Indicateurs de qualité des passes au football ; de l'intérêt de "casser des lignes"

Laurie Dussère[1] & Alexandre Gendreau[1] & Sébastien Déjean[2] & Javier Lopez Sanchez[3] & Philippe Saint Pierre[2]

[1] *Institut National des Sciences Appliquées, {dussere, gendreau}@insa-toulouse.fr*
[2] *Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse et CNRS, {sebastien.dejean, philippe.saint-pierre}@math.univ-toulouse.fr*
[3] *Universidad Catolica de Murcia, Spanish Sports University et Toulouse Université Club Tennis, 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 se consacrent, assez logiquement, à la probabilité de marquer un but, avec la popularisation relativement récente de la notion d'*expected goal*. En revanche, peu de travaux semblent se consacrer à la probabilité de réussir une passe qui est pourtant un élément majeur dans tout sport collectif. À partir des données de type *freeze frame* donnant la position des joueurs visibles dans le champ de la caméra durant une action, nous nous sommes ainsi attachés à définir des indicateurs de qualité des passes avec un intérêt particulier pour celles qui "cassent des lignes". Et nous avons ainsi étudié les liens entre les performances d'une équipe en termes de résultats obtenus et sa capacité à réaliser des passes difficiles.

Mots-clés. analyse de données sportives, modèle linéaire, forêts aléatoires.

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 been dedicated to scoring a goal, with the increasing popularity of the *expected goal*, translating the probability of scoring a goal for a player (or a team) in a given situation. On the contrary, few works focused on the passes, whereas it is of major importance in collective sports. Based on freeze frame data which provide player position's during one event, we have focused our work on defining quality indicators of passes, with a specific attention on "line-breaking passes". And we have studied the links between resultats obtained by a team and its ability to complete difficult passes.

Keywords. sports data analysis, linear modelling, random forests.

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 baseball. Dans la presse sportive, des articles de plus en plus nombreux font état de l'apport de ce domaine de la statistique 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 dans une situation donnée. Les *expected goal* sont maintenant utilisés comme indicateur de la performance d'un joueur[1] ou 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 à partir de données du site `statsbomb.com`[2]. De nombreuses informations sont disponibles pour caractériser 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. Depuis 2021, certains jeux de données, estampillés StatsBomb360, contiennent également la position des joueurs ou des joueuses visibles dans le champ de la caméra au moment d'un événement, même si ceux-ci ou celles-ci ne participent pas effectivement à l'action en cours. C'est en exploitant cette nouvelle source de données que nous nous sommes intéressés à l'élaboration d'indicateurs de difficulté des passes réalisées par les joueurs d'un match ; un joueur ou une joueuse performant étant capable de réaliser des passes difficiles. Nous nous sommes en particulier attachés à évaluer le principe de "casser des lignes", terminologie correspondant à la situation dans laquelle une passe réalisée par un joueur parvient à un partenaire en étant passé au milieu de plusieurs joueurs adverses globalement alignés dans la largeur du terrain.

Dans cet article, nous présentons d'abord brièvement les données qui nous ont servi pour réaliser ce travail. Dans une deuxième partie, nous définissons des indicateurs de difficulté d'une passe. Nous proposons ensuite une analyse descriptive de ces indicateurs

[1] https://www.lequipe.fr/Football/Actualites/Erling-haaland-un-ovni-statistique-dans-l-histoire-de-la-premier-league/1372012

[2] www.statsbomb.com STATSBOMB

au niveau des joueurs, puis au niveau des équipes. Enfin, nous présentons une modélisation de la performance des équipes évaluée via le nombre de buts, le nombre de tirs ou les *expected goals*, en fonction des indicateurs de difficultés des passes que nous avons définis.

2 Les données

2.1 StatsBomb et StatsBomb360

Les données que nous analysons dans ce travail proviennent du site StatsBomb. Elles sont disponibles gratuitement dans le cadre du *Free Data Offerings from StatsBomb Services* et peuvent être récupérées en utilisant le package `StatsBombR`[3] en vue d'analyses réalisées avec le logiciel R. À noter que le package `statbombpy` existe aussi pour les utilisateurs de Python. Parmi l'ensemble des données disponibles, nous nous sommes restreints à celles relatives aux rencontres de l'Euro 2020.

Ces données intègrent le principe du *freeze frame* qui permet, pour un événement donné (passe, tir...) de localiser les joueurs présents dans le champ de la caméra (StatsBomb, 2021)

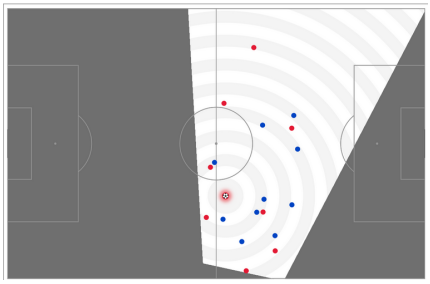

Figure 1: Illustration des données freeze frame. Source StatsBomb (2021)

Nous avons récupéré les données dont l'analyse est proposée dans ce travail grâce au script R suivant :

```
> devtools::install_github("statsbomb/StatsBombR")
> library(StatsBombR)
> Comp <- FreeCompetitions()
> Matches <- FreeMatches(Comp)
> Matches <- Matches %>% filter(competition.competition_name == "UEFA Euro")
> data360 <- StatsBombFree360Events(MatchesDF = Matches])
> events <- StatsBombFreeEvents(MatchesDF = Matches)
```

[3] github.com/statsbomb/StatsBombR

2.2 Description succincte des données

Les données concernent les 51 matchs de l'Euro masculin 2020. Elles représentent 192692 événements (soit environ 4000 événements par match) caractérisés par 188 variables (données `events`). Parmi ces évènements, 166892 sont également caractérisés par les informations relatives au *freeze frame* (données `data360`). Ces données peuvent être liées aux données `events` grâce à un identifiant des événements. Elles fournissent les limites de la zone observée au moment de l'action ainsi que la position des joueurs, partenaires ou adversaires, visibles dans cette zone.

3 Définition d'indicateurs de difficulté des passes

L'objectif de cette section est d'exploiter le jeu de données décrivant les *freeze frame* à chaque évènement. En effet, dans la base de données `events`, seule la variable *under pressure* contient une information condensée sur la position des autres joueurs. Cette information n'est pas suffisamment précise pour envisager une modélisation fine de la probabilité de réussite d'une passe. Nous cherchons donc à dériver des indicateurs de difficulté d'une passe des données *freeze frame* disponibles pour chaque passe.

3.1 Indicateurs standards

Dans un premier temps, nous déclinons quelques indicateurs de difficulté de passes sur des principes relativement simples de distance gagnée et de nombre de joueurs adverses éliminés :

- distance gagnée en direction du but adverse : calculée en soustrayant la position du ballon à l'arrivée de la passe à la position au départ de la passe, dans le sens de la longueur du terrain. Cet indicateur a été décliné en 2 variables: l'une prenant en compte seulement les passes vers l'avant ; l'autre considérant à la fois les passes vers l'avant et en retrait.
- nombre de joueurs éliminés par une passe réussie ; un joueur est considéré comme éliminé par une passe si cette passe était en direction du but adverse et si, après la passe, le ballon se trouve entre lui et le but qu'il défend après la passe.
- position du ballon à l'arrivée de la passe
- nombre de joueurs restant à éliminer i.e. les joueurs présents dans le *freeze frame*, situés entre le point d'arrivée de la passe et le but adverse

3.2 Qu'est-ce qu'une passe qui casse une ligne ?

Afin de compléter ces premiers indicateurs, nous nous sommes intéressés à caractériser des passes qui "cassent des lignes" selon un vocabulaire qui tend à s'imposer dans le monde du football[4]. En effet, cet indicateur semble pertinent pour mesurer la difficulté d'une passe et par conséquent la performance d'un joueur qui réussit une telle passe.

Qu'est-ce qu'une ligne ? Dans notre contexte, une ligne (de joueurs) n'est pas une chose précisément définie. On peut en trouver différentes définitions[5] [6] sans qu'aucune ne soit unanimement reconnue. Dans le cadre de notre travail, nous avons utilisé la définition suivante :

- Une ligne est constituée par au moins 2 joueurs de la même équipe.
- Les joueurs sont éloignés d'au plus 3 mètres dans la longueur du terrain.
- Les 2 joueurs les plus éloignés doivent être à une distance d'au moins 8 mètres dans la largeur du terrain.

La figure 2 illustre cette définition.

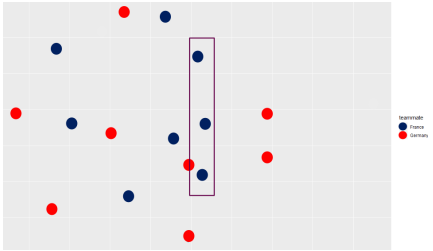

Figure 2: Illustration de notre définition d'une ligne de joueurs.

Cette définition présente l'avantage d'être paramétrable même si nous n'aborderons pas ici la sensibilité de nos résultats par rapport aux paramètres que nous avons fixés.

Quand une passe casse-t-elle une ligne ? De même, le fait de "casser une ligne" n'a pas de définition précise. Nous utilisons la définition suivante, illustrée sur la Figure 3. Une passe cassant une ligne :

- est réussie, c'est à dire est reçue par un partenaire

[4] "*Deschamps m'a demandé de continuer à casser des lignes*" affirme Upamecano qui veut "*apprendre*" de son sélectionneur, 01/09/2020

[5] https://statsbomb.com/articles/soccer/statsbomb-360-exploring-line-breaking-passes

[6] https://github.com/albizup/Line-Breaking-Passes-Algorithm

- est longue d'au moins 7 mètres
- a une trajectoire passant entre 2 joueurs adverses appartenant à une ligne
- est réalisée en direction du but adverse.

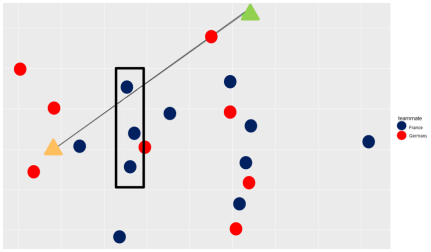

Figure 3: Illustration d'une passe qui casse une ligne de joueurs. Le triangle jaune, à gauche, indique le départ de la passe, le triangle vert, en haut, son arrivée.

De même que pour la définition d'une ligne, nous n'abordons pas ici la sensibilité de nos résultats par rapport à cette définition. Cela pourra naturellement faire l'objet de travaux ultérieurs. Ces définitions nous permettent de caractériser les passes réalisées selon qu'elles cassent une ligne ou pas.

4 Analyse descriptive des indicateurs de qualité des passes

En mettant en œuvre le calcul des indicateurs précédemment évoqués sur les données de l'Euro 2020, nous avons constitué un jeu de données de 484 lignes (1 ligne par joueur) et 27 variables contenant des indicateurs en lien avec la qualité moyenne des passes réalisées par chaque joueur. Ces indicateurs sont par exemple :

- nombre de passes ayant cassé une ligne
- nombre de passes réalisées
- nombre de passes réussies
- nombre de passes vers l'avant
- position moyenne du ballon à l'arrivée de la passe
- distance gagnée en direction du but adverse
- nombre de joueurs éliminés qu'ils fassent partie d'une ligne ou pas

- nombre de joueurs restant à éliminer (joueur se trouvant entre le ballon et le but qu'il défend).

Nous avons également construit un jeu de données avec les équipes en tant qu'individus en agrégeant les informations à l'échelle des joueurs. Cette démarche nous permet d'envisager plus naturellement de lier les indicateurs de qualité des passes avec un indicateur de performance. En effet, à l'échelle d'une équipe, la performance s'évalue facilement via notamment les buts marqués, les tirs réalisés ou encore les *expected goals* durant une rencontre. La seule prise en compte des buts marqués nous a semblé trop restrictive car le fait de marquer un but reste sujet à une part trop importante de facteurs incontrôlables.

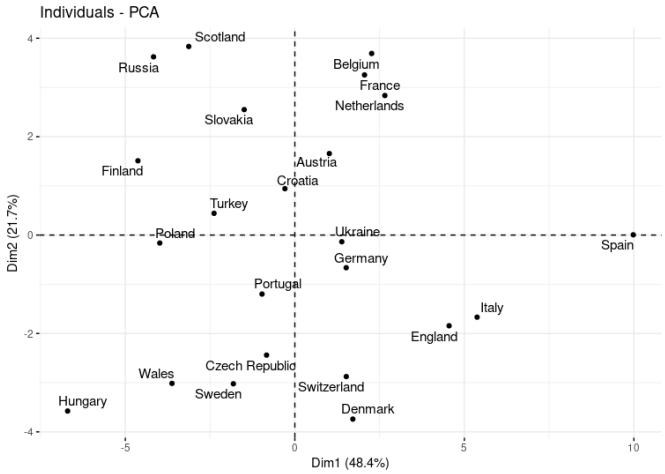

Figure 4: Représentation des équipes sur le premier plan principal.

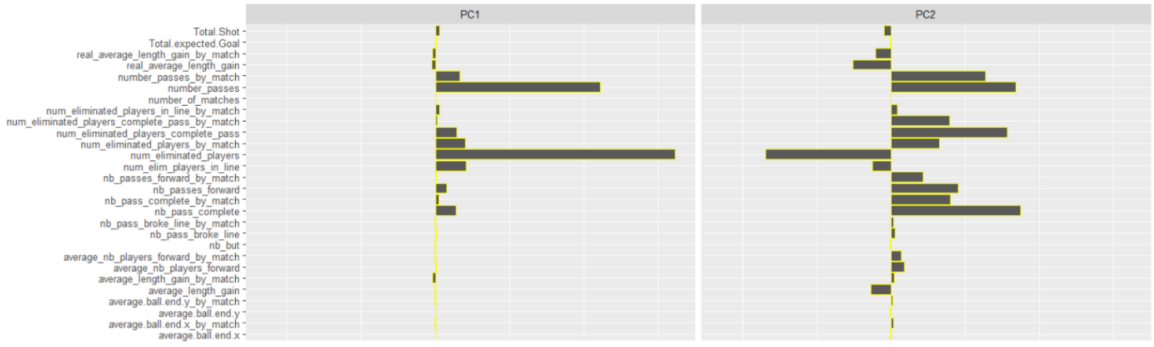

Figure 5: Coordonnées des variables sur les deux premières composantes principales.

Dans un premier temps, une Analyse en Composantes Principales des indicateurs est effectué à partir du jeu de données `Equipes` afin d'étudier les liens entre les indicateurs. Les résultats illustrent par exemple le fait que l'Espagne, positionnée à droite sur la

figure 4, se caractérise par un nombre de passes et un nombre de joueurs éliminés par passe importants (Fig. 5).

5 Modélisation de la performance des équipes par la qualité des passes

Dans cette partie, nous cherchons à établir un lien entre les performances d'une équipe et la qualité de son jeu de passes. Pour cela, nous évaluons les performances d'une équipe grâce à trois indicateurs que nous considérons séparément : le nombre de buts, le nombre de tirs et les *expected goals*. Chacune de ces trois caractéristiques est la variable à expliquer dans les modèles que nous mettons en œuvre. Ces modèles sont d'une part une régression linéaire avec sélection de variables, et d'autre part des forêts aléatoires. Les variables explicatives considérées sont les indicateurs de qualité de passes définis dans ce qui précède.

Le tableau 1 rassemble les variables sélectionnées soit par critère AIC pour la régression linéaire, soit par importance pour les forêts aléatoires, pour l'explication des trois indicateurs de performance des équipes.

Table 1: Synthèse des variables sélectionnées par les deux méthodes pour les trois variables à expliquer (xG pour *expected goals*).

	Nb buts	Nb tirs	xG
Régr. lin.	nb de joueurs à éliminer ; nb joueurs éliminés ; longueur moyenne gagnée ; nb passes cassant des lignes	nb de joueurs à éliminer ; nb joueurs éliminés ; longueur moyenne gagnée	nb de joueurs à éliminer ; nb joueurs éliminés ; longueur moyenne gagnée ; nb passes
Forêt	nb passes ; nb joueurs éliminés	nb passes ; nb joueurs éliminés	nb passes ; nb joueurs éliminés

En considérant l'algorithme des forets aléatoires, on observe que le nombre de passes et le nombre de joueurs éliminés semblent être pertinents pour expliquer les trois indicateurs de performances considérés. Le modèle de régression linéaire fait aussi ressortir le nombre de joueurs éliminés pour expliquer les trois indicateurs. Le nombre de passes est pertinent pour expliquer les *expected goals* alors que le nombre de passes cassant des lignes permet d'expliquer le nombre de buts marqués.

6 Conclusion et perspectives

Notre travail n'a pas de prétention à 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 est de contribuer à évaluer les performances d'un joueur ou d'une équipe à travers sa qualité de passe en exploitant les données de *freeze frame* fournissant la position de certains joueurs, partenaires et adversaires, autour du ballon. Il s'agit d'une utilisation relativement simple de ces données et de nombreuses perspectives sont envisageables pour les exploiter de façon plus approfondie. Par exemple, ne plus s'intéresser seulement à des lignes de joueurs, mais à des formes plus complexes, traduisant par exemple la compacité d'un bloc défensif ou la capacité d'une équipe à justement s'adapter face à une défense compacte.

7 Remerciements

Les auteurs remercient Julien Demeaux pour ces commentaires et conseils avisés sur l'analyse de ces données.

Bibliographie

- StatsBomb, (2021) StatsBomb 360 Freeze Frame Viewer: A New Release In StatsBomb IQ, `statsbomb.com/news/statsbomb-360-freeze-frame-viewer-a-new-release-in-statsbomb-iq`
- 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