Maths et sports, quand les statistiques sont plus que des nombres

Christophe Ley

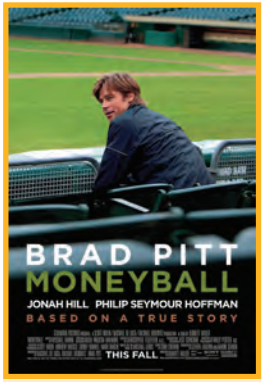
Professeur de statistique à l'Université de Gand
Président de la Société luxembourgeoise de statistique

Mathématiques et sports – à première vue, cette combinaison semble fort peu naturelle. D'accord, il faut savoir compter pour suivre les scores lors de matchs de football ou de tennis, et il faut savoir lire les chronomètres et additionner des temps pour savoir qui porte le maillot jaune lors du Tour de France. Mais nous sommes encore bien loin de « vraies » maths, et donc nous restons dans le schéma de pensées que maths et sports, ça ne va pas vraiment bien ensemble. Et en effet, pendant très longtemps les deux domaines ont vécu en parallèle, sans véritable intersection.

États-Unis, 2002 : le match de baseball qui a tout changé

La situation a complètement changé à la suite du succès inattendu de l'équipe de baseball californienne Oakland Athletics lors de la saison 2002. Dû à des contraintes monétaires, leur directeur général Billy Beane (né en 1962) avait adopté une toute nouvelle manière pour recruter des nouveaux joueurs : il s'était basé sur la *sabermétrie*, une approche statistique du baseball. Le mot tire son origine de l'acronyme SABR (pour Society for American Baseball Research). En d'autres termes, Beane utilisait des statistiques avancées et un raisonnement mathématique pour identifier quels joueurs sous-évalués étaient sur le marché et pourraient au mieux renforcer son équipe.

Cette approche scientifique a été couronnée de succès et a valu à Beane de faire l'objet du livre *Moneyball: The Art of Winning an Unfair Game* (Michael Lewis, W.W. Norton and Company Inc., 2003), livre qui aura servi par la suite comme base pour le film *Moneyball* (en français, *le Stratège*) qui est sorti en salles de cinéma en 2011 avec, comme acteur principal dans le rôle de Billy Beane, nul autre que Brad Pitt.



Le film a été six fois nominé aux Oscars 2012.

On peut imaginer la redondance qu'ont eue le livre et le film sur la mentalité des professionnels du milieu sportif. Depuis lors, la plupart des clubs cherchent à s'assurer les services de statisticiens ou analystes de données, de mathématiciens, d'informaticiens. Non seulement la manière de constituer une équipe, mais aussi les tactiques, les entraînements, la santé des joueurs, bref de nombreux aspects divers ont été repensés et basés sur une analyse de données scientifique. Le domaine des *sports analytics* est en pleine effervescence depuis plus d'une décennie ! C'est en particulier le cas au football.

Le football : un sport populaire aux règles faciles à comprendre

Lors d'un match de football, le résultat est clair : victoire, défaite ou match nul. Le résultat peut certes être contesté entre fans après le match, mais il se lit facilement et c'est là l'un des points forts de ce sport : tout le monde peut y jouer et comprendre les règles. Le classement des ligues nationales est aussi assez simple : une équipe reçoit trois points pour une victoire, un pour match nul, aucun pour une défaite, et les points sont additionnés. À la fin de la saison, l'équipe en première division nationale avec le plus de points se voit attribuer le titre de champion, les meilleures équipes peuvent jouer la Ligue des champions (respectivement la Ligue d'Europe en Europe), et les équipes terminant en derniers sont reléguées vers la seconde division nationale. Cet engrenage a toujours très bien fonctionné… jusque l'an dernier, où la pandémie de Covid-19 a marqué un stop net à toute activité sportive.

Les ligues ont dû s'arrêter de jouer, et on se demandait comment faire : peut-on continuer à un moment donné avec les ligues, ou faut-il déclarer la saison terminée ? Alors que des ligues comme la Bundesliga ou Premier League ont repris leur service, la France a déclaré le 28 avril 2020 que la saison était terminée, et qu'on allait considérer le classement au moment où la saison a dû être stoppée comme final (comme toutes les équipes n'avaient pas joué un nombre équivalent de matchs, les instances ont basé le classement sur le ratio points/matchs joués). Cette décision a suscité le courroux de plusieurs clubs, notamment l'Olympique lyonnais, privé d'une participation en coupe d'Europe.

De fait, il n'y avait pas de solution miracle toute prête, parce que personne n'avait imaginé pareil scénario. Du coup, des chercheurs de plusieurs pays européens ont entamé des recherches pour voir comment évaluer de manière plus correcte une saison stoppée prématurément.

Vers un classement alternatif : probabilités et lois de Poisson

Comment donc peut-on envisager un classement alternatif ? Explorons une approche, basée sur un classement probabiliste, c'est-à-dire un classement qui indique, pour chaque équipe, la probabilité d'arriver à chaque position en fin de saison. Pour y parvenir, il faut modéliser le résultat d'un match de football par une loi probabiliste.

Une *loi probabiliste* est une formule mathématique qui cherche à décrire le résultat d'un évènement aléatoire (pensez à la célèbre cloche de Gauss pour décrire bon nombre de phénomènes naturels, comme la distribution de la taille des gens). Un match de football est caractérisé par deux équipes adverses qui peuvent marquer chacune un certain nombre de buts pendant un intervalle de temps bien déterminé (quatre-vingt-dix minutes, plus arrêts de jeux). Un choix naturel pour décrire un phénomène aléatoire X qui peut se produire plusieurs fois sur un intervalle de temps déterminé est la *loi de Poisson*, avec comme formule $P(X=k)=\exp(-\lambda)\times\lambda^k/k!$, où l'entier positif k représente le résultat, le paramètre $\lambda\geq 0$ est la moyenne attendue de X et $k!=k\times(k-1)\times(k-2)\times\ldots\times 2\times 1$ est la factorielle de k. Une version légèrement plus élaborée, qui tient compte de l'interaction entre les deux équipes, permet de modéliser le résultat d'un match de football en disant que X_i est le nombre de buts marqués par l'équipe i dans le match. En assignant *une force de jeu* s_i à chaque équipe et en la reliant au paramètre λ_i, il devient possible d'estimer la force de jeu de toute équipe en déterminant, sur tous les matchs joués lors d'une saison, les paramètres s_i qui « collent le mieux » avec les résultats obtenus jusque-là.

Ainsi, à tout instant de la saison, on connaîtra les forces de chaque équipe, et ces paramètres tiennent compte des adversaires déjà rencontrés. Quand une saison doit être stoppée, il suffit alors de simuler un grand nombre de fois les matchs restants via la formule de Poisson et les forces s_i calculées au moment de l'arrêt de la saison, et on obtient ainsi pour chaque équipe le nombre de fois qu'elle est placée sur chaque position au classement final. Une simple division par le nombre de simulations donne alors les pourcentages mentionnés. En guise d'illustration, ci-dessous, la première figure contient le classement officiel de la Ligue 1 pour la saison 2019–2020, à comparer avec le classement probabiliste (seconde figure) obtenu pour la même saison.

	Team	Points	Win	Draw	Loss	Goals	Goals against	Goal difference	Matches	Points per match
1	PSG	68	22	2	3	75	24	51	27	2.52
2	Marseille	56	16	8	4	41	29	12	28	2.00
3	Rennes	50	15	5	8	38	24	14	28	1.79
4	Lille	49	15	4	9	35	27	8	28	1.75
5	Nice	41	11	8	9	41	38	3	28	1.46
6	Reims	41	10	11	7	26	21	5	28	1.46
7	Lyon	40	11	7	10	42	27	15	28	1.43
8	Montpellier	40	11	7	10	35	34	1	28	1.43
9	Monaco	40	11	7	10	44	44	0	28	1.43
10	Strasbourg	38	11	5	11	32	32	0	27	1.41
11	Angers	39	11	6	11	28	33	-5	28	1.39
12	Bordeaux	37	9	10	9	40	34	6	28	1.32
13	Nantes	37	11	4	13	28	31	-3	28	1.32
14	Brest	34	8	10	10	34	37	-3	28	1.21
15	Metz	34	8	10	10	27	35	-8	28	1.21
16	Dijon	30	7	9	12	27	37	-10	28	1.07
17	St. Etienne	30	8	6	14	29	45	-16	28	1.07
18	Nîmes	27	7	6	15	29	44	-15	28	0.96
19	Amiens	23	4	11	13	31	50	-19	28	0.82
20	Toulouse	13	3	4	21	22	58	-36	28	0.46

Le classement officiel de la Ligue 1.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
PSG	100																			
Marseille		77	18	5																
Rennes		12	41	36	8	2	1													
Lille		11	37	39	9	3	1													
Lyon			3	10	30	18	13	9	6	4	3	2	1							
Reims				3	12	15	15	13	12	10	8	6	4	2						
Montpellier				3	11	13	13	12	12	10	9	7	5	3	1					
Bordeaux				2	7	11	12	12	12	11	10	9	7	4	1					
Nice				1	7	10	12	12	12	12	11	9	7	4	1					
Strasbourg				1	6	9	11	11	12	12	11	10	9	5	2	1				
Monaco				1	4	7	9	10	12	12	13	12	10	6	3	1				
Nantes				1	4	7	8	10	11	12	12	12	11	7	3	1				
Angers					2	4	5	7	9	11	13	15	15	11	5	2	1			
Metz								1	1	2	4	7	11	21	22	16	10	4		
Brest								1	1	2	4	6	10	19	23	18	10	5	1	
Dijon											1	2	4	10	16	22	23	15	5	
St. Etienne											1	1	3	7	14	22	25	20	7	
Nîmes													1	3	6	12	22	36	18	
Amiens															1	3	8	20	66	2
Toulouse																			2	98

Le classement probabiliste obtenu (en pourcentages) après cent mille simulations. Les nombres sont arrondis au pourcent le plus proche, et les vides représentent des 0. Ainsi, le PSG a 100 % de chances de finir premier, Marseille a 0 % de chances de finir premier, 77 % de chances de finir deuxième, 18 % de chances de finir troisième, et ainsi de suite.

Une intelligence artificielle au retentissement médiatique inopiné

Cette modélisation mathématique d'un match de football n'est pas nouvelle. Comme la formule de Poisson se prête très bien à la prédiction de matchs de foot, il n'est pas surprenant qu'elle ait aussi été utilisée comme ingrédient essentiel d'une intelligence artificielle (IA) pour prédire la Coupe du monde 2018, que la France a remportée. Avec des collègues chercheurs, nous avons combiné cette formule avec des données économiques de chaque pays participant,

des données sportives comme l'âge moyen des joueurs, le niveau des clubs pour qui ils jouent, le nombre de joueurs évoluant à l'étranger, ou encore des données sur le coach. Nous avons entraîné un nouveau type d'IA, à savoir une forêt aléatoire hybride, avec ces données sur base des Coupes du monde 2002 à 2014, et puis avons simulé cent mille fois la Coupe du monde 2018 pour obtenir nos prévisions.

Le retentissement médiatique sur nos prédictions fut tout à fait inattendu et a montré à quel point les gens sont friands d'une prédiction mathématique d'un jeu réel comme le football. Alors que notre favori, l'Espagne, est sorti en huitièmes de finale (partiellement dû au licenciement de leur coach un jour avant le début de la compétition, fait que notre IA n'a plus pu prendre en compte faute de temps), nous avons terminé deuxièmes d'une compétition internationale de prédiction, soulignant la force de la combinaison maths et sports.

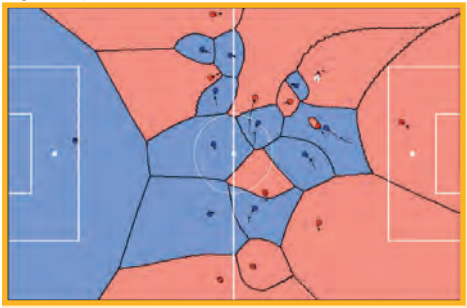
De nombreuses branches des mathématiques mobilisées pour le foot

Quand on parle de ligues nationales, il ne faut pas oublier qu'avant le début de saison, il fallait mettre sur pied le calendrier des matchs en tenant compte de multiples contraintes comme les coupes d'Europe, les coupes nationales ou encore la situation spéciale que deux équipes d'une même ville partagent le même stade (pensez au cas italien de San Siro, à Milan, qui héberge les grandes équipes du Milan AC et de l'Inter de Milan).

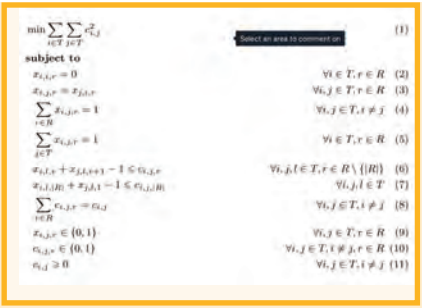
Même sans ces contraintes, il n'est pas facile de trouver un calendrier pour dix-huit équipes qui assure qu'en première moitié de saison chaque équipe rencontre chaque autre exactement une fois. Bien sûr, une fois que l'on dispose d'un modèle, c'est facile, mais essayez de vous prêter au jeu ! Le domaine des mathématiques qui traite ce genre de problèmes s'appelle la *recherche opérationnelle* ou l'*optimisation combinatoire*. L'exemple classique est le problème du voyageur de commerce, qui, pour une liste de lieux et des distances entre toute paire de lieux, doit trouver le chemin le plus court pour visiter chaque lieu une et une seule fois.

Trouver le meilleur calendrier d'une saison de football est un tel problème d'optimisation. Ce problème pose des défis aux chercheurs car le meilleur calendrier possible se définit également en termes d'équité : il faudrait éviter que les équipes aient trop de matchs à domicile de suite, et il faudrait également éviter *un effet de transfert*. Cet effet arrive quand par exemple une équipe rencontre systématiquement des adversaires qui, le tour précédent, ont dû souffrir une défaite cuisante contre une équipe du top comme par exemple le Paris Saint-Germain. Pareilles défaites peuvent démoraliser une équipe, qui

jouera moins bien le match suivant. Un tel effet de transfert doit être partagé entre toutes les équipes, idéalement. Pour parvenir à tenir compte de toutes ces contraintes et constamment améliorer les calendriers des ligues, les chercheurs ont recours à la *théorie des graphes* afin de formuler en termes mathématiques le problème d'optimisation. Ci-dessous la figure droite propose un exemple de telle formulation (sans expliquer les concepts en détails, ici $x_{i,j,r}$ vaut 1 si les équipes *i* et *j* se rencontrent au tour *r* et 0 sinon, $c_{i,j}$ est l'effet transfert de l'équipe *i* sur *j*, T est le nombre total d'équipes et R le nombre total de tours à jouer).



Extrait du chapitre
« Analysing Positional Data »



$$\min \sum_{i \in T} \sum_{j \in T} c_{i,j}^2 \quad (1)$$

subject to

$$x_{i,j,r} = 0 \qquad \forall i \in T, r \in R \quad (2)$$
$$x_{i,j,r} = x_{j,i,r} \qquad \forall i,j \in T, r \in R \quad (3)$$
$$\sum_{i \in R} x_{i,j,r} = 1 \qquad \forall i,j \in T, i \neq j \quad (4)$$
$$\sum_{j \in T} x_{i,j,r} = 1 \qquad \forall i \in T, r \in R \quad (5)$$
$$x_{i,l,r} + x_{j,l,r+1} - 1 \leq c_{i,j,r} \qquad \forall i,j,l \in T, r \in R \setminus \{|R|\} \quad (6)$$
$$x_{i,l,|R|} + x_{j,l,1} - 1 \leq c_{i,j,|R|} \qquad \forall i,j,l \in T \quad (7)$$
$$\sum_{r \in R} c_{i,j,r} = c_{i,j} \qquad \forall i,j \in T, i \neq j \quad (8)$$
$$x_{i,j,r} \in \{0,1\} \qquad \forall i,j \in T, r \in R \quad (9)$$
$$c_{i,j,r} \in \{0,1\} \qquad \forall i,j \in T, i \neq j, r \in R \quad (10)$$
$$c_{i,j} \geqslant 0 \qquad \forall i,j \in T, i \neq j \quad (11)$$

Extrait du chapitre
« Fairness trade-offs in Sports Timetabling »

de l'ouvrage *Science Meets Sports: When Statistics Are More Than Numbers,* pensé et édité pour un public très large par Christophe Ley et Yves Dominicy.

Même les tactiques au football sont améliorées grâce aux maths. Ci-dessus la figure gauche indique, sur une phase de jeu, les zones de contrôle de chaque joueur, c'est-à-dire les endroits qui, selon sa position et sa vitesse, seront couverts par lui en premier. Ces zones sont obtenues en voyant le terrain sous l'angle d'un système de coordonnées et en enregistrant les positions (*x*, *y*) des joueurs, de l'arbitre et de la balle ainsi que des évènements comme les passes, les tirs… L'analyse de ces données de position se fait par des techniques d'apprentissage de machine et des concepts mathématiques comme les diagrammes de Voronoï et les triangulations de Delaunay.

C. L.

Pour en savoir (un peu) plus

Science Meets Sports: When Statistics Are More Than Numbers. Christophe Ley et Yves Dominicy, Cambridge Scholars Publishing, 2020.

« Le basket à l'épreuve des stats. » Conférence de Rémy Mahfouf au Mathematic Park, Institut Henri Poincaré, Paris, samedi 7 octobre 2017, disponible en ligne.