.175	0.00***

D'après le tableau 7, comme pour la base favori, la variable Odd_S est corrélée aux autres variables explicatives.

TABLE 8 – Statistiques descriptives des variables quantitatives

		Moyenne	Mediane	Ecart-type	Min	Max	Q1	Q3
Base brute	Dmatch	-0,5	0	2,26	-7	5	-2	1
	DLpts	0,839	0,684	0,639	0,004	2,84	0,348	1,236
	D_Class	1,464	1	1,129	0	4***	1	2
	D_W_L5	-0,01458	0	0,263	-0,8	0,8	-0,2	0,2
	Odd_S	152 =0 ; 40= 1	-	-	-	-	-	-
Base sans outliers	Dmatch	-0,442	0	2,25	-7	5	-2	1
	DLpts	0,778	0,639	0,597	0,004	2,81	0,33	1,09
	D_Class	1,309	1	0,968	0	3	1	2
	D_W_L5	-0,012	0	0,262	-0,8	0,8	-0,2	0,2
	Odd_S	141 =0 ; 40= 1	-	-	-	-	-	-

$^{***}p < 0.001$, p-value du test de Grubbs

On note d'après le tableau 8 que le retrait de notre valeur atypique de la variable D_Class ne modifie que très peu les distributions de nos différentes variables.

TABLE 9 – Matrice des corrélation entre les variables quantitatives

	D_Class	Dmatch	D_W_L5	DLpts
D_Class	1			
Dmatch	-0.148	1		
D_W_L5	-0.094	0.496	1	
DLpts	0.63	-0.1522	-0.024	1

Enfin, d'après le tableau 9, les variables quantitatives ne semblent pas fortement corrélées entre elles ce qui est un bon point pour la stabilité des modèles estimés ultérieurement.

4 Modélisation

4.1 Modèle Biprobit Favori

Le tableau 10 résume les différents modèles biprobit effectués sur la base Favori. Nous constatons que chacun des modèles estimé a un ρ non significatif. En ce sens, d'après nos différents modèles, l'évènement le favori gagne et le nombre de set du matchs ne sont pas liés entre eux. Par conséquent nous devrons estimés deux modèles probit séparément. On note également que pour l'équation estimant la victoire du favori, toutes nos variables sont significatives exepté la différence de victoire sur les 5 derniers match (D_W_L5). En ce qui concerne le nombre de sets, aucune variable est significative à part la constante.

Bien que le Modèle 2 et le Modèle 3 possède un ρ avec une p-value proche de 10%, le modèle nous indique que seul la constante modélise le nombre de set, et non nos variables explicatives. Deux perspectives s'offrent à nous : ou bien le nombre de set d'un match de tennis est aléatoire et donc imprévisible ou alors nous n'avons pas mis les bonnes variables dans le modèle. En changeant notre Y_1, peut être que les résultats différeront. Pour rappel, la deuxième série de données consiste à prédire la victoire du joueur le mieux classé.

Table 10 – Tableau récapitulatif des modèles biprobit sur la base favori

	Model 1		Model 2		Model 3		Model 4	
Variables	FavW	sets	FavW	sets	FavW	sets	FavW	sets
Dmatch	0.161***	-0.0482	0.155***	-0.0606	0.158***	-0.0539	0.157***	-0.0541
	(0.0521)	(0.0495)	(0.0517)	(0.0489)	(0.0467)	(0.0406)	(0.0467)	(0.0407)
D_W_L5	-0.0643	-0.148	0.0329	-0.161				
	(0.424)	(0.431)	(0.428)	(0.431)				
DLpts	0.468***	0.177					0.451***	
	(0.165)	(0.154)					(0.166)	
Odd_S	0.722**	0.510	0.766**	0.239	0.740**		0.672**	
	(0.332)	(0.323)	(0.332)	(0.323)	(0.331)		(0.333)	
D_Class			0.286***	-0.0101	0.286***			
			(0.0943)	(0.0876)	(0.0933)			
Constant	-0.0537	0.198	-0.0809	0.370**	-0.0733	0.399***	-0.0374	0.399***
	(0.176)	(0.173)	(0.175)	(0.174)	(0.171)	(0.0938)	(0.175)	(0.0938)
ρ (p-value)	0.22		0.1286		0.1253		0.2091	
Observations	192		192		192		192	

Significativité : p-value < 0.01***, p-value < 0.05**, p-value < 0.01*

4.2 Modèle Biprobit Best Rank

Dans le tableau 11 on peut observer les résultats des différents modèles biprobit effectués sur la base Rank. Nous constatons que dans chaque modèle estimé, aucun n'a un ρ significatif. De ce fait, nous pouvons supposer que nos deux variables expliquées, à savoir l'événement le favori gagne et le nombre de sets du match, ne sont pas liées entre elles. Cela signifie que le nombre de sets du match n'influence pas le fait que le favori gagne. Par conséquent nous devons estimer ces différents phénomènes par deux probits séparés. Dans le modèle 1, le nombre de variables qui sont significatives nous fait douter de la qualité du modèle. En revanche dans les 3 autres modèles, nous constatons que pour l'équation estimant la victoire du favori, toutes nos variables sont significatives sauf la constante. Cependant en ce qui concerne l'équation estimant le nombre de sets aucune variable n'est significative à l'exception de la constante. C'est donc la constante qui modélise le nombre de set et non nos variables explicatives. Dans le modèle 2 et 3 la p-value de ρ est proche de 10%, mais cela n'est pas suffisant. On conclut donc que, comme dans le cas de la base Favori : soit le nombre de set d'un match de tennis est aléatoire et donc il n'est pas modélisable, soit nous n'avons pas les variables adéquates permettant de modéliser ce phénomène.

Table 11 – Tableau récapitulatif des modèles biprobit sur la base Best Rank

	Modèle 1		Modèle 2		Modèle 3		Modèle 4	
Variables	FavW	sets	FavW	sets	FavW	sets	FavW	sets
D_Class	0.130	-0.135			0.254**	-0.0696	0.255**	-0.0600
	(0.134)	(0.134)			(0.107)	(0.103)	(0.108)	(0.103)
Odd_S	-0.615**	0.126	-0.630**	0.0604	-0.663***	0.0116	-0.655**	0.104
	(0.260)	(0.257)	(0.249)	(0.243)	(0.250)	(0.243)	(0.259)	(0.256)
Dmatch	0.156***	-0.0743	0.151***	-0.0403	0.145***	-0.0488	0.143***	-0.0786
	(0.0533)	(0.0516)	(0.0473)	(0.0449)	(0.0471)	(0.0447)	(0.0524)	(0.0511)
D_W_L5	0.00660	0.559					0.0487	0.569
	(0.442)	(0.453)					(0.440)	(0.452)
DLpts	0.351	0.194	0.480***	0.0518				
	(0.223)	(0.215)	(0.182)	(0.164)				
Constant	-0.160	0.357*	-0.0889	0.306*	-0.0472	0.445**	-0.0505	0.410**
	(0.195)	(0.192)	(0.178)	(0.173)	(0.178)	(0.180)	(0.181)	(0.183)
ρ (p-value)	0.1661		0.2104		0.1371		0.1348	
Observations	181	181	181	181	181	181	181	181

Significativité : p-value < 0.01***, p-value < 0.05**, p-value < 0.01*

4.3 Modèle probit séparés base Favori

4.3.1 *Modélisation du vainqueur*

Le tableau 12 présente les différents modèles probit sur la variable dépendante Favori Win. Comme les variables explicatives DLpts et D_Class sont colinéaires, elles n'ont pas été introduites en même temps dans les modèles. Pour chaque modèle on constate que la différence de victoire sur les 5 derniers match n'est pas significative (D_W_L5). En revanche les autres variables sont significatives et les coefficients sont conformes à nos attentes.

Ces modèles sont supposés homoscédastiques. Nous avons vérifié cette hypothèse pour les modèles 2 et 4 (meilleures au sens des prédictions in sample présentées dans le tableau 14). Le tableau 13 résume les p-value des variables pouvant expliquer la présence d'hétéroscédasticité dans les modèles. Dans le modèle 2, il n'y a pas d'hétéroscédasticité à priori (la variable Dmatch était suspecte, mais mise seule, elle n'est plus significative). Pour le modèle 4 en revanche il semble que 2 variables soient responsables de l'hétéroscédasticité des erreurs. Toutefois, à cause de la multicolinéairité le modèle ne converge pas quand on inclut toutes les variables explicatives comme "coupable" de l'hétéroscédasticité, de même le VIF des modèles estimés en prenant en compte cette hypothèse sur les erreurs sont élevés (voir Annexe 1 et 2). Enfin, la ligne Lrtest présente les tests statistiques mesurant la différence entre le modèle 4 supposé homoscédastique et les 2 modèles hétéroscédastiques ; avec les variables D_class et Dmatch comme explicatives de l'hétéroscédasticité. Nous ne trouvons aucunes différences entre ces modèles et les modèles supposés homoscédastiques. En ce sens, les modèle 2 et 4 peuvent être utilisé comme tel.

Le tableau 14 résume les performances de prédictions des différents modèles estimés. On note que ces modèles sont très sensibles, i.e qu'ils prédisent très bien Y=1, ici le favori des bookmakers gagne le match. En revanche, ils prédisent beaucoup moins bien Y=0, i.e la victorie de l'outsider. D'après ces données, le modèle 4 est le plus performant avec le meilleur taux d'erreur de 35,08%.

TABLE 12 – Tableau récapitulatif des modèles probit sur le vainqueur du match

	Modele 1	Modele 2	Modele 3	Modele 4
(Intercept)	−0.06	−0.06	−0.08	−0.08
	(0.18)	(0.17)	(0.17)	(0.17)
Dmatch	0.16**	0.16***	0.15**	0.16**
	(0.05)	(0.05)	(0.05)	(0.05)
DLpts	0.47**	0.47**		
	(0.16)	(0.16)		
D_W_L5	−0.06		0.04	
	(0.43)		(0.43)	
Odd_S1	0.73*	0.73*	0.77*	0.77*
	(0.33)	(0.33)	(0.33)	(0.33)
D_Class			0.28**	0.28**
			(0.09)	(0.09)
AIC	243.20	241.22	241.98	239.99
BIC	259.46	254.23	258.24	253.00
Log Likelihood	-116.60	-116.61	-115.99	-116.00
Deviance	233.20	233.22	231.98	231.99
Num. obs.	191	191	191	191

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE 13 – Modèles hétéroscédastiques

Variables	Modele 2		Modèle 4		
DLpts	0.259				
Dmatch	0.002***	0.3	0.0174*		
Odd_S	0.77			0.106	
D_Class					0.0133*
Lrtest			0.2625	0.103	

TABLE 14 – Prédictions in sample des modèles probit estimés

	Modèle 1	Modèle 2	Modèle 3	Modèle 4
Sensitivité (Y=1)	86.89%	87.7%	86.07%	86.07%
Spécificité (Y=0)	21.74%	20.29%	26.09%	27.54%
Taux d'erreur	36.65%	36.65%	35.6%	35.08%
Nbr obs	191	191	191	191

4.3.2 *Modélisation du Nombre de set*

Le tableau 15 résume les différents modèles fait sur la variable dépendante Nombre de sets. On constate qu'aucunes variables explicatives n'est significatives sauf la constante quand on prend la variable D_class à la place de DLpts. Toutefois, en verifiant l'hypothèse

d'homoscédasticité, certains modèles hétéroscédastiques sont statistiquement meilleurs que les modèles homoscédastiques (p-value [5] Lrtest inférieur à 0.05). Le tableau 16 en présente 4. Malgré un meilleur Log likelihood de ces modèles, on constate que ces modèles sont très sensibles (on prédit très bien Y = 1, mais ils ne prédisent pas Y=0). En ce sens, nos variables n'expliquent pas le nombre de set d'après le tableau 17.

TABLE 15 – Tableau récapitulatif des modèles probit sur le Nombre de set

	Modele 1	Modele 2	Modele 3	Modele 4
(Intercept)	0.20	0.17	0.37*	0.36***
	(0.17)	(0.17)	(0.17)	(0.11)
D_W_L5	−0.17		−0.18	
	(0.43)		(0.43)	
DLpts	0.19	0.23		
	(0.16)	(0.15)		
Odd_S1	0.50	0.47	0.24	0.23
	(0.33)	(0.32)	(0.32)	(0.24)
Dmatch	−0.04		−0.05	−0.06
	(0.05)		(0.05)	(0.04)
D_Class			−0.00	
			(0.09)	
AIC	252.26	249.60	253.71	249.89
BIC	268.52	259.36	269.98	259.65
Log Likelihood	-121.13	-121.80	-121.86	-121.94
Deviance	242.26	243.60	243.71	243.89
Num. obs.	191	191	191	191

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE 16 – Modèles hétéroscédatsiques pour la variable nombre de set

variables	Modèle 1		Modèle 3	
DLpts	0.0501*	0.00286***		
Dmatch	0.00023***	0.00186***	0.2358	
Odd_S	0.86			
D_Class			0.439	0.4195
D_W_L5	0.45		0.0089***	0.0015***
lrtest		0.0345*		0.017*

5. Les p-value du lrtest correspond au test entre le modèle supposé homoscédastique et le modèle prenant en compte l'hétéroscédasticité des erreurs.

TABLE 17 – Predictions in sample des modèles estimés

	Modèle 1	Modèle 2	Modèle 3	Modèle 4
Sensitivité (Y=1)	98.4%	100%	98.4%	100%
Spécificité (Y=0)	1.515%	0%	0%	0%
Taux d'erreur	35.08%	34.55%	35.6%	34.55%
Nbr obs	191	191	191	191

4.4 Modèle Probit séparés Best Rank

Dans le tableau 18, les résultats des probit séparés sur la base Best Rank sont représentés. Nous ne retenons pas le modèle 1 car les variables DLpts et D_W_L5 sont corrélées. Cela peut expliquer le fait qu'elles ne soient pas significatives dans ce modèle. Le modèle ayant le taux de prédiction le plus élevé est le modèle 4. Les R^2 d'un modèle à l'autre varient très peu. Nous allons donc interpréter plus en détail les probit séparés du modèle 4.

TABLE 18 – Tableau récapitulatif des modèles probit séparés sur la base Best Rank

	Probit 1		Probit 2		Probit 3		Probit 4	
Variables	FavW	sets	FavW	sets	FavW	sets	FavW	sets
D_Class	0.126	-0.133			0.250**	-0.0668	0.251**	-0.0579
	(0.130)	(0.127)			(0.104)	(0.101)	(0.103)	(0.102)
Odd_S	-0.615**	0.127	-0.629**	0.0619	-0.660***	0.0149	-0.654***	0.106
	(0.257)	(0.249)	(0.250)	(0.246)	(0.245)	(0.242)	(0.253)	(0.246)
Dmatch	0.156***	-0.0743	0.150***	-0.0409	0.145***	-0.0492	0.143***	-0.0784
	(0.0497)	(0.0500)	(0.0457)	(0.0441)	(0.0462)	(0.0440)	(0.0508)	(0.0496)
D_W_L5	-0.00604	0.554					0.0380	0.563
	(0.457)	(0.430)					(0.457)	(0.424)
DLpts	0.355*	0.192	0.479***	0.0541				
	(0.209)	(0.221)	(0.169)	(0.176)				
Constant	-0.158	0.354*	-0.0889	0.304*	-0.0441	0.441**	-0.0471	0.406**
	(0.188)	(0.194)	(0.175)	(0.181)	(0.180)	(0.178)	(0.178)	(0.180)
R2	0.1162	0.0169	0.1126	0.0054	0.1058	0.0068	0.1058	0.0135
Correctly classified	63.54%	65.19%	62.43%	64.64%	62.98%	64.64%	62.98%	65.19%
Observations	181	181	181	181	181	181	181	181

Significativité : p-value 0.01*** , p-value 0.05** , p-value 0.1*

4.4.1 *Modélisation du vainqueur*

Voici dans la Figure 3 les effets marginaux du probit 4, modélisant le vainqueur du match. Les effets marginaux des variables Odd_S, Dmatch, D_Class, D_W_L5 sont respectivement de -0.254 , 0.056 , 0.099 et 0.0151. Par conséquent, le fait que le joueur le mieux classé ai une cote supérieur à 1.90, fait diminuer sa probabilité de victoire de -0.2549. En outre, si le joueur le mieux classé a joué un match de plus que son adversaire, alors sa probabilité de remporté le match augmente de 0.056.

FIGURE 3 – Tableau récapitulatif des effets marginaux du probit 4, avec la variable FavW

```
. mfx

Marginal effects after probit
      y  = Pr(FavW) (predict)
         =  .52930247
```

variable	dy/dx	Std. Err.	z	P>|z|	[95% C.I.]		X
Odd_S*	-.2549291	.09335	-2.73	0.006	-.43789	-.071968	.220994
Dmatch	.0568085	.02023	2.81	0.005	.017158	.096459	-.441989
D_Class	.0998813	.04094	2.44	0.015	.019636	.180126	1.30939
D_W_L5	.0151126	.18166	0.08	0.934	-.340927	.371152	-.012155

(*) dy/dx is for discrete change of dummy variable from 0 to 1

Dans 62,98% des cas, le modèle prédit correctement le fait que le favori gagne. La sensibilité est de 72,92%. C'est à dire (pour les cas où FavW =1), le modèle prédit correctement le fait que le favori gagne dans 72,92% des cas. La spécificité est de 51,76% (pour les cas où FavW=0).

FIGURE 4 – Tableau récapitulatif des prédictions du probit 4 sur la variable FavW

```
. estat classification

Probit model for FavW

                 True
Classified |         D          ~D |      Total
-----------+--------------------------+-----------
     +     |        70          41 |        111
     -     |        26          44 |         70
-----------+--------------------------+-----------
   Total   |        96          85 |        181

Classified + if predicted Pr(D) >= .5
True D defined as FavW != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   72.92%
Specificity                     Pr( -|~D)   51.76%
Positive predictive value       Pr( D| +)   63.06%
Negative predictive value       Pr(~D| -)   62.86%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   48.24%
False - rate for true D         Pr( -| D)   27.08%
False + rate for classified +   Pr(~D| +)   36.94%
False - rate for classified -   Pr( D| -)   37.14%
--------------------------------------------------
Correctly classified                        62.98%
--------------------------------------------------
```

4.4.2 *Modélisation du Nombre de set*

Les effets marginaux du probit 4, avec la variable sets sont représentés ci-dessous. Comme la qualité de prédicition du modèle n'est pas bonne, nous n'allons pas interpréter les valeurs suivantes.

FIGURE 5 – Tableau récapitulatif des effets marginaux du probit 4, avec la variable sets

```
. mfx

Marginal effects after probit
      y  = Pr(sets) (predict)
         =  .64874805
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.      z    P>|z|  [    95% C.I.   ]       X
---------+--------------------------------------------------------------------
   Odd_S*|    .0388511      .08898     0.44   0.662  -.135546  .213248   .220994
   Dmatch|   -.0290749      .01837    -1.58   0.113  -.065081  .006931  -.441989
  D_Class|   -.0214657      .03785    -0.57   0.571  -.095659  .052727   1.30939
    D_W_L5|    .2087715      .15729     1.33   0.184  -.099503  .517046  -.012155
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Dans le tableau suivant, nous avons les prédictions du probit 4 sur la variable FavW. Dans 65,19% des cas, le modèle prédit correctement le fait que le favori gagne. La sensibilité est de 100% (pour les cas où FavW =1) et la spécificité est de 1.58% (pour les cas où FavW=0). Ces valeurs extrêmes confirme que la qualité de prédiction du modèle n'est pas très bonne.

FIGURE 6 – Tableau récapitulatif des prédictions du probit 4 sur la variable sets

```
. estat classification

Probit model for sets

                 True
Classified |         D          ~D |      Total
-----------+----------------------+-----------
     +     |       117          63 |        180
     -     |         0           1 |          1
-----------+----------------------+-----------
   Total   |       117          64 |        181

Classified + if predicted Pr(D) >= .5
True D defined as sets != 0
-------------------------------------------------
Sensitivity                     Pr( +| D)  100.00%
Specificity                     Pr( -|~D)    1.56%
Positive predictive value       Pr( D| +)   65.00%
Negative predictive value       Pr(~D| -)  100.00%
-------------------------------------------------
False + rate for true ~D        Pr( +|~D)   98.44%
False - rate for true D         Pr( -| D)    0.00%
False + rate for classified +   Pr(~D| +)   35.00%
False - rate for classified -   Pr( D| -)    0.00%
-------------------------------------------------
Correctly classified                        65.19%
-------------------------------------------------
```

5 Validation out of sample et simualtion de paris

5.1 Validation out of sample

Le tableau 19 nous résume les performances out of sample des 4 meilleurs modèles estimés. Nous en avons retenu 2 pour la base favori et 2 pour la base rank. L'évaluation out of sample permet de retrouver nos résultats des estimations, i.e que l'approche basé

sur le classement prédit mieux la spécificité de l'évènement, mais l'approche basée sur les favori des book a une prédiction plus sensible. D'après nos résultats, l'approche basée sur le favori des bookmakers est la plus performante en terme de taux d'erreur.

Toutefois, pour évaluer la performance réelle du modèle, il faut voir si nous pouvons en tirer un bénéfice. En ce sens, nous allons simuler une prise de paris à partir du modèle 4 ayant la meilleur prédiction out of sample pour l'approche par les cotes et le probit 4 ayant la meilleur prédiction out of sample pour l'approche avec le classement. Nous reprenons l'échantillon test pour effectuer la simulation.

TABLE 19 – Résultats out of sample des 4 meilleures modèles

	base rank		base favori	
	probit 2	probit 4	modele 2	modele 4
sensitivité (Y=1)	57,14%	68,57%	68,29%	73,17%
spécificité (Y=0)	50%	38,89%	16,67%	16,67%
Taux d'erreur	45,28%	41,51%	43,40%	39,62%

5.2 Application des modèles aux paris

Nous devons définir des conditions sur la prise des paris. Pour ce qui est du modèle 4, nous allons nous focaliser sur la prise des paris sur les joueurs favoris des bookmakers. Prendre des paris sur les outsiders ne serait pas rentable car la spécificité du modèle est très faible. En revanche pour le probit 4, la spécificité est meilleure. En ce sens, nous prendrons les paris pour les 2 catégories.

Pour prendre les paris nous fixons comme condition que :

— La probabilité calculé de victoire du favori soit d'au moins 55% afin d'éviter les paris de type 50/50 trop incertains.

— La cote minimale du paris est 1.3 afin d'éviter de prendre un risque pour peu de rendement potentiel pour Y=1.

— Enfin, la cote estimée par le modèle doit être inférieure à la cote proposée par les bookmakers [6]. En ce sens, le paris est dis "value" car il peut nous rapporter plus que ce que nous espérons d'après notre prédiciton.

6. Comme les bookmakers le font, nous augmentons artificiellement la probabilité de victoire des deux joueurs. En moyenne les bookmakers augmentent de 6% ces probabilités d'après Lisi et Ganella (2017).Ceci leur permet de dégagé une marge

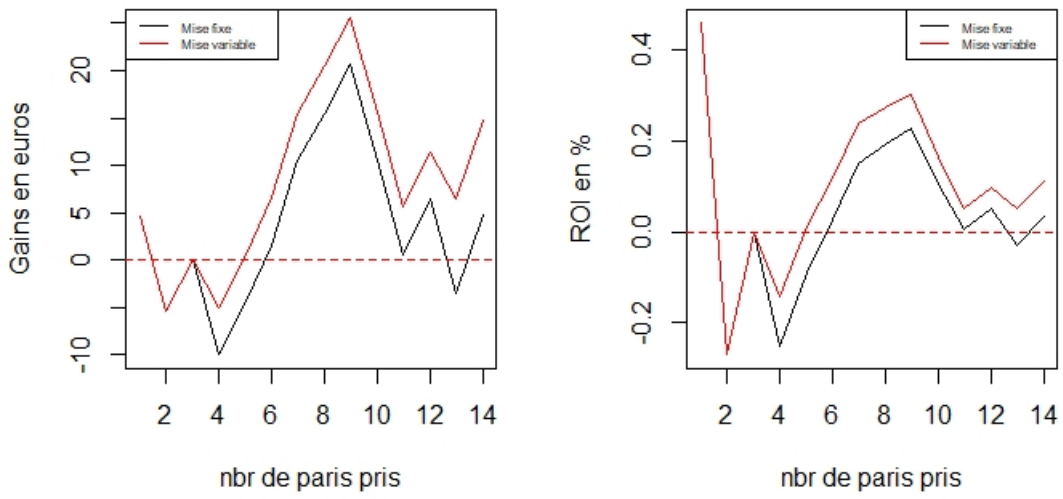
Dans le cadre de la prise de paris avec le probit 4, nous avons comme conditions supplémentaires de sélection

— La probabilité calculée que le joueur le moins bien classé gagne doit être également de 55% au minimum.

— La cote minimale pour la prise de paris dans le cas ou Y= 0 est prédit doit être de 1.5 et la cote maximale doit être de 3.5.

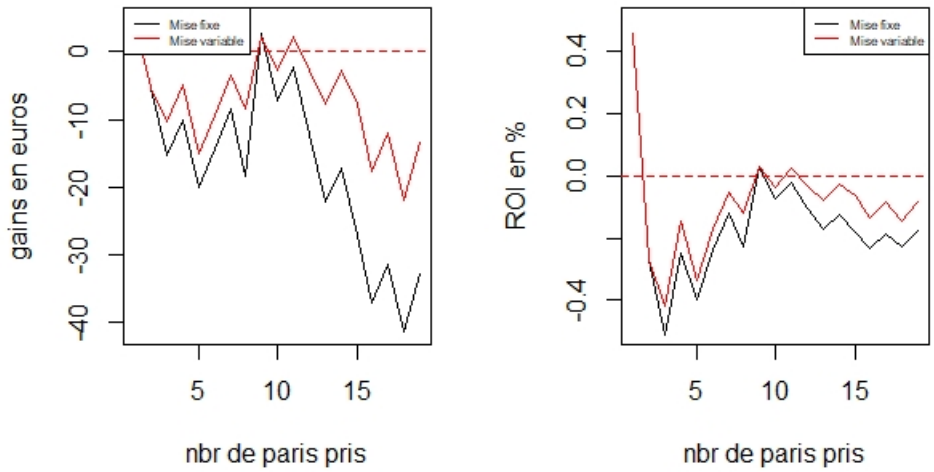
En outre, dans le milieu des parieurs professionnels, la gestion de mise est importante pour maximiser les gains. En ce sens, nous appliquerons également une gestion de mise sur notre simulation que nous comparerons à une mise fixe. Dans le cadre de la prise de paris des joueurs favori au sens des bookmakers, la mise s'effectura en fonction du spread des cotes, i.e l'écart entre la cote prédite et la cote des bookmakers. Plus le spread est grand plus la mise sera grande. Nous prenons deux niveaux de mise fictives, 5€ si le spread est inférieur à 0.1 et 10€ sinon. Dans l'approche par le classement, nous prenons une autre approche ou l'on mise 10€ quand le mieux classé est prédit mais également quand l'outsider est prédit vainqueur sous condition que sa cote soit inférieur à 1.9. Nous miserons 5€ pour les cotes supérieures à 1.9.

FIGURE 7 – Gains et retour sur investissement avec le Modèle 4



La figure 7 montre les gains ainsi que le retour sur investissement obtenu [7] avec l'utilisation du modèle 4. On constate que l'utilisation de la mise variable génère plus de bénéfice. D'après nos critères de sélections, sur 53 matchs disputés, nous avons retenus 14 paris et 9 d'entres eux se sont avérés gagnants.

FIGURE 8 – Gains et retour sur investissement avec le Probit 4



La figure 8 montre les gains ainsi que le retour sur investissement obtenu avec l'utilisation du probit 4. D'après nos critères de sélections, sur les 53 matchs disponibles, 19 paris ont étés pris et 9 ont été gagnants soit moins de 50%. Par conséquent, peu importe le choix du type de mise, ici nous sommes en perte nette si nous utilisons ce modèle selon les critères de sélection vu précédemment. Néanmoins, ces résultats doivent être confimer sur un plus grand échantillon out of sample. Il est possible que les résultats soit positif avec un plus grand nombre de matchs sélectionnés. Nous pouvons également chercher d'autres critères de sélections des paris ou bien miser avec 3 niveaux de confiance également.

Le tableau 20 résume les différents résultats de la simulation. En définitive, le modèle 4 est le plus performant en terme de prédiction out of sample, de réussite sur les paris prit, et de gains monétaire.

7. Le retour sur investissement est le bénéfice totale sur la somme totale engagée

TABLE 20 – Résumé des résultats de la simulation

	Modèle 4	Probit 4
Nbr paris retenus	14	19
% réussite	64.3%	47.36%
Cote moyenne	1.598	1.835
ROI mise fixe	3.4%	-17.36%
ROI mise variable	11.4%	-8,5%

6 Conclusion et Discussion

Notre objectif était de modéliser l'issue des matchs de tennis en prédisant à la fois le vainqueur et le nombre de set. Nous avons pris deux approches différentes afin de prédire l'issue du match : nous avons cherché à predire le vainqueur du match en partant de la différence entre le favori et l'outsider au sens des bookmakers, mais aussi en fonction du classement. Pour ceci, nous avons décidé de réaliser un modèle biprobit, car nous pensions que la victoire d'un joueur et le nombre de sets disputés au cours d'un match étaient en partie liés. Il s'avère au final que non ; le ρ n'est significatif dans aucun des modèles estimés. En revanche, dans certains modèles , le paramètre ρ tend à être significatif [8].

Nous avons ensuite estimé deux probit séparés pour prédire les deux évènements. Avec nos données, nous n'avons pas réussi à prédire correctement le nombre de set disputés. Les modèles ne prédisent pas Y=0. En somme, nos variables explicatives ne sont pas adéquates pour prédire cet aspect de la rencontre. Un questionnement plus approfondie est nécéssaire afin de trouver les indicateurs pour obtenir une meilleure modélisation de cet évènement.

En ce qui concerne la prédiction du vainqueur du match ; comme nous l'attendions, nos variables explicatives sont toutes significatives et les coefficients associés sont conformes à nos attentes à l'exeption du nombre de victoire sur les 5 derniers matchs qui n'est pas significative. Cependant, nos modèles diffèrents dans leur performance. En fonction des approches, le taux de sensitivité et de spécificité n'est pas le même. L'approche par le favori au sens des bookmakers nous donne des modèles estimés très sensibles et peu spécifiques. En changeant d'approche et en cherchant à prédire la victoire du joueur le mieux classé, nous obtenons des modèles plus équlibrés, i.e moins sensible et plus spécifique. On note par ailleurs que nous obtenons un taux d'erreur d'environ 40% pour chaque modèle en validation out of sample, ce qui diffère peu des résultats in sample. En

8. les p-value sont proche de 0.1.

ce sens, les résultats semblent robuste. Nous avons testé les 2 meilleurs modèles de chaque approche pour la prise de paris, en se fixant des critères de sélection des paris à prendre. Selon le modèle prédisant le favori au sens des bookmakers, nous avons retenu 14 paris à prendre. En comparant un système de mise fixe et d'une mise variable en fonction du spread entre les cotes prédites par le modèle et les cotes réelles, nous obtenons un retour sur investissement de 11.3%. Enfin, le modèle prédisant la victoire du joueur le mieux classé obtient un résultat inférieur. Avec ce modèle nous retenons 19 paris, et un retour sur investissement de -8.4% avec un système de mise variable.

Toutefois ces réultats peuvent être améliorés en rajoutant des facteurs explicatifs, en prenant un échantillion plus grand ou en changeant le type de modélisation. Il serait intéréssant de tester des modèles non linéaires sur nos variables. Par exemple, la variable Dmatch peut influer de manière non linéaire sur l'issue du match ou bien le nombre de set disputés.

7 Bibliographie

Lisi F., Zanella G. (2017). Tennis betting : Can statistics beat bookmakers ?, *Electronic Journal of Applied Statistical Analysis . 2017, Vol. 10 Issue 3*, p790-808.

Klaasen F., Magnus J. (2003). Forecasting the winner of a tennis match, *European Journal of Operational Research 148*, p257–267.

Kovalchik S. (2016). Searching for the GOAT of tennis win prediction, *Journal of Quantitative Analysis in Sports*, p127-138.

8 Annexes

Annexe 1 : VIF des modèles probit du vainqueur homoscédastiques

	Modèle 1	Modèle 2	Modèle 3	Modèle 4
Dmatch	1.44	1.18	1.44	1.19
DLpts	2.16	2.16		
D_class			2.13	2.12
D_W_L5	1.3		1.32	
Odd_S	1.97	1.96	1.94	1.94

Annexe 2 : VIF des modèles probit du vainqueur hétéroscédastique

	Modèle 2 (all)	Modèle 2 (Dmatch)	Modèle 4 (D_Class)	Modèle 4 (Dmatch)
Dmatch	11.98	3.56	9.89	3.59
DLpts	60.04	2.94		
D_class			4.34	12.22
Odd_S	91.17	2.65	1.14	8.20

Note : entre parenthèse est précisé la variable potentiellement responsable de l'hétéroscédasticité des erreurs du modèle

Annexe 3 : Test Rosner et Grubbs sur les variables quantitatives base favori

```
Alternative Hypothesis:          Up to 5 observations are not
                                 from the same Distribution.

Type I Error:                    5%

Number of Outliers Detected:     0

  i    Mean.i     SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
1 0 0.1458333 2.310818     7      89 2.966122   3.593225   FALSE
2 1 0.1099476 2.262615    -6      50 2.700392   3.591646   FALSE
3 2 0.1421053 2.224400     6     104 2.633472   3.590057   FALSE
4 3 0.1111111 2.188785    -5      58 2.335136   3.588458   FALSE
5 4 0.1382979 2.162396    -5      65 2.376205   3.586849   FALSE
```

(a) Variable Dmatch

```
Alternative Hypothesis:          Up to 6 observations are not
                                 from the same Distribution.

Type I Error:                    5%

Number of Outliers Detected:     0

  i    Mean.i      SD.i     Value Obs.Num    R.i+1 lambda.i+1 Outlier
1 0 0.5782480 0.8832464 -2.439395      14 3.416536   3.593225   FALSE
2 1 0.5940472 0.8579347 -1.773227      10 2.759271   3.591646   FALSE
3 2 0.6065065 0.8426978  2.848750      20 2.660792   3.590057   FALSE
4 3 0.5946428 0.8288743  2.812797      76 2.676104   3.588458   FALSE
5 4 0.5828441 0.8150186  2.737485      54 2.643670   3.586849   FALSE
6 5 0.5713220 0.8017067 -1.493989      89 2.576142   3.585230   FALSE
```

(b) Variable DLpts

```
Alternative Hypothesis:          Up to 4 observations are not
                                 from the same Distribution.

Type I Error:                    5%

Number of Outliers Detected:     0

  i     Mean.i      SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
1 0 0.06458333 0.2556090   0.8      45 2.877116   3.593225   FALSE
2 1 0.06073298 0.2506360   0.8      47 2.949564   3.591646   FALSE
3 2 0.05684211 0.2454464  -0.6      46 2.676112   3.590057   FALSE
4 3 0.06031746 0.2413656  -0.6     112 2.735756   3.588458   FALSE
```

(c) Variable D_W_L5

Annexe 4 : Test Rosner et Grubbs sur les variables quantitatives base rank

```
        Grubbs test for one outlier

data:  base_rank$Dmatch
G = 2.87545, U = 0.95648, p-value = 0.7082
alternative hypothesis: lowest value -7 is an outlier
```

(a) Variable Dmatch

```
Alternative Hypothesis:          Up to 3 observations are not
                                 from the same Distribution.

Type I Error:                    5%

Number of Outliers Detected:     0

  i    Mean.i      SD.i    Value Obs.Num    R.i+1 lambda.i+1 Outlier
1 0 0.8389297 0.6393416 2.848750      20 3.143578   3.593225   FALSE
2 1 0.8284071 0.6241297 2.812797      76 3.179451   3.591646   FALSE
3 2 0.8179629 0.6088139 2.737485      54 3.152888   3.590057   FALSE
```

(b) Variable DLpts

```
        Grubbs test for one outlier

data:  base_rank$D_Class
G = 2.24531, U = 0.97347, p-value < 2.2e-16
alternative hypothesis: highest value 4 is an outlier
```

(c) Variable D_Class

```
        Grubbs test for one outlier

data:  base_rank$D_W_L5
G = 3.09401, U = 0.94962, p-value = 0.3357
alternative hypothesis: highest value 0.8 is an outlier
```

(d) Variable D_W_L5

Table des figures

Liste des tableaux