Modélisation de la probabilité de réussite d'une passe au football

Sébastien Déjean [1] & J avier Lopez Sanchez [2] & Philippe Saint Pierre [1]

[1] *Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse et CNRS,* {*sebastien.dejean, philippe.saint-pierre*}*@math.univ-toulouse.fr*

[2] *Toulouse Université Club Tennis et Athletics Coaching Club Ramonville, jlsone1@hotmail.com*

Résumé. Depuis quelques années, le sport offre un environnement très dynamique pour l'analyse statistique de données. Que ce soit pour accompagner le développement de systèmes de paris sportifs, pour agrémenter les diffusions télévisées de rencontres sportives ou en vue d'optimiser les performances d'un athlète ou d'une équipe, l'analyse de données est une aide précieuse dans de tels contextes. Pour le football, plusieurs études ont conduit à populariser la notion d'*expected goal* traduisant la probabilité de marquer un but pour un joueur (ou une équipe) dans une situation donnée. À ce jour, cependant, peu de travaux semblent se consacrer à la probabilité de réussite d'une passe pourtant élément majeur dans tout sport collectif. C'est avec l'objectif de contribuer à ce sujet que nous avons utilisé la régression logistique pour modéliser la probabilité de réussite d'une passe en fonction de différentes variables explicatives. Nos analyses montrent des résultats qui, loin de révolutionner le football (plus une passe est courte et vers l'arrière, plus il est probable de la réussir) mettent en évidence de nombreuses perspectives pour l'optimisation de la performance d'une équipe.

Mots-clés. données sportives, statistique descriptive, régression logistique, visualisation de données

Abstract. In recent years, sport has provided a very dynamic environment for statistical data analysis. Whether it is to accompany the development of sports betting systems, to enhance television broadcasts of sports games or to optimise the performance of an athlete or a team, data analysis is a valuable aid in such contexts. For football, several studies have led to the popularisation of the notion of *expected goal* translating the probability of scoring a goal for a player (or a team) in a given situation. To date, however, few studies seem to focus on the probability of a successful pass, which is a major element in any team sport. With the aim of contributing to this subject, we used logistic regression to model the probability of a successful pass as a function of different explanatory variables. Our analyses show results which, far from revolutionising football (the shorter the pass and the further back it is, the more likely it is to be successful), highlight numerous perspectives for the optimisation of a team's performance.

Keywords. sports data, descriptive statistics, logistic regression, data visualization

1 Introduction

Qu'elle porte le nom de Big Data ou d'intelligence artificielle, l'analyse de données est en plein essor dans le domaine du sport. Cette approche a été popularisée notamment depuis 2011 avec la sortie du film Le Stratège (*Moneyball*) qui s'inspire d'une histoire vraie d'un analyste statisticien recruté par un club de base-ball. Dans la presse sportive, des articles de plus en plus nombreux font état de l'apport de ce domaine de l'IA dans le management des équipes, et de plus en plus de sociétés se positionnent sur ce marché très porteur pour proposer leurs services d'analystes.

À notre échelle, nous nous intéressons depuis quelques années à ce domaine en proposant des sujets de projets à des élèves ingénieurs de l'Institut National des Sciences Appliquées de Toulouse. Pour alimenter ces sujets, nous nous appuyons sur des bases de données disponibles gratuitement sur des sites comme `whoscored.com` ou `statsbomb.com`.

Pour le football, plusieurs études ont conduit à populariser la notion d'*expected goal* traduisant la probabilité de marquer un but pour un joueur (ou une équipe) dans une situation donnée. Les *expected goal* sont maintenant utilisés comme indicateur de la performance d'une équipe et intéressent également le domaine des paris sportifs (Steffen *et al.*, 2019). À ce jour, cependant, peu de travaux semblent se consacrer à la probabilité de réussite d'une passe pourtant élément majeur dans tout sport collectif (Ievoli *et al.*, 2021). Dans le travail présenté ici, nous nous sommes plus particulièrement intéressés aux passes lors d'un match de football. De nombreuses caractéristiques sont disponibles pour représenter une passe : le nom du joueur ou de la joueuse qui l'a effectuée, le moment du match, l'endroit du terrain d'où est partie la passe, l'endroit où elle est arrivée, la partie du corps (pied gauche, pied droit, tête, main si c'est un gardien de but...), etc... et surtout, une information concernant la réussite ou l'échec de la passe. C'est sur cette dernière information que nous nous sommes focalisés pour modéliser la probabilité de réussite d'une passe pendant un match de football. Une fois les données recueillies et pré-traitées, la mise en œuvre d'une régression logistique, déjà utilisée par Pollard and Reed (1997) pour mesurer l'efficacité de certaines stratégies au football, permet d'aboutir à un tel modèle.

Dans cet article, nous présentons d'abord les données qui nous ont servi pour réaliser ce travail. Dans une deuxième partie, nous abordons la partie modélisation pour laquelle nous proposons une visualisation originale des résultats basée sur une représentation schématique d'un terrain de football. Enfin, nous exprimons quelques perspectives de ce travail en lien avec la recherche d'indicateurs de performance susceptibles d'être révélés par des analyses statistiques.

2 Les données

2.1 Récupération des données

Les données que nous analysons dans ce travail proviennent du site StatsBomb ([1]). Elles sont disponibles gratuitement dans le cadre du *Free Data Offerings from StatsBomb Services* et peuvent être récupérées en utilisant le package `StatsBombR` ([2]) en vue d'analyses réalisées avec le logiciel R. Parmi l'ensemble des données disponibles, nous nous sommes restreints à celles relatives aux 64 rencontres de la Coupe du monde de football 2018.

2.2 Description succincte des données

Lors de ces 64 rencontres, **227 886** passes ont été enregistrées. Elles sont caractérisées par différentes variables comme le nom de l'équipe, la phase de jeu (*Regular play, Free kick, throw in...*), le joueur réalisant la passe, le joueur la recevant, la position du ballon au début de la passe, la longueur de la passe, l'angle que fait la passe avec la droite qui relie les 2 buts, la hauteur de la passe (3 modalités : au sol, à mi-hauteur, en hauteur), le fait d'être sous la pression de l'adversaire ou pas et, point important pour notre étude, la réussite (le ballon est arrivé dans les pieds d'un partenaire) ou l'échec (ballon récupéré par l'adversaire ou sorti hors des limites du terrain) de la passe. Le nombre de passes réussies est de **215 172** soit un taux de réussite de l'ordre de 95%.

Pour illustrer ces données, la figure 1 représente sur la localisation du début (à gauche) et de la fin (à droite) des passes avec un codage couleur indiquant la réussite (en vert) ou l'échec (en marron) de la passe.

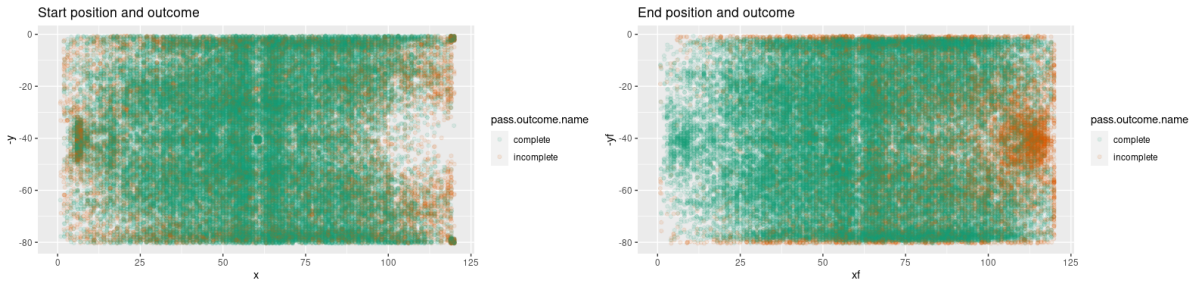

Figure 1: Représentation de la position du début (à gauche) et de la fin (à droite) d'une passe. La couleur indique la réussite (en vert) ou l'échec de la passe (en marron). Les points sont localisés en supposant que l'équipe en possession du ballon défend le but situé à gauche de la figure et attaque vers la droite.

On peut noter sur cette figure que peu de passes ont pour origine la surface de

[1] www.statsbomb.com STATSBOMB

[2] github.com/statsbomb/StatsBombR

réparation adverse (zone proche des buts adverses) et que les passes qui arrivent dans cette zone sont généralement ratées. Ceci est une évidence compte tenu de la forte densité de joueurs adverses dans une telle zone défensive.

Une autre illustration des données est proposée sur la figure 2. Cette fois-ci, la représentation en coordonnées polaires utilisant la longueur et l'angle de la passe, illustre là aussi une évidence footballistique : une passe vers l'arrière ou sur les côtés est plus susceptible d'être réussie qu'une passe vers l'avant.

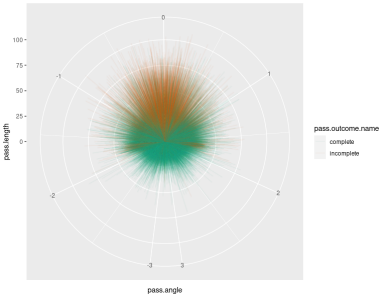

Figure 2: Représentation en coordonnées polaires de la longueur des passes (en yards) en fonction de l'angle de la passe (en radian, 0 indiquant une passe en direction des buts adverses, π ou $-\pi$ une passe en retrait). La couleur indique la réussite (en vert) ou l'échec de la passe (en marron).

3 Modélisation de la réussite d'une passe

Soit Y une variable à valeurs dans $0, 1$ à expliquer par p variables explicatives $\mathbf{X} = (X_1, ..., X_p)$. Le modèle logistique propose une modélisation de la loi de $Y|X = x$ par une loi de Bernoulli de paramètre $p(x) = \mathbb{P}(Y = 1|X = x)$ telle que :

$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{e^{\alpha+\beta\mathbf{x}}}{1 + e^{\alpha+\beta\mathbf{x}}}$$

où α et $\beta = (beta_1, ..., beta_p)$ sont les coefficients de régressions. Supposons qu'on observe un échantillon $(Y_i, \mathbf{X}_i)_{i=1,...,n}$ de n passes réussies ou non. Les estimateurs de β sont obtenus par maximum de vraisemblance. Les coefficients de régressions sont interprétable en terme d'odds ratio ($OR_k = e^{\beta_k}$). Le test du rapport de vraisemblance peut être utilisé pour tester la nullité des coefficients et comparer des modèles emboîtés afin de sélectionner un modèle. La probabilité de réussir une passe dans une situation donnée se déduit aisément du modèle logistique.

4 Application

4.1 Résultats d'un modèle de régression logistique

Nous présentons ici les résultats d'un modèle prenant en compte la position de départ de la passe dans l'axe du terrain (x), la valeur absolue de la position latérale ($abs(y)$) afin de considérer de façon équivalente les parties gauche et droite du terrain, le cosinus de

l'angle de la passe afin là aussi de respecter la symétrie du terrain, et enfin le fait d'être sous la pression de l'équipe adverse.

Table 1: Résultats de la régression logistique modélisant la probabilité de réussite d'une passe en fonction de la position d'origine de la passe, de la longueur de la passe, du cosinus de l'angle de la passe et du fait d'être sous la pression de l'équipe adverse

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	4.3090	0.0460	93.67	< 0.00001
x	-0.0164	0.0004	-36.99	< 0.00001
abs(y40)	-0.0190	0.0010	-19.71	< 0.00001
cos(pass.angle)	-1.5683	0.0228	-68.72	< 0.00001
pass.length	-0.0357	0.0007	-52.37	< 0.00001
under_pressurePressure_YES	-0.4048	0.0265	-15.30	< 0.00001

Le tableau 1 indique que toutes les variables du modèle sont significatives. Les estimations des coefficients s'interprètent logiquement en termes footballistiques. Par exemple, l'*odd ratio* associé au fait d'être sous pression de l'équipe adverse ($OR = e^{-0.4048} < 1$) indique qu'une pression diminue la probabilité de réussite d'une passe.

4.2 Visualisation des résultats

Le modèle permet d'estimer la probabilité de réussite d'une passe dans différentes situations. Par exemple, sachant la position d'un joueur qui fait une passe, on peut représenter sur un terrain de football la probabilité de réussite d'une passe en fonction de l'arrivée de la passe. Les probabilités de succès d'une passe peuvent ensuite être comparé dans différentes situations. La figure 3 représente les probabilités de succès dans une situation où le joueur qui fait la passe est dans une situation offensive (but adverse à droite de l'image) en position axiale (figure de gauche) ou décalé sur un côté du terrain (figure de droite).

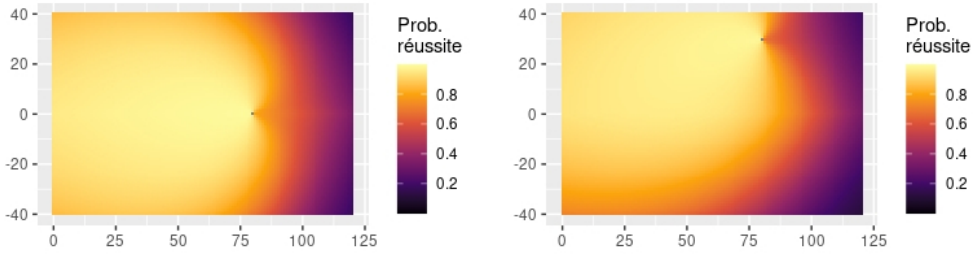

Figure 3: Représentation de la probabilité de réussite d'une passe à partir de 2 positions offensives : en position axiale (à gauche), une position excentrée (à droite). Les dimensions du terrain sont exprimées en yards.

5 Conclusion et perspectives

Nous n'avons pas révolutionné le football avec un modèle qui indique qu'une passe courte en retrait a plus de chances d'aboutir dans les pieds d'un partenaire qu'une passe longue vers l'avant. Il n'est pas non plus question de remplacer l'expérience et l'expertise d'un entraîneur chevronné qui a passé de longues années sur le bord des terrains. En revanche, nous pensons que l'analyse statistique de données sportives peut apporter un complément d'informations important permettant d'optimiser les performances aussi bien d'une équipe que d'un joueur ou d'une joueuse. L'objectif de ce travail était d'expliquer l'échec d'une passe et de proposer des résultats facilement utilisables pour analyser la performance d'un joueur ou d'une équipe.

D'un point de vue pratique, la prédiction de la réussite d'une passe dans chacune des situations rencontrées dans un match peut être déduite d'un tel modèle. Il est par exemple possible de définir, pour chaque joueur ou pour chaque équipe, un indicateur de performance basé sur la comparaison du nombre de passes prédites et effectivement observées. Cet indicateur peut être calculer sur des intervalles de temps afin de représenter l'évolution des performances au cours d'un match. Une autre perspective est de récupérer des données plus complètes afin de disposer de plus d'informations sur le déroulement du match (la position des autres joueurs par exemple) et ainsi inclure un plus grand nombre de variables dans l'analyse. Des méthodes de *machine learning* peuvent alors être mise en œuvre pour améliorer les qualités prédictives du modèles.

6 Remerciements

Les auteurs remercient Zoé Philippon et Lauriane Kiersnowski étudiantes à l'INSA de Toulouse qui ont contribué à ce travail dans le cadre d'un projet tutoré.

Bibliographie

Steffen, P. and Gerville-Réache, L. Bisoffi, N. (2019). Paris sportifs au football : l'intérêt des expected goals. Journées de Statistique, Nancy, France. ⟨hal-02150047⟩

Ievoli, R., Palazzo, L. and Ragozini, G. On the use of passing network indicators to predict football outcomes (2021), *Knowledge-Based Systems*, 222, 106997

Pollard, R. and Reep, C. (1997). Measuring the Effectiveness of Playing Strategies at Soccer, *Journal of the Royal Statistical Society. Series D (The Statistician)*, 46(4), pp. 541-550