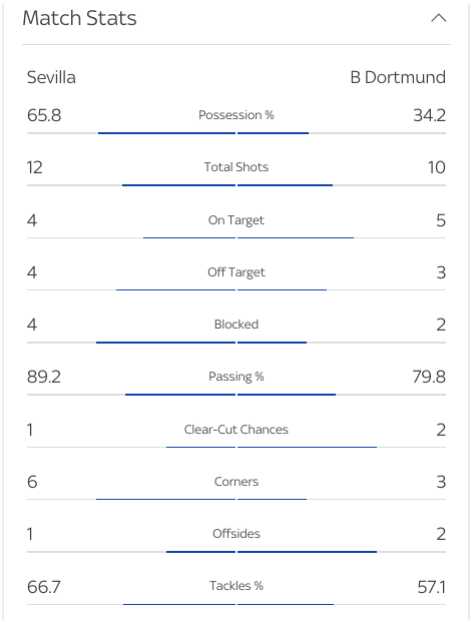
TRAVAIL ENCADRE DE RECHERCHE



Construction d'une mesure de dissimilarité sur des données de joueurs de football : saison 2014-2015

Achraff ADJILEYE,
Encadré par Christine KERIBIN dans le cadre du TER de M1 du master "Mathématiques Appliquées"

Juin-Août 2022

Table des matières

Note importante : Mon travail encadré de recherche (TER) est une reprise d'une partie de la thèse du docteur Serhat EMRE AKHANLI, lecteur en statistiques à "Muğla Sıtkı Koçman University" en Turquie, réalisée sur le thème : "**Distance construction and clustering of football player performance data**". Le rapport de la thèse est téléchargeable directement sous forme de document pdf en cliquant sur ce lien. Les idées et méthodes présentées dans les lignes qui vont suivre sont extraites des chapitres 2, 3 et 4 de ce rapport de thèse

1 Introduction

Les clubs de football professionnel investissent beaucoup de ressources dans le recrutement de nouveaux joueurs talentueux. Nous pensons que les méthodes traditionnelles de repérage par observation directe sur le terrain peuvent être améliorées par une interprétation intelligente de la grande quantité de données footballistiques sur les performances des joueurs aujourd'hui disponible et facile d'accès. Cette étude consiste à concevoir une méthodologie pour quantifier les similitudes entre les joueurs de football en se basant sur leurs données. Ce type d'étude peut être très utile pour les recruteurs et les managers de football lors de l'évaluation des joueurs, la détection de talents partout dans le monde et plus généralement le recrutement de nouveaux joueurs. Par exemple, certains managers souhaitent construire un effectif le plus stable possible sur le long terme, et face à un mercato (terme désignant le marché de transferts des joueurs de football) souvent très mouvementé et sujet à des surprises chaque année, avoir un tel système de comparaison de joueurs leur permettrait de vite trouver le remplaçant le plus adapté à un joueur sur le départ.

Les données dont nous disposons pour cette étude sont assez complexes avec de nombreux types de variables qui nécessitent chacun un traitement individuel. Elles seront pré-traitées de manière à ce que les informations des joueurs soient le mieux représentées et que les résultats soient le mieux interprétables d'un point de vue analyse footballistique. Ensuite, différentes mesures de dissimilarités seront construites sur les différents types de variables. Enfin, ces mesures seront agrégées pour avoir une mesure finale qui permettra de construire une matrice de dissimilarité sur les données.

Avant de rentrer dans le vif du sujet, nous fournissons quelques informations générales concernant le football pour les lecteurs qui ne sont pas familiers avec ce sport, puis nous évoquerons le rôle important des statistiques dans le recrutement de joueurs de nos jours.

2 A propos du football

2.1 Petite histoire

Le football est l'un des plus grands (peut-être le plus grand) sport mondial. Des millions de personnes se rendent régulièrement dans les stades de football, tandis que des milliards d'autres regardent les matchs à la télévision. Le sport le plus populaire au monde a une histoire longue et intéressante. Des preuves historiques et des sources suggèrent que le football a été pratiqué en Égypte, dans la Chine ancienne, en Grèce et à Rome. Vers 2500 av. J.-C., les Égyptiens jouaient à un jeu de football lors des fêtes de la fertilité. Vers 400 av. J.-C., une forme différente de jeu semblable au football, appelée « Cuju » (traduit « frapper le ballon avec le pied ») était populairement pratiquée en Chine. Le jeu était utilisé par les chefs militaires comme un sport de compétition pour garder les soldats en bonne forme physique. Dans la Rome antique, le jeu est devenu si populaire qu'il a été inclus dans les premiers Jeux Olympiques.

L'histoire contemporaine du football a été codifiée pour la première fois en 1863 à Londres, en Angleterre. Douze clubs londoniens ont créé des règles de football plus strictes, puis ont formé la Football Association, la même FA qui organise la populaire FA Cup (coupe de la FA) d'aujourd'hui. Les Britanniques ont également été considérés comme essentiels à la diffusion du jeu dans d'autres pays européens, tels que l'Espagne, la France, les Pays-Bas et la Suède, et dans le monde entier. Finalement, un organe directeur du football a été formé par ces pays et la FIFA (Fédération internationale des associations de Football) a été fondé. En 1930, la FIFA a organisé la première Coupe du monde de football en Uruguay avec 13 équipes. A partir de cette époque, le tournoi est organisé tous les quatre ans sauf en 1942 et 1946 où il n'a pas eu lieu en raison de la Seconde Guerre mondiale. Outre la Coupe du monde, plusieurs compétitions internationales de football entre équipes nationales existent, telles que les Championnats d'Europe (Euro), la Copa America (coupe d'amérique) et la Coupe d'Afrique des Nations (CAN). Au niveau national, les ligues les

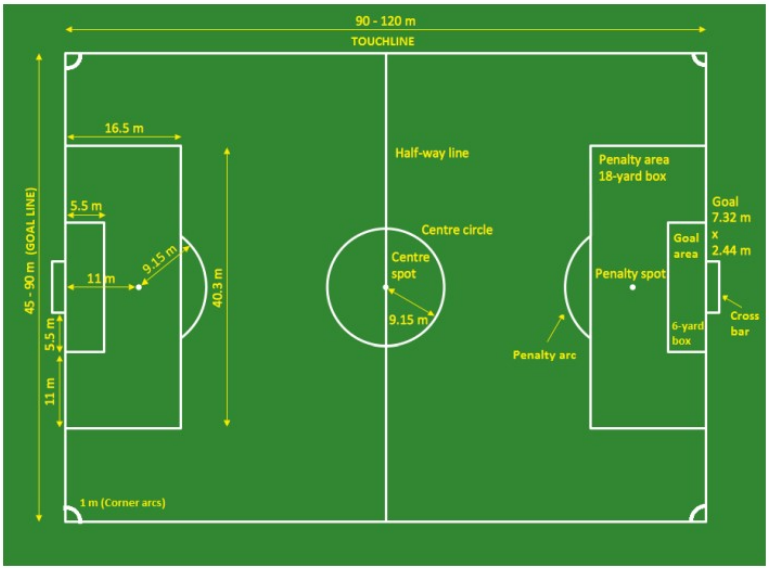


Figure 1 – Le terrain de Football et ses dimensions

plus relevées sont en Angleterre (Barclays Premier League), en Espagne (La Liga), en Allemagne (Bundesliga), en Italie (Serie A) et en France (Ligue 1 Uber Eats).

Bien que le football soit particulièrement connu comme un sport masculin, il est le sport d'équipe le plus important à avoir été joué par des femmes depuis l'époque des premiers matchs féminins enregistrés à la fin des années 1960 et au début des années 1970. La Coupe du Monde Féminine de la FIFA est organisée tous les quatre ans depuis 1991.

Pour les sources, lire cette page wikipédia : Histoire du football.

2.2 Le jeu du football

Le football se joue entre deux équipes de onze joueurs chacune, avec un ballon sphérique, et l'objectif principal est de marquer en faisant entrer le ballon dans le but adverse avec n'importe quelle partie du corps à l'exception des bras et des mains. L'équipe qui marque le plus de buts gagne. Si les deux équipes ont le même nombre de buts ou si aucune d'entre elles n'a marqué de but, cela est considéré comme un match nul. Les règles du football sont officiellement appelées les "Lois du Jeu" et sont disponibles sur la page wikipédia : Lois du jeu, mais nous allons les résumer brièvement ici.

Terrain de jeu : Le jeu se joue sur des surfaces naturelles ou artificielles, qui doivent être vertes et de forme rectangulaire. La figure 1 illustre les zones du terrain de football avec ses dimensions et ses marques (indications en anglais).

Ballon : Il doit être sphérique avec une circonférence de 68-70 cm et un poids de 400 à 450 grammes et en cuir (ou matière similaire) et d'une certaine pression.

Nombre de joueurs : Un match de football est disputé par deux équipes de onze joueurs maximum chacune, dont l'un est le gardien de but. Un match ne peut pas être joué si l'une ou l'autre des équipes a moins de sept joueurs. Dans les compétitions officielles de football, le nombre maximum de remplacements est de trois. Mais il est passé à 5 depuis la crise du covid-19. Les gardiens de but sont les seuls joueurs autorisés à prendre le ballon de la main, et ce dans la surface de réparation (y compris la surface de but). Chaque équipe aura un capitaine désigné.

Équipement du joueur : Les joueurs doivent porter un maillot, un short, des chaussettes, des protège-tibias et des chaussures de football. Les gardiens de but porteront en outre des gants rembourrés et devront porter un maillot qui les distinguera des joueurs de champ et des arbitres.

Arbitre en chef : L'arbitre doit appliquer les Lois du Jeu au cours d'un match.

Arbitres assistants : Le rôle des arbitres assistants est principalement d'assister l'arbitre en chef, et ils doivent également faire respecter les Lois du Jeu pendant un match. Il y a deux arbitres assistants, un à chaque ligne de touche. De nos jours, des matchs sont joués avec deux autres arbitres assistants sur chaque ligne de but dans certaines compétitions. Depuis la coupe du monde 2018, le revisionnage des actions de matchs par la vidéo, communément désigné par l'acronyme anglais VAR (video assistance review) a été introduit dans toutes les grandes compétitions de football pour améliorer l'arbitrage.

Durée du match : La durée d'un match de football est de 90 minutes, disputées en deux mi-temps de 45 minutes chacune avec une pause de 15 minutes entre les mi-temps. Un temps additionnel est souvent joué à la fin de chaque mi-temps pour compenser le temps perdu par des remplacements, des joueurs blessés nécessitant une attention ou d'autres arrêts. Lors des matchs à élimination directe, en cas de score nul, les équipes jouent en plus 2 mi-temps de 15 minutes chacune (les prolongations), puis toujours en cas d'égalité, jouent des tirs au buts pour déterminer un vainqueur.

Début et reprise du jeu : Un coup d'envoi commence le jeu au début de la première et de la deuxième mi-temps du match ou après qu'un but a été marqué. L'équipe qui commence le jeu est déterminée par un tirage au sort au début du match. Lors du coup d'envoi, seuls deux joueurs de l'équipe qui commence le jeu sont autorisés à se trouver à l'intérieur du cercle central : celui qui donne le coup de pied et celui qui reçoit le ballon.

Ballon en jeu et hors jeu : Le ballon est hors jeu lorsqu'il a entièrement franchi une ligne de but ou une ligne de touche, au sol ou dans les airs. Le ballon reste en jeu à tout autre moment, sauf si le jeu est arrêté par l'arbitre dans des circonstances légitimes.

Méthode de marquage : Un but est marqué si le ballon franchit entièrement la ligne de but, que ce soit au sol ou dans les airs entre les deux poteaux de but et sous la barre transversale, tant qu'aucune infraction aux règles n'a été commise.

Hors-jeu : Un joueur est en position de hors-jeu lorsqu'au moment où son coéquipier lui passe le ballon, il y a moins de deux joueurs (gardiens compris) entre lui et la ligne de but. Un joueur ne peut pas être pris hors-jeu dans sa propre moitié de terrain. Un coup franc est alors accordé à l'adversaire. La figure 2 donne une illustration pour comprendre la règle.

Des fautes sont sifflées si un joueur utilise une force excessive contre un adversaire tout en jouant le jeu, délibérément ou non, ou touche le ballon de la main (à l'exception des gardiens de but dans la surface de réparation). Lorsque cela se produit, en fonction de la gravité de la faute, l'arbitre peut lui adresser un carton jaune pour avertissement ou un carton rouge pour renvoi définitif du terrain. Deux cartons jaunes équivalent à un carton rouge.

Les coups francs sont donnés par l'arbitre après qu'une faute ou une infraction au règlement a été commise. Un coup franc peut être soit "direct" dans lequel un tireur peut marquer directement, soit "indirect", dans lequel un autre joueur doit toucher le ballon avant qu'un but ne puisse être marqué. L'équipe adverse doit être à au moins 9,15 mètres du ballon lorsque le coup franc est exécuté.

Les penaltys sont accordés si un joueur de l'équipe adverse commet une faute à l'intérieur de sa propre surface de réparation. Le coup de pied doit être direct et tiré du point de penalty. Tous

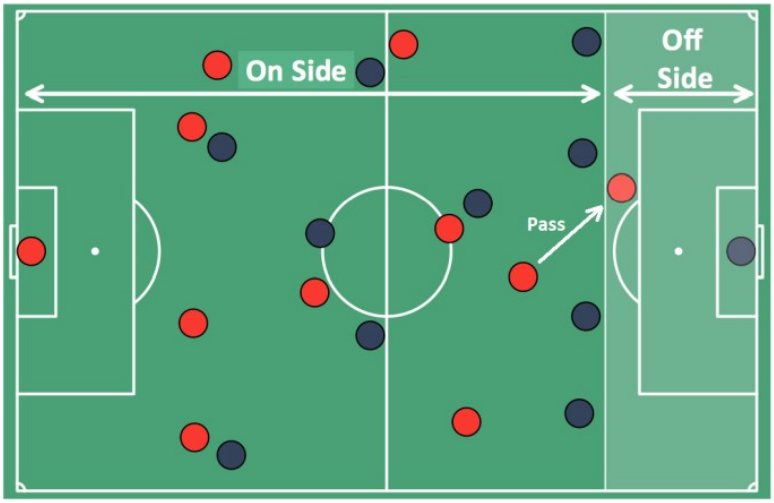


Figure 2 – Illustration d'un hors-jeu

les joueurs (à l'exception du tireur et du gardien de but adverse) doivent être à l'extérieur de la surface de réparation et de l'arc de la surface jusqu'à ce que le penalty soit tiré.

Une remise en jeu ou touche est accordée à une équipe si le ballon a entièrement franchi la ligne de touche que ce soit au sol ou dans les airs. Le ballon est alors donné à l'équipe adverse à l'équipe qui a touché le ballon en dernier. Le lanceur doit utiliser les deux mains, avoir chaque pied soit sur la ligne de touche soit à l'extérieur de la ligne de touche, et lancer le ballon par derrière et au-dessus de sa tête à partir du point où le ballon a quitté le terrain de jeu. Un but ne peut pas être marqué directement à la remise en jeu. Un joueur n'est pas pénalisé pour un hors-jeu lorsqu'il reçoit le ballon directement après une remise en jeu.

Un corner est donné à une équipe qui attaque si le ballon a entièrement franchi la ligne de but (le but lui-même n'est pas pris en compte ici) que ce soit au sol ou dans les airs (sans qu'un but ne soit marqué), et a été touché en dernier par un joueur de l'équipe qui défend. Un coup de pied de coin est alors tiré de l'intérieur de l'arc de coin le plus proche du point où le ballon franchit la ligne de but. Tous les joueurs en défense doivent être à au moins 9,15 mètres de l'arc de coin jusqu'à ce que le coup de pied soit exécuté. Un but peut être marqué directement à partir d'un corner, et un attaquant qui reçoit directement le ballon d'un corner ne peut être pénalisé pour hors-jeu, à la différence d'un coup-franc.

2.3 Football : Statistiques et recrutement de joueurs

Si présentes dans notre quotidien qu'elles sont devenues une banalité pour un suiveur assidu du ballon rond, les statistiques et les données révolutionnent le football moderne depuis la fin des années 1990. Avant cette période, ces technologies n'existaient pas et il fallait alors faire de longs déplacements pour observer un joueur. Aujourd'hui, il est possible juste avec quelques clics, de trouver les meilleurs joueurs correspondant à des critères donnés avec un programme d'analyse de données. Très rapidement, les clubs de football ont trouvé dans ce système une opportunité d'améliorer leurs recrutements, comprenant que ce système était rentable et qu'il est désormais facile de trouver le bon joueur à un coût minime. Branco van den Boomen, footballeur néerlandais évoluant au poste de milieu de terrain au club de Toulouse FC (France) depuis 2020, a affirmé dans une interview quelques semaines après sa signature pour un montant de 350 000€, avoir été recruté grâce à des données statistiques, beaucoup utilisées par le président du club Damien Comolli : « Toulouse a entré dans l'ordinateur le profil de joueur qu'il recherchait, et mon nom est sorti dans la liste ». Lors de la saison de football 2021-2022, il a marqué 12 buts et délivré 20

passes décisives en 37 matchs de ligue 2 BKT (la deuxième division des championnats de France de football), chiffres faisant de lui un joueur décisif presque une fois par match et contribuant largement ainsi à l'ascension de son équipe en première division.

Cependant, les données ne remplaceront pas les professionnels du métier. Le chiffre à lui seul ne parle pas beaucoup, il peut même devenir dangereux si l'analyse qui l'accompagne est trop simpliste. Il est donc important de faire appel à son sens de l'interprétation en matière de football. Dans le cas du recrutement, les départements spécialisés rassemblent des données à ne pas négliger mais le scouting terrain reste l'élément principal quant à la phase finale du processus.

Pour plus de lecture sur les multiples intérêts des statistiques dans l'analyse de performances en football et dans le sport en général, consulter le site www.sportperformanceanalysis.com.

3 Présentation du jeu de données

Nous disposons d'un jeu de données sur la saison de football 2014-2015 de 3003 joueurs professionnels dans 8 championnats de première division de football européens : l'Angleterre, l'Espagne, l'Allemagne, l'Italie, la France, la Russie, les Pays-Bas et la Turquie. Pour des raisons de crédibilité et d'interprétabilité, cette étude prend en compte les footballeurs ayant **joués au moins** 200 **minutes** (soit à peu près 3 matchs en intégralité) de jeu sur la saison 2014-2015. Les gardiens de buts sont exclus de l'étude car ne disposant pas des mêmes données que les joueurs de champs. Les données ont été recueillies sur le site www.whoscorred.com, un site mondialement reconnu pour fournir de la donnée sportive, utilisé par les médias, les clubs de football, les recruteurs, etc.

Différents types de variables sont présentes dans les données et on en dénombre un peu plus d'une centaine au total. Elles seront regroupées en 3 grandes catégories : **les variables de profil des joueurs, les variables de position et les variables de performance**.

3.1 Les variables de profil des joueurs

- **Ligue :** Elle sera notée x_l. Il s'agit de la ligue à laquelle appartient le joueur. Les ligues sont classées en fonction du score officiel de l'UEFA (Union des associations européennes de football) sur la saison 2014-2015 (voir tableau 1).

Pays	Score
Espagne	99.427
Angleterre	80.391
Allemagne	79.415
Italie	70.510
France	52.416
Russie	50.498
Pays-Bas	40.979
Turquie	32.600

Table 1 – Pays et coefficients UEFA 2014-2015

- **Equipe :** Deux variables seront utilisées pour l'analyse : x_{tp} qui représentera le nombre de points acquis par l'équipe sur la saison et x_{tc} qui représentera son coefficient UEFA. Le coefficient UEFA d'une équipe est un score qui permet de la classer par rapport aux autres équipes européennes, toutes ligues nationales confondues, en fonction de ses résultats locaux (dans son pays) et de ses résultats dans les compétitions européennes. Ce classement est mis à jour chaque saison de football et est disponible sur le site de l'UEFA.
- **Nom :** Le nom du joueur ne sera pas pris en compte dans les calculs, mais sera très utile dans les visualisations.

(a) Position-1 (b) Position-2

FIGURE 3 – Illustration des variables de positions

- **Age, Taille, Poids :** L'âge, le poids et la taille du joueur seront traités comme des variables continues.

3.2 Les variables de position

Dans cette partie, les sigles (en notations anglaises) suivants seront utiles :
DC : défenseur central, DL (resp DR) : défenseur latéral gauche ou plus simplement latéral gauche (resp droit), DMC (resp DML et DMR) : milieu défensif central (resp gauche et droit), MC : milieu central, ML (resp MR) : milieu gauche (resp droit), AMC : milieu offensif, AML (resp AMR) : ailier gauche (resp droit), FW : avant centre, FWL (resp FWR) : avant gauche (resp droit).

Deux groupes de variables seront définis :

- **Position 1** : Il s'agit de 15 variables qui indiqueront les proportions relatives de présence sur l'ensemble de ses matchs, du joueur aux différentes positions (voir figure 3.(a)) que l'on ordonnera de la façon suivante : (DC, DL, DR, DMC, DML, DMR, MC, ML, MR, AMC, AML, AMR, FW, FWL, FWR). On les notera Y_{15}. Par exemple, pour le joueur **Lionel Messi**,

$$Y_{15} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0.025, 0, 0, 0.463, 0, 0.512).$$

Cela veut dire que ce joueur a joué 2.5% **de ses matchs** en milieu offensif central, 46.3% en tant qu'avant centre et 51.2% en tant qu'avant droit.

- **Position 2** : Il s'agit de 11 variables binaires qui indiqueront 11 différentes positions connues ou pas pour le joueur en se basant sur les données de la saison de football précédente (2013-1014). On les notera Y_{11}. Dans l'ordre des positions, on aura : (DC, DL, DR, DMC, MC, ML, MR, AMC, AML, AMR, FW), voir figure 3.(b). Dans le cas de Lionel Messi,

$$Y_{11} = (0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1).$$

Cela veut dire qu'en se référant à la saison précédente, Lionel Messi peut jouer en milieu offensif, en ailier droit et en avant centre.

3.3 Les variables de performances

Toutes ces variables sont discrètes et seront regroupées en 4 grandes catégories en fonction de leur sens : les variables de temps, les variables de performances défensive et offensive et les variables de passes de balles. Ces variables sont regroupées dans le tableau de la page suivante :

Temps	Défensives	Offensives	Passes
Apps	**Tacles tentés**	**Tirs**	**Passes**
Mins	•Dribbles subis	*1. Zones*	*1. Taille*
	•Tacles réussis	•Hors surface de réparation (HSR)	•LP-Pré
	Hors-jeu	•Six mètres (6m)	•LP-NPré
	Interceptions	•En Surface de réparation (ESR)	•CP-Pré
	Fautes commises	*2. Situations*	•CP-NPré
	Fautes subies	•Dans le jeu (DJ)	*2. Type*
	Dégagements	•Contre attaque (CA)	•Cr-Pré
	Contres	•Coup de pieds arrêté (CPA)	•Cr-NPré
	•Tirs contrés	•Pénalty (Pen)	•Crn-Pré
	•Centres contrés	*3. Partie du corps*	•Crn-NPré
	•Passes contrées	•Pied droit (PD)	•CF-Pré
		•Pied gauche (PG)	•CF-NPré
		•Tête (Tê)	**Passes clées**
		•Autre (Au)	*1. Taille*
		4. Précision	•Longues
		•Cadrés (Cad)	•Courtes
		•Non cadrés (N-Cad)	*2. Type*
		•Contrés (Con)	•Centres (Cr)
		Buts	•Corner (Crn)
		1. Zones	•Passes en profondeur (PenP)
		•Hors surface de réparation	•Coups-Francs (CF)
		•Six mètres	•Touches (Tou)
		•En Surface de réparation	•Autre
		2. Situations	**Passes décisives (Passes dé)**
		•Dans le jeu	•Cr
		•Contre attaque	•Crn
		•Coup de pieds arrêté	•PenP
		•Penalty	•CF
		3. Partie du corps	•Tou
		•Pied droit	•Autre
		•Pied gauche	
		•Tête	
		•Autre	
		Contrôles ratés	
		Ballons perdus	
		Duels aériens (Duels ae)	
		•Gagnés	
		•Perdus	
		Dribbles tentés	
		•Réussis	
		•Non réussis	

Table 2 – Récap des variables de performances, pour la légende, CP : courte passe, LP : longue passe, Pré : précis, NPré : non précis

Certaines variables comme le nombre de buts disposent de sous-catégories qui donnent des informations plus précises sur le joueur. Pour le nombre de buts, on distingue notamment le nombre de buts marqués par zones de terrain (six mètres, hors surface de réparation, en surface de réparation), ou le nombre de buts par situations de jeu (Dans le jeu, contre-attaque, penalty, coups de pieds arrêtés), ou encore le nombre de buts en fonction de la partie du corps. Comme toutes les variables disposant **d'une ou plusieurs sous-catégories**, le nombre total de buts est égal à la somme des nombres de buts marqués dans chaque sous-catégorie (s.c). Par exemple, le nombre total de tirs est égal à la somme des nombres de tirs à la fois en fonction des zones du terrain (s.c 1), en fonction des parties du corps (s.c 3), en fonction des situations (s.c 2) et enfin en fonction de la précision (s.c 4).

La description des variables de performances est disponible dans le tableau 3 et sa compréhension est fondamentale pour le reste de l'analyse.

Catégorie	Variables	Descriptions
Temps	Apps	Nombre de matchs ou le joueur est monté sur le terrain durant la saison (remplacement compris).
	Mins	Nombre total de minutes jouées au cours de la saison.
Défensives	Dribble subi	Le joueur ne réussit pas à récupérer le ballon après un tacle sur un adversaire, ce dernier a donc conservé la possession du ballon.
	Tacle réussi	Le joueur dépossède un adversaire du ballon d'une façon légale.
	Hors-jeu	Le joueur est pris en position de hors-jeu, ce qui revient à une faute pour l'équipe adverse.
	Interception	Le joueur intercepte une passe de l'équipe adverse.
	Faute	Une manoeuvre illégale d'un joueur qui offre une faute à l'équipe adverse.
	Dégagement	Le joueur annule légalement une attaque de l'équipe adverse et enlève la pression (même pendant un très bref moment) sur son équipe.
	Contre	Le joueur intercepte un tir, un centre ou une passe vers son but d'un adversaire.
Offensives	Tir & But	Le joueur tente d'envoyer le ballon dans la cage adverse, depuis n'importe quelle position du terrain et à travers une partie légale du corps & Il y a but si le ballon franchit entièrement la ligne de la cage adverse.
	But dans le jeu	Un but qui résulte d'une attaque construite par une équipe depuis son camp.
	Contre-attaque	Une tentative rapide de marquer un but juste après annulation d'une attaque adverse.
	Coups de pied-arrêté	Une tentative de marquer un but sur coups de pied arrêté (coup-franc, corner, touche).
	Tir cadré	Un tir non dévié qui, soit aboutit à un but, soit est stoppé par le gardien adverse.
	Contrôle raté	Le joueur réceptionne mal une passe d'un coéquipier.
	Duel aérien	Le joueur gagne ou perd un ballon dans les airs dans un duel direct avec un adversaire.
	Dribble	Le joueur réussit à passer un adversaire sans que ce dernier n'arrive point à lui prendre le ballon (souvent avec un geste technique).
Passes	Longue passe	Une tentative de passe de 25 mètres ou plus, sinon il s'agit d'une passe courte.
	Centre	Une tentative de passe offensive d'une position excentrée vers le centre (généralement vers la surface de réparation adverse).
	Passe clée	Une passe finale qui amène un but ou un tir d'un coéquipier.
	Passe en profondeur	Le joueur passe la balle passe entre deux ou plusieurs joueurs adverses dans leur ligne défensive pour trouver un coéquipier bien démarqué (qui se dirige vers le but adverse).
	Passe décisive	Une passe qui amène directement un but d'un coéquipier.

Table 3 – Description des variables du tableau 2

4 Construction d'une matrice de dissimilarité sur les données des joueurs de football

4.1 Pré-traitement des données

Quatre grandes étapes de pré-traitement ont été effectuées sur les données : la représentation, la transformation, la standardisation et la pondération.

4.1.1 Représentation des variables

→ **Variables de profil**

- Ligue : x_l représente la ligue à laquelle appartient le joueur, ligue qui sera représentée par son coefficient UEFA (voir tableau 1 page 7). Ainsi, pour un joueur qui a évolué dans le championnat allemand, $x_l = 79.415$.
- Equipe :
  - Etant donné que tous les championnats ne contiennent pas le même nombre d'équipes, et donc pas le même nombre de matchs joués, x_{tp} représente le ratio nombre de points obtenus par une équipe/nombre de matchs joués. Par exemple, l'équipe anglaise **Chelsea** a joué 38 matchs pour 87 points, donc $x_{tp} = 2.289$, alors qu'en Allemagne, le **Bayern Munich** a joué 34 matchs pour 79 points, soit $x_{tp} = 2.323$.
  - Pour la variable x_{tc} représentant le coefficient UEFA des équipes, elle sera prise telle quelle. Par exemple, en 2014-2015, le coefficient UEFA de **Chelsea** était de 126 points, celui du **Bayern Munich** de 139 points.
- N.B. Il est à souligner que certains joueurs ont joué pour plusieurs équipes dans la même saison. Dans ce cas, chacune des variables x_l, x_{tp} et x_{tc} pour ces joueurs est calculée en la moyennant par le nombre de minutes jouées dans chaque équipe. Soit un joueur i ayant joué pour n équipes au cours de la saison 2014-2015. Alors on a la moyenne pondérée :

$$x_l^i = \frac{\sum_{j=1}^{n} x_{lj}^i m_j^i}{\sum_{j=1}^{n} m_j^i}$$

  , où x_{lj}^i représente le score UEFA de la j^{eme} ligue où le joueur i a évolué et m_j^i le nombre de minutes jouées dans la j^{eme} ligue. Le même principe s'applique pour x_{tp} et x_{tc}.
- Les variables âge, taille et poids seront représentées telles quelles.

→ **Variables de positions**

Tous joueur de football possède une ou plusieurs positions de préférence. Ainsi, pour tout joueur i de notre jeu de données, Y_{11} n'est pas nulle partout. Cependant, et bien qu'ayant joué plus de 200 minutes durant la saison, il est possible qu'un joueur n'ait jamais été titulaire d'entrée dans le match et soit toujours rentré en cours de jeu. Dans ce cas figure, il n'y a pas d'informations sur les positions occupées par ce type de joueur dans le jeu de données. La variable Y_{15} est donc nulle partout dans ce cas. Cependant, il n'y aurait pas de sens de dire qu'un joueur de football ayant été sur terrain pendant au moins 200 minutes sur une saison, n'a jamais joué à aucun poste, parce que être sur le terrain implique automatiquement d'occuper une position sur celui-ci. Mathématiquement, pour tout joueur, avoir $Y_{15} = 0$ partout est une valeur interdite. Ce problème sera contourné de la façon suivante, qui s'explique facilement par un exemple.

Imaginons un joueur i connu pour jouer aux postes de DMC, MC ou AMC, et ayant joué tous ses matchs de la saison comme remplaçant. Alors la valeur de Y_{15} pour i sera de 0 partout sauf pour les positions DMC, MC, et AMC où il sera de $1/3$ pour chaque. Cette manière de faire est basée sur notre connaissance du football et tient tout son sens dans son interprétation.

→ **Variables de performance**

- Les variables de temps "Apps" et "Mins" seront utilisées telles quelles.
- Les autres variables de performances seront regroupées en deux niveaux. On distinguera les variables de niveau supérieur (VNS) et les variables de niveau inférieur (VNI).

  - **les variables de niveau supérieur (VNS)**

  Elles comptent le nombre total des différents types d'action effectuées par le joueur. Par exemple, le nombre total de tirs est égal à la somme des nombres des tirs en dehors de la surface de réparation, au six mètres et dans surface de réparation. Ces variables seront représentées par 90 minutes de jeu (car un match de football dure 90 minutes).

$$y_{ij} = 90 \times \frac{x_{ij}}{m_i},$$

  où x_{ij} est la j^{eme} VNS du joueur i, m_i son nombre total de minutes jouées. La figure 4 (page 15) résume les distributions de chacune de ces variables. Certaines ont beaucoup de zéros, car elles représentent des actions plutôt rares (buts, passes décisives, etc.) tandis que d'autres (tacles, passes, etc.) ont des distributions plus ou moins gaussiennes.

  - **les variables de niveau inférieur (VNI)**

  Elles comptent le nombre total de différents types d'actions effectuées par le joueur, en fonction des zones du terrain ou des catégories de certaines actions. Ce sont toutes les variables marquées d'un "•" dans le tableau 2 (page 9). Par exemple, pour le nombre de tacles tentés, on distingue deux sous-variables selon qu'un tacle ait été réussi ou non (Dribble subi). Ces variables seront représentées en pourcentage, comme des proportions de la variable de niveau supérieur correspondante. Par exemple, un joueur peut réussir 75% de ses tacles tentés chaque 90 minutes et en perdre 25%. De même, un joueur peut, chaque 90 minutes, effectuer d'une part, 50% de ses tirs du pied droit, 30% du pied gauche, 10% de la tête et 10% d'une autre partie du corps, ou d'autre part, effectuer 90% de ses tirs dans la surface de réparation et 20% en dehors. Il est indispensable de comprendre l'intérêt de cette représentation en pourcentage des VNI.

  En effet, imaginons qu'un joueur effectue 7 tirs toutes les 90 minutes, dont 4 dans la surface de réparation, 2 dans les six mètres et 1 en dehors de la surface. Quand on évalue la dissimilarité entre lui et un autre joueur, deux aspects doivent être pris en compte : la différence du nombre total de tirs par match et les fréquences de tirs en fonction des **zones du terrain**. Ainsi, si on choisissait de représenter ces VNI telles que présentées dans cet exemple, des joueurs avec une grande différence en nombre total de buts (VNS) auraient une grande différence en nombre de buts par zones et la dissimilarité finale serait dominée par la VNS. On perdrait alors l'information des fréquences de tirs par zones du terrain dans le calcul de la dissimilarité.

  Les VNI seront donc des proportions d'un tout, mais certaines VNI seront également vues comme des taux de succès. Par exemple, le nombre de buts marqués du pied droit peut, en plus d'être vu comme proportion du nombre total de buts, être vu comme le taux de succès du nombre total de tirs effectués du pied droit, parce que marquer un but est l'objectif principal d'un tir. Ces deux représentations de cette VNI sont intéressantes pour caractériser un joueur de football et ont chacune son intérêt. Le tableau 4 (page 14) résume toutes les variables de niveau inférieur qui seront également vues comme taux de succès.

  Un autre détail à prendre en compte dans la représentation des VNI concerne les catégories "Autre" des quatre VNS suivantes : nombre de tirs, nombre de buts, nombre de passes clées et nombre de passes décisives. En effet, il peut arriver dans le football que l'on marque un but d'une partie du corps autre que les pieds ou la tête, ou que l'on fasse une passe décisive (et donc une **passe clée**) qui ne soit d'aucun type parmi ceux cités dans le tableau 2. Il est facile de se faire une idée d'un but marqué par une autre partie

VNI	Taux de succès de
Buts par sous-catégorie	Total de tirs par sous-catégorie
Total Buts (VNS)	Total tirs cadrés (VNI)
Total Buts	Total tirs
Tacles/Duels ae/Dribbles (réussis)	Total Tacles/Duels ae/Dribbles
LP (Pré)	Total LP (LP-Pré+LP-NPré)
CP (Pré)	Total CP (CP-Pré+CP-NPré)
Passes par Type (Pré)	Total Passes par type (Pré+NPré)
Total Passes (Pré)	Total de Passes (Pré+NPré)
Passes décisives par type	Passes clées par type
Total passes décisives	Total passes clées

TABLE 4 – Liste des VNI vues comme taux de succès

du corps, mais pour les passes, la catégorie "Autre" est bien plus compliquée à spécifier. Ainsi, dans un souci d'interprétabilité de la mesure de dissimilarité construite sur ces variables, on ne prendra pas en compte les catégories "Autre" des passes clées et des passes décisives dans l'évaluation de la dissimilarité.

En conclusion, les variables de niveau supérieur seront considérées comme un taux sur 90 minutes et les variables de niveau inférieur seront représentées comme des compositions.

Définition : Un vecteur $x = (x_1, ..., x_D)$ est appelé une variable compositionnelle si pour tout $1 \leq i \leq D$, $x_i \geq 0$ et $x_1 + ... + x_D = 1$.

L'information portée par ce type de vecteur est contenu dans les rapports entre ces composantes, et non dans la valeur de chaque composante prise à part.

→ **Traitement des zéros**

Certaines mesures de dissimilarité que nous construirons sur les compositions nécessitent des transformations logarithmiques pour lesquelles la valeur 0 est interdite. Mais il peut arriver qu'une de ces variables vaille 0, ce qui peut être problématique. Ces mesures doivent donc être revues de manière à contourner ce problème, mais nous ne discuterons de la construction de matrice de dissimilarité que dans la prochaine section. Par ailleurs, l'objectif principal de ce traitement des données est que les variables soient représentées d'une manière en adéquation avec l'interprétation qu'on a d'elles dans la réalité footballistique. Ainsi, avoir des composants à zéro dans les compositions peut aussi être problématique pour l'interprétation des performances d'un joueur par rapport à un autre. Par exemple, considérons pour la VNS "nombre de tirs", la catégorie "précision" et supposons deux joueurs dont l'un a effectué 100 tirs pour 0 cadré et l'autre 10 tirs pour 0 cadré. Quand on représente cette catégorie sous forme de composition, la précision de tirs cadrés de ces deux joueurs est de 0%, mais le second joueur a fait beaucoup moins de tirs que le premier. Avec 90 tirs en plus, il a des chances non négligeables d'avoir au moins 1 tir cadré et donc sa composante "cadré" non nulle dans la représentation de sa précision de tir. Nous allons donc proposé une approche qui permet de remplacer les zéros dans une composition par une petite valeur non nulle ϵ tout en respectant les ratios entre les autres composantes non nulles et en essayant de prendre en compte au mieux les rapports d'un joueur à l'autre.

-L'approche multiplicative bayésienne (AMB) :

Cette méthode a été introduite par Pepus Daunis-i-Estadella et al. dans leur article publiée en 2008 : "Bayesian tools for count zeros in compositional data" (voir référence 1).

Soit Ω un ensemble partitionné de façon exhaustive en k catégories ω_j. On suppose que toute observation dans Ω est incluse dans une et une seule catégorie ω_j de Ω et on tire indépendamment N observations dans Ω. On note n_j le nombre total d'observations dans

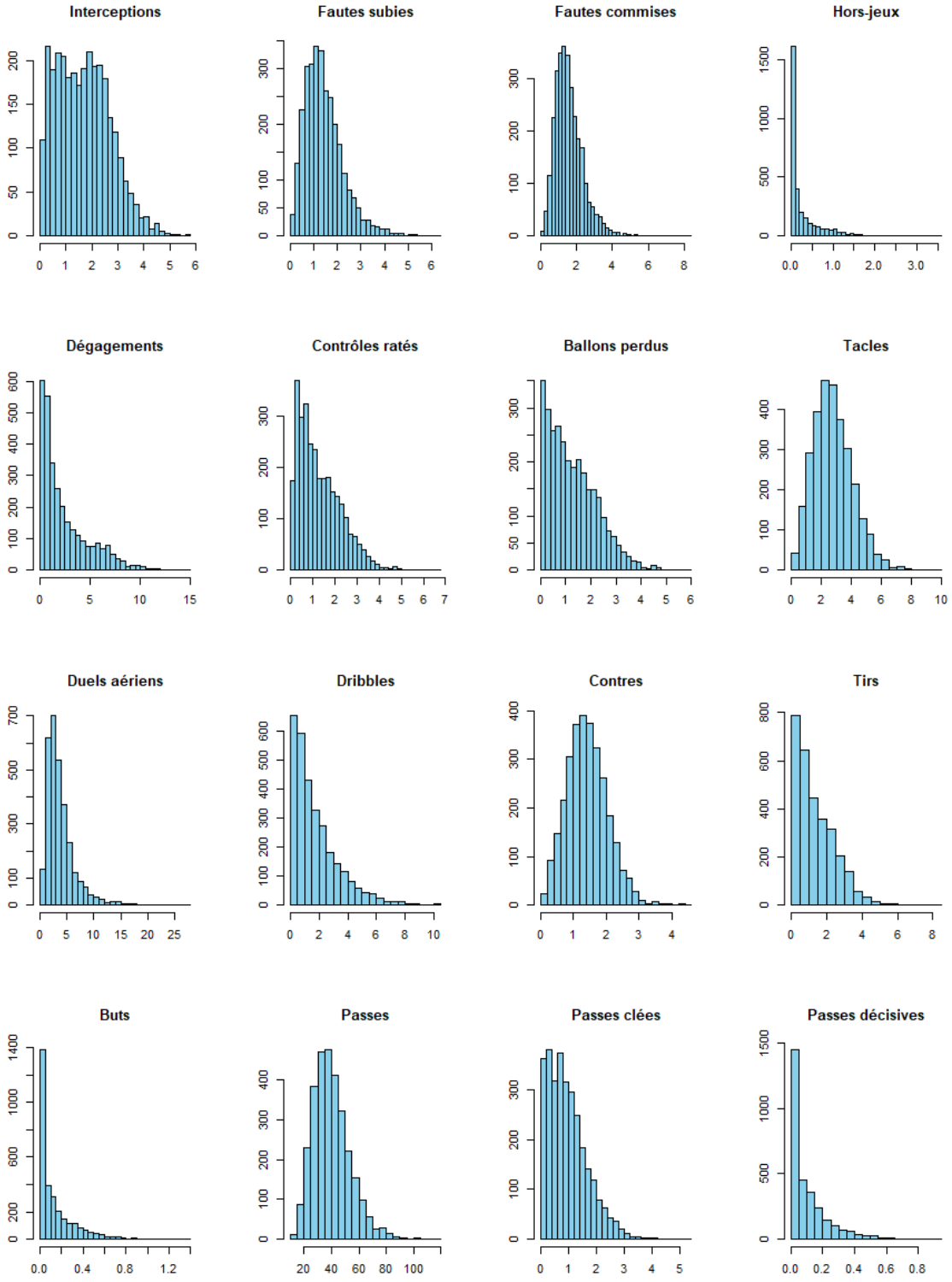


FIGURE 4 – Histogramme des VNS

la catégorie ω_j, alors $n = (n_1, ..., n_k)$ suit une distribution multinomiale de paramètre $\theta_j = \mathbb{P}(\omega_j)$, $1 \leq j \leq k$.
Dans une approche bayésienne, on suppose que $\theta = (\theta_1, ..., \theta_k)$ possède une loi à priori de Dirichlet (conjugué de la loi multinomiale) de paramètre st, avec s le poids total de la loi à priori et t son espérance : $t = (t_1, ..., t_k)$, $t_j > 0$ (cette condition force l'estimation à priori de chaque θ_j à être différent de 0) pour tout j et $\sum_{j=1}^{k} t_j = 1$. On note $\alpha_j = st_j$ le poids à priori de la catégorie ω_j. L'espérance à priori de θ_j est t_j.
D'après le théorème de Bayes, la loi à posteriori de θ s'écrit en fonction de sa loi à priori

$$p(\theta) \propto \prod_{j=1}^{k} \theta_j^{st_j}$$

et de sa vraisemblance sachant une réalisation x :

$$\mathcal{L}(\theta|x) \propto \prod_{j=1}^{k} \theta_j^{x_j}$$

. Il en découle aisément que la loi à posteriori de θ est une loi de Dirichlet de paramètre $x + st$. Ainsi, l'espérance à posteriori de θ_j est :

$$\mathbb{E}(\theta_j|x) = \frac{x_j + st_j}{N + s} = \frac{x_j + \alpha_j}{N + s}$$

. Pour un N fixé, différentes lois à priori de Dirichlet ont été proposées avec l'estimation de θ par l'espérance de la loi à posteriori correspondante :

loi à priori	s	α_j	estimation $\hat{\theta}_j$
Haldane	0	0	$\frac{x_j}{N}$
Perks	1	$1/k$	$\frac{x_j+1/k}{N+1}$
Jeffreys	$k/2$	$1/2$	$\frac{x_j+1/2}{N+k/2}$
Bayes-Laplace	k	1	$\frac{x_j+1}{N+k}$

TABLE 5 – Quelques lois à priori classique

Par exemple, prenons $N = 5$ pour le nombre de tirs par match d'un joueur donné, $k = 3$ (catégorie précision) et $x = (3, 0, 2)$ resp pour tirs cadrés, non cadrés et contrés. Alors en choisissant la loi à priori de Haldane, on a l'estimation à posteriori de $\theta = (\frac{3}{5}, 0, \frac{2}{5})$, ce qui revient en fait à une approche fréquentiste. En choisissant plutôt Bayes-Laplace, on obtient $\theta = (\frac{4}{8}, \frac{1}{8}, \frac{3}{8})$. La valeur nulle est donc légèrement modifiée dans la composition.
Cependant, nous cherchons à préserver les ratios entre les composantes non nulles et cette estimation à posteriori de θ par l'espérance ne permet pas de respecter cette condition :

$$(\frac{3}{2} \neq \frac{4}{3}).$$

Pour contourner ce problème, nous avons utilisée la formule suivante proposée par Martin-Fernandez et al. dans leur article publiée en 2003 : "Dealing with zeros and missing values in compositional data sets using nonparametric imputation." (voir référence 2) où les composantes nulles sont remplacées par l'espérance de la loi à posteriori et les composantes non nulles par une approche multiplicative. La formule proposée est la suivante :

$$\begin{cases} \theta_j = \frac{\alpha_j}{N+s} \text{ si } x_j = 0 \\ \theta_j = \frac{x_j}{N}(1 - \sum_{j,x_j=0} \frac{\alpha_j}{N+s}) \text{ si } x_j > 0 \end{cases} \quad (1)$$

Dans l'exemple précédent, l'estimation de θ donne :

$$(\frac{3}{5}(1-\frac{1}{8}), \frac{1}{8}, \frac{2}{5}(1-\frac{1}{8})) = (\frac{21}{40}, \frac{5}{40}, \frac{14}{40}).$$

On peut vérifier que les proportions entre les valeurs non nulles sont respectées cette fois :

$$(\frac{3}{2} = \frac{21}{14}).$$

-La loi à priori d'Akhanli :
Il est important de rappeler que l'utilisation de l'AMB est motivée par la volonté d'ajuster les VNI, qui décrivent par un pourcentage la qualité d'un joueur dans une certaine action de jeu, et non par la volonté d'appliquer des transformations logarithmiques. Ainsi, le choix de la loi à priori doit être fait de sorte que les ajustements des compositions soient interprétables en pratique d'un point de vue footballistique, comme décrit dans le contexte de la méthode. C'est ainsi que Serhat Akhanli dans ses recherches, a proposé une loi à priori de Dirichlet avec un choix de paramètres s et t adapté au jeu de données.
La valeur de s est fixée à 1. Le paramètre le plus important ici est t, qui représente l'espérance de la loi à priori, et donc l'espérance à priori estimée de la proportion θ_j de la VNI j. Une estimation bien connue de l'espérance est la moyenne, et utiliser la moyenne sur tous les joueurs (dont la VNI j est non nulle) peut être un bon choix de loi à priori pour aller déterminer l'estimation à posteriori de θ_j, et est plus cohérent dans les interprétations. Ainsi, la loi à priori proposée par Serhat est donnée par les paramètres :

$$s = 1, \quad \alpha_j = \frac{1}{n}\sum_{i=1}^{n}\frac{c_{ij}}{N_i}$$

où c_{ij} est la j^{eme} VNI du joueur i (en nombre et non en pourcentage), N_i sa VNS correspondante et n le nombre total de joueur dans le jeu de données. Le ratio $\frac{c_iJ}{N_i}$ donne le pourcentage de l'action décrite par la VNI j effectué par le joueur i.
Considérons l'exemple suivant sur la variable nombre de tirs avec la catégorie précision, sur un groupe de 3 joueurs avec des valeurs choisies uniquement à titre illustratif. Les valeurs de la VNS et des VNI sont données dans le tableau suivant :

Joueurs	nombre de tirs	cadrés	non cadrés	contrés
1	100	0	80	20
2	50	10	30	10
3	10	0	7	3

Table 6 – Exemple illustratif d'un jeu de données de 3 joueurs

On veut appliquer l'AMB pour remplacer les zéros de la VNI "nombre de tirs cadrés". Les paramètres de la loi à priori de Serhat sont donnés par :

$$s = 1, \quad \alpha = \frac{1}{3}(\frac{0}{100} + \frac{10}{50} + \frac{0}{10}) = \frac{1}{15}$$

.
Le tableau suivant résume les estimations à posteriori obtenues pour la composition de la précision de tirs avec les différentes lois à priori :

Loi à priori	Joueurs	tirs cadrés	tirs non cadrés	tirs contrés
Haldane	1	0	0.8	0.2
	2	0.2	0.6	0.2
	3	0	0.7	0.3
Perks	1	0.083	0.733	0.184
	2	0.2	0.6	0.2
	3	0.083	0.641	0.275
Jeffreys	1	0.166	0.666	0.167
	2	0.2	0.6	0.2
	3	0.167	0.583	0.25
Bayes-Laplace	1	0.5	0.4	0.1
	2	0.2	0.6	0.2
	3	0.5	0.35	0.15
Serhat	1	0.017	0.786	0.197
	2	0.2	0.6	0.2
	3	0.017	0.688	0.295

TABLE 7 – Résultats sur les joueurs 1, 2 et 3 pour les lois à priori

On observe qu'avec la loi à priori de Bayes-Laplace, l'estimation à posteriori des précisions de tirs dégénère totalement et ne réflète pas du tout la réalité des joueurs 1 et 3. Celle Jeffreys accorde plus de 16% de précision de tirs cadrés à des joueurs qui n'ont cadré aucun tir, ce qui est aussi assez loin de la réalité. Perks semble plus cohérent que les deux précédents, mais les pourcentages de tirs cadrés calculés pour les joueurs 1 et 3 restent tout de même élevés (8%). La loi à priori de Serhat donne la meilleure estimation à posteriori avec seulement 1.6% de tirs cadrés pour les joueurs 1 et 3, ce qui est bien plus proche de la réalité. De plus, les proportions dans ce cas sont les plus proches des fréquences observées (Haldane). C'est cette loi qui sera donc utilisée pour traiter les valeurs nulles dans les compositions par approche multiplicative bayésienne.

→ **Les zéros essentiels**

Lorsque qu'une variable de niveau supérieur vaut 0, il serait imprudent d'assigner une valeur nulle à toutes les variables de niveau inférieur correspondantes, car nous n'avons dans ce cas de figure aucune information sur comment performent les joueurs concernés dans les catégories d'actions décrites par ces VNI. La dissimilarité pour cette variable entre un joueur j dans ce cas de figure et un joueur i quelconque sera donc construite en considérant uniquement la variable de niveau supérieur, pondérée par la somme des poids des VNI correspondantes. La formule suivante résume le procédé :

$$d(x_i, x_j) = \begin{cases} d(x_i, x_j) \sum_{t=1}^{D} \omega_t \text{ si } \sum_{t=1}^{D} x_{it} = 0 \\ d(x_{it}\omega_t, x_{jt}\omega_t) \text{ sinon} \end{cases} \quad (2)$$

où x_i est la VNS en question, x_{it} est sa t^{eme} VNI et ω_t le poids de cette VNI (voir section 4.1.4 pour plus de détails sur la pondération).

4.1.2 Transformation des VNS

Pour construire une mesure de dissimilarité réflectant au mieux la réalité, il est parfois indispensable de transformer certaines variables. Les distributions de certaines VNS sont plus ou moins asymétriques (voir figure 4 page 15). Par exemple, très peu de joueurs (les avants) sont responsables de la majorité des tirs et très peu de défenseurs sont responsables de la majorité des dégagements. Cela veut dire qu'il peut y avoir de grands écarts en valeur absolue entre des

joueurs qui tirent ou qui dégagent beaucoup, alors que parallèlement, les écarts entre ceux qui tirent ou qui dégagent moins souvent seraient faibles. Par exemple, si deux joueurs A et B tirent respectivement 10 et 7 fois toutes les 90 minutes, et que deux autres C et D tirent $0,4$ et $0,1$ fois toutes les 90 minutes, alors l'écart entre A et B est 10 fois plus grand que l'écart entre C et D. Or cela n'a pas vraiment de sens d'un point de vue footballistique. En effet, la différence entre deux joueurs qui tirent souvent comparée à celle entre deux joueurs qui ne tirent presque jamais est plus ou moins la même. A et B peuvent être regroupés dans une classe de tireurs habituels, et sont dans cette classe aussi proche que le sont C et D dans leur classe de rares tireurs.

Cela suggère donc qu'il faut appliquer sur les VNS une transformation qui rend plus ou moins proportionnels les écarts entre les grandes valeurs par rapport aux écarts entre les petites valeurs : le logarithme ou la racine carrée sont deux bons candidats. Cependant, un problème avec la fonction logarithme est qu'elle n'est pas définie en 0 alors que des VNS valent 0 pour certains joueurs. Il faudrait alors rajouter une constante c à ces VNS afin de pallier ce problème. Toutes les VNS n'ayant pas les mêmes significations, la constante ne sera pas la même pour toute.

Pour choisir les constantes, 9 joueurs populaires (3 défenseurs, 3 milieux de terrain, 3 avants) en se basant sur les informations de la saison 2014-2015 ont été sélectionnées. Il s'agit de : David Alaba (D.A), Sergio Ramos (S.R), Gérard Piqué (G.P), Andres Iniesta (A.I), Paul Pogba (P.P), James Rodriguez (J.R), Lionel Messi (L.M), Cristiano Ronaldo (C.R), Neymar (Ney.). Notre connaissance des joueurs de football et notamment des profils de ces joueurs particuliers jouera un grand rôle dans l'interprétation des résultats. L'idée est de comparer pour chaque VNS, après avoir appliquer une transformation du type $\log(x+c)$ ou $\sqrt{x+c}$ et après centrage et standardisation (voir section 4.1.3), les écarts entre les joueurs ayant de grandes valeurs par rapport à celle ayant des petites valeurs, pour différentes valeurs de la constante c. Nous choisirons ensuite une valeur de c pour laquelle la transformation rétrécit au mieux les écarts entre les grandes valeurs par rapport à ceux entre les petites, et pour laquelle la comparaison des écarts entre ces 9 joueurs reflète au mieux la réalité, en se basant sur notre connaissance des profils de ces derniers.

Un autre critère de sélection de la valeur de c est de regarder la forme des distributions (histogrammes) des VNS transformées. En appliquant des transformations du type logarithme ou racine carrée, les grandes valeurs se rapprochent des petites, ce qui tend à rendre les histogrammes asymétriques de certaines VNS plus ou moins normaux. Plus la distribution d'une VNS transformée tendra à être gaussienne pour une valeur de c, plus cette valeur sera un bon choix.

La figure suivante est une illustration des critères présentés ci-dessus sur la VNS nombre de tirs. C'est une capture d'écran d'une application web $R\ Shiny$ développée pour illustrer les choix des constantes c. Dans le rectangle en haut à droite, nous avons tracé les écarts entre les 9 joueurs cités plus haut, lorsqu'on n'applique aucune transformation (none), une transformation racine carrée (sqrt) puis une transformation logarithmique (log). La valeur de c prise est 0.42. L'histogramme de la transformation logarithmique montre une courbe plus ou moins en cloche, ce qui nous fait croire que cette valeur de c répond au besoin. Par ailleurs, quand on regarde le rectangle des écarts, on peut voir d'une part que le logarithme affecte plus les grandes valeurs que la racine carrée, et d'autre part que les écarts entre les grandes valeurs sont en général bien plus proche des écarts entre les petites valeurs (par comparaison avec la ligne none). De plus, quand on s'y connaît bien au football, on sait par exemple qu'en terme de nombre de tirs par match (un match de foot dure 90 minutes), la différence entre Cristiano Ronaldo (C.R) et Lionel Messi (L.M) est assez proche, voir même plus petit que celle entre Sergio Ramos (S.R) et David Alaba (D.A). Ainsi, pour cette VNS, une transformation logarithmique avec $c = 0.42$ peut être un bon choix, même si cette valeur n'est sûrement pas la seule.

Pour finir, il est important de noter que les VNS nombre de buts et nombre de passes décisives

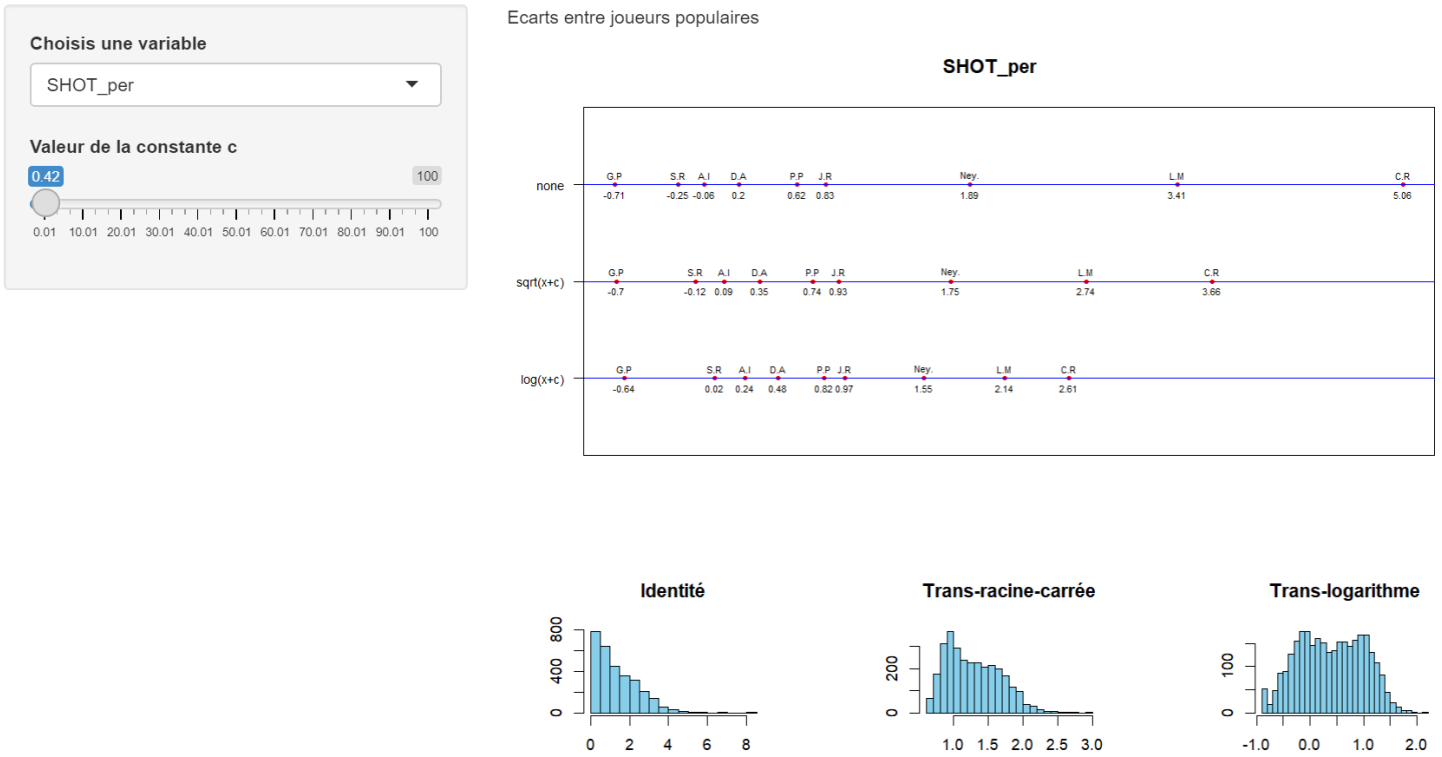


FIGURE 5 – Transformation de la VNS nombre de tirs

ne sont pas concernées par cette étape de pré-traitement de données. En effet, le football est gouverné par le but et une passe décisive implique directement un but, donc ces variables ne devraient pas être transformées en raison de leur impact direct sur les résultats d'un match de foot.

4.1.3 Standardisation

Contrairement à la transformation qui permet d'avoir des valeurs comparables entre elles au sein d'une même variable, la standardisation permet de comparer les variables entre elles.

→ **Les VNS** seront standardisées par l'écart absolu médian (EAM) qui représente la moyenne des écarts absolus à la médiane. L'EAM se calcule avec la formule :

$$EAM(x_j) = \frac{1}{n}\sum_{i=1}^{n} |x_{ij} - med(x_j)|$$

où x_j désigne la variable j, x_{ij} la valeur du joueur i pour la variable j, n le nombre total de joueurs et $med(x_j)$ la médiane de la variable j. La raison de ce choix est que la valeur absolue permet de bien prendre en compte les écarts entre les joueurs (voir section 4.2 sur la construction de la dissimilarité) et est moins sensible aux valeurs extrêmes. En effet, la plupart des variables n'ont pas des maximums naturels. Ces valeurs peuvent beaucoup varier d'une année à une autre, de ce fait, il est important de réduire leur impact dans l'étude si on espère l'adapter à des données d'une autre saison de football. Par exemple, sur les données de 2014-2015, on observe un maximum de 1.38 pour le nombre de buts par 90 minutes, ce qui témoigne d'une saison très prolifique en terme de buts marqués pour un certain joueur, mais ce chiffre peut considérablement baisser/augmenter pour une autre saison. C'est pour cette même raison que les méthodes de standardisation min-max sont écartées au détriment des méthodes qui prennent en compte les variations globales des variables. De plus, nous choisissons la médiane car elle minimise l'EAM.

→ **Les variables âge, poids, taille, mins et apps** seront également standardisées par l'EAM.

→ **Pour ce qui est des VNI**, la méthode est un peu différente. Toutes les variables d'une même composition seront standardisées par la même quantité qui est la moyenne des écarts absolus médians (EAM_s) des variables de cette composition. On a

$$EAM_s = \frac{1}{k}\sum_{j=1}^{k} EAM_j$$

, avec EAM_j l'écart absolu médian de la j^{eme} variable de la composition et k le nombre de variable de cette composition. Par exemple, pour les pourcentages de buts par partie du corps, les variables pied droit, pied gauche, tête et autre seront toutes standardisées par la moyenne des écarts absolus médians de toutes les 4. La raison de ce choix de standardisation est qu'une certaine différence de pourcentage entre deux joueurs a la même signification dans chaque catégorie d'une composition, indépendamment des écarts absolus médians des variables représentant chaque catégorie. Plus formellement, on veut que la propriété suivante soit vérifiée par chaque composition de VNI standardisées :

Proposition : Soit $x = (x_1, ..., x_D)$ une composition, $D \geq 2$, $x^1 = (x_1 + \epsilon, x_2 - \frac{\epsilon}{D-1}, ..., x_D - \frac{\epsilon}{D-1})$ et $x^q = (x_1 - \frac{\epsilon}{D-1}, ..., x_{q-1} - \frac{\epsilon}{D-1}, x_q + \epsilon, x_{q+1} - \frac{\epsilon}{D-1}, ..., x_D - \frac{\epsilon}{D-1})$, $1 \leq q \leq D$, avec $\epsilon > 0$ choisi de sorte que x^1 et x^q soit encore des compositions. Notons $s_1 > 0$, ..., $s_D > 0$ des constantes de standardisation respectives pour les variables x_1, ..., x_D. Considérons la distance de Manhattan (ou distance en norme 1) standardisée

$$d_M(x, y) = \sum_{j=1}^{D} |\frac{x_j}{s_j} - \frac{y_j}{s_j}|$$

. Alors pour tout $1 \leq q \leq D$, $d_M(x, x^1) = d_M(x, x^q)$ si et seulement si $s_l = s$, pour tout $1 \leq l \leq D$, avec s une constante donnée.

Preuve :

$$d_M(x, x^1) - d_M(x, x^q) = \sum_{j=1}^{D} |\frac{x_j}{s_j} - \frac{x_j^1}{s_j}| - \sum_{j=1}^{D} |\frac{x_j}{s_j} - \frac{x_j^q}{s_j}| \tag{3}$$

$$= (\frac{\epsilon}{s_1} + \frac{\epsilon}{D-1}\sum_{l=2}^{D}\frac{1}{s_l}) - (\frac{\epsilon}{s_q} + \frac{\epsilon}{D-1}\sum_{l=1, l\neq q}^{D}\frac{1}{s_l}) \tag{4}$$

$$= \epsilon\{(\frac{1}{s_1} - \frac{1}{s_q}) + \frac{1}{D-1}(\frac{1}{s_q} - \frac{1}{s_1})\} \tag{5}$$

$$= \epsilon(\frac{1}{s_1} - \frac{1}{s_q})(1 - \frac{1}{D-1}) \tag{6}$$

$$= 0 \iff s_1 = s_q \qquad [cqfd]. \tag{7}$$

Comme annoncé dans la section 4.1.1 (-**les variables de niveau inférieur**), les variables "autre" des compositions "type de passes clées" et "passes décisives" ne seront pas prises en compte dans la moyenne des écarts absolus médians de ces compositions.

→ **Les variables de positions, d'équipe et de ligue** ne sont pas prises en compte dans cette étape. Les traitements de ces variables seront discutés respectivement dans les sections 4.2.3 et 4.2.4 .

4.1.4 Pondération

La pondération est le concept de multiplication des variables par différentes constantes. Certaines variables peuvent être plus importantes que d'autres, et la pondération permet d'en tenir compte. En effet, l'attribution d'un poids différent à chaque variable conduit à ce que certaines

variables soient plus dominantes que d'autres dans la construction de dissimilarité, et par conséquent influence la signification des résultats.

Toutes les variables de niveau supérieur ne contenant aucune sous-catégorie (composition), les variables de temps, ainsi que l'âge, le poids et la taille des joueurs se verront attribuer un poids d'une unité (1).

Pour les variables de niveau supérieur n'ayant qu'une seule composition (tacles, contres, duels aériens, dribbles), chaque variable de la composition représente un aspect de l'action comptée par la VNS. Ainsi, Le même poids sera attribué à la VNS qu'à l'ensemble de la composition. Le poids de la VNS sera d'une unité et par conséquent chaque VNI de la composition se verra attribuer un poids de $1/D$ avec D le cardinal de la composition. Ces variables auront donc un poids total de 2 unité chacune.

Un autre cas est celui des VNS possédant plus d'une composition (Tirs, Buts, passes et passes clées). Certaines VNI des VNS dans ce cas sont représentées à la fois comme proportion d'un total (composition) et comme taux de succès d'une autre variable (voir tableau 4 page 14). Les pondérations appliquées à ces variables de niveau inférieur sont résumées dans les tableaux suivants (se référer au tableau 2 page 9 pour les abréviations) :

Catégories	Représentation	Poids des VNI				Poids total
Zones		HSR	6m	ESR		
	pro-Tirs	1/3	1/3	1/3		1
	pro-Buts	1/6	1/6	1/6		1/2
	suc-Buts	1/12	1/12	1/12		1/4
Situations		DJ	CA	CPA	Pen	
	pro-Tirs	1/4	1/4	1/4	1/4	1
	pro-Buts	1/8	1/8	1/8	1/8	1/2
	suc-Buts	1/16	1/16	1/16	1/16	1/4
Partie du corps		PD	PG	Tê	Au	
	pro-Tirs	1/4	1/4	1/4	1/4	1
	pro-Buts	1/8	1/8	1/8	1/8	1/2
	suc-Buts	1/16	1/16	1/16	1/16	1/4
Précision		Cad	N-Cad	Con		
	pro-Tirs	1/3	1/3	1/3		1
	suc-Buts 1	3/8				3/8
	suc-Buts 2	3/8				3/8

Table 8 – Pondération des VNI pour les VNS nombre de tirs et nombre de buts ; pro-Tirs (resp Buts) est la proportion du nombre de tirs (resp Buts) par catégorie, suc-Buts est le taux de succès nombre de Buts/nombre de Tirs par catégorie, suc-Buts 1 est le taux global de succès toute catégorie comprise nombre total de buts/ nombre total de tirs, suc-Buts 2 est le taux global de succès toute catégorie comprise nombre total de buts/ nombre total de tirs cadrés

Catégories	Représentation	Poids des VNI			Poids total
Taille		Longue	Courtes		
	pro-Passes	1/2	1/2		1
	pré-Passes	1/4	1/4		1/2
Type		CF	Cr	Crn	
	pro-Passes	1/3	1/3	1/3	1
	pré-Passes	1/6	1/6	1/6	1/2
Taux global de passes réussies		1			1

Table 9 – Pondération des VNI pour la VNS nombre de passes ; pro-Passes est la proportion du nombre de passes par catégorie, pré-Passes le taux de réussite de passes par catégories

Catégories	Représentation	Poids des VNI					Poids total
Taille	pro-Passes	Longue			Courtes		1
		1/2			1/2		
Type		CF	Cr	Crn	PenP	Tou	
	pro-Passes clées	1/5	1/5	1/5	1/5	1/5	1
	pro-Passes dé	1/10	1/10	1/10	1/10	1/10	1/2
	suc-Passes dé	1/20	1/20	1/20	1/20	1/20	1/4
Taux global de passes-dé réussies		1/4					1/4

TABLE 10 – Pondération des VNI pour les VNS nombre de passes clées et nombre de passes décisives ; suc-Passes dé est le taux de succès nombre de passes décisives/nombre de passes clées par catégories

Nous avons établi les poids des VNI pour toutes les VNS possédant plus d'une catégorie. La question maintenant est de savoir **quels poids attribuer à ces VNS elles-mêmes** : doit-on attribuer à chacune d'elle un poids d'une unité (1) comme pour les VNS ayant au plus une seule sous-catégorie, ou doit-on à chacune d'elle attribuer un poids égal au nombre de sous-catégorie qu'elle contient (par ex 4 pour le nombre tirs) ? Dans le premier cas, deux joueurs avec des nombres de buts assez éloignés serait assez proches s'ils ont des compositions assez similaires pour les VNI, tandis que dans le second cas, deux joueurs avec des nombres de buts assez proches serait assez proches, bien qu'ils aient des compositions très différentes. Une bonne manière d'équilibrer les choses serait de moyenner les deux possibilités. Ainsi, la VNS nombre de tirs se verrait attribuer un poids de $(1 + 4)/2 = 2.5$ par exemple.

Le tableau suivant résume les poids totaux accordés à chaque variable, toute représentation (niveau supérieur/inférieur) comprise :

Variable	Poids			Total
	VNS	VNI	Néant	
Age	-	-	1	1
Poids	-	-	1	1
Taille	-	-	1	1
Mins	-	-	1	1
Apps	-	-	1	1
Hors-jeu	1	-	-	1
Interception	1	-	-	1
Fautes commises	1	-	-	1
Fautes subies	1	-	-	1
Dégagements	1	-	-	1
Contrôles ratés	1	-	-	1
Ballons perdus	1	-	-	1
Tacles	1	1	-	2
Contres	1	1	-	2
Duels aériens	1	1	-	2
Dribbles	1	1	-	2
Tirs	2.5	4	-	6.5
Buts	2	3	-	5
Passes	2.5	4	-	6.5
Passes clées	1.5	2	-	3.5
Passes décisives	1	1	-	2

TABLE 11 – Pondération totale par variable

Notons enfin que toutes les pondérations dans ces tableaux sont subjectives car basées sur

notre connaissance personnelle du football et peuvent varier en fonction des applications. Ainsi, un recruteur qui serait par exemple intéressé par des joueurs similaires à un défenseur donné peut décider d'augmenter les poids des variables défensives comme il le souhaite. Une application web $R\ Shiny$ offrant la possibilité de choisir la pondération a été développée à cet effet.

4.2 Conception d'une mesure de dissimilarité sur les classes de variables

Le but de cette partie est de construire une mesure de dissimilarité entre les joueurs pour chacune des 5 classes de variables que sont : les variables de niveau supérieur (VNS), les variables de niveau inférieur (VNI), les variables de ligue et d'équipe, les variables de **position 1**, les variables de **position** 2 et les autres variables quantitatives (âge, poids, taille, apps et mins). Notre principale motivation sera de construire à chaque fois une "dissimilarité interprétable", i.e qui reflète au mieux les différences entre les joueurs dans la vie réelle. Nous agrègerons ensuite les cinq mesures de dissimilarité pour avoir une mesure finale interprétable.

4.2.1 VNS et autres variables quantitative

Chacune des VNS compte un taux par 90 minutes d'une action déterminée du jeu de football, et afin de bien prendre en compte les écarts entre les joueurs pour chacune de ces variables, nous utiliserons la distance de Manhattan, qui n'est rien d'autre que la somme des écarts en valeur absolue, pour comparer les 16 VNS entre deux joueurs quelconques. Cette même distance sera utilisé pour les autres variables quantitatives.

Définition : La distance de Manhattan entre deux vecteurs $x = (x_1, ..., x_k)$ et $y = (y_1, ..., y_k)$ est donnée par

$$d_M(x, y) = \sum_{j=1}^{k} |x_j - y_j|.$$

4.2.2 VNI

La distance de Manhattan sera utilisée pour mesurer la dissimilarité entre deux joueurs quelconque sur l'ensemble des variables de niveau inférieur. Elle permet de traiter de la même manière les différences entre les pourcentages, indépendamment de la valeur des pourcentages où ces différences se produisent.

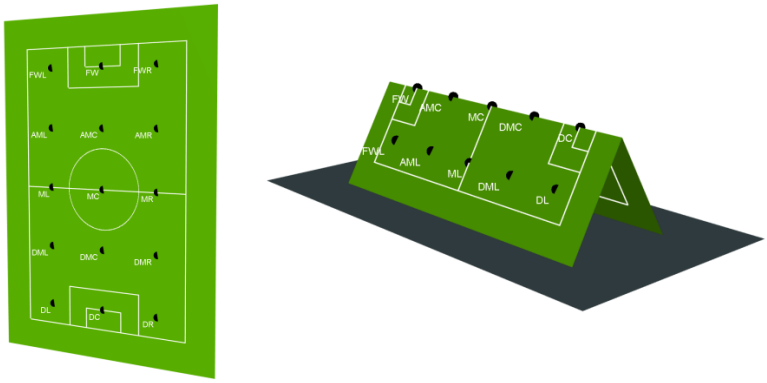
4.2.3 Variables d'équipe et de ligue

La variable de ligue x_l et les variables d'équipe x_{tp} et x_{tc} ont été introduites dans la section 3.1. Pour rappel, pour un joueur donné, x_l représente le score de la ligue à laquelle il appartient, x_{tp} le nombre de points acquis par son équipe lors de la saison et x_{tc} le coefficient UEFA de son équipe. Il semble approprié d'adopter la distance de Manhattan pour ces variables, mais les variables x_l et x_{tp} sont reliées entre elles, car le nombre de point qu'obtient une équipe dépend fortement du niveau de la ligue dans laquelle il joue. Plus ce niveau est élevé (x_l grand), plus il est difficile pour une équipe qui joue dans cette ligue d'obtenir des points. En termes de performances d'équipe, considérer x_l et x_{tp} séparément pourrait éloigner deux équipes similaires l'une de l'autre.

Ainsi, pour évaluer une dissimilarité interprétable entre deux joueurs i et j sur les variables x_l et x_{tp}, Serhat A. propose la formule suivante, qui incorpore les deux variables dans une même mesure de différence :

$$d_1(i, j) = |\frac{x_{il} - x_{jl}}{s_l} + \frac{x_{itp} - x_{jtp}}{s_{tp}}|.$$

où x_{il} est le score de la ligue à laquelle appartient le joueur i, s_l (resp s_{tp}) l'écart absolu médian de la variable x_l (resp x_{tp}), qui sert de constante de normalisation afin que les variables soient comparables entre elles.



(a) 2 dimensions (b) 3 dimensions

FIGURE 6 – Représentation du terrain de football : 2D vs 3D

Par ailleurs, nombre d'équipes ne jouent pas en compétition européenne (seuls les meilleurs y sont qualifiées), et de ce fait, ne disposent pas de score UEFA. Il y a donc beaucoup de valeurs manquantes pour la variable x_{tc}. Mais x_{tc} peut être interprétée comme une combinaison entre x_l et x_{tp} basée sur les matchs ayant lieu entre les équipes jouant en compétitions européennes. Ainsi, pour évaluer la dissimilarité finale pour les variables de ligue et d'équipe entre deux joueurs i et j dans le cas de la non existence de la valeur de x_{tc} pour i ou j, Serhat A. propose de pondérer à la hausse la dissimilarité des variables x_l et x_{tp} par rapport à celle de x_{tc}. La formule est la suivante :

$$d(i,j) = \begin{cases} d_1(i,j) \text{ si } x_{itc} \text{ ou } x_{jtc} \text{ est manquante} \\ \frac{2}{3}d_1(i,j) + \frac{1}{3}d_2(i,j) \text{sinon} \end{cases} \tag{8}$$

avec $d_2(i,j) = |\frac{x_{itc} - x_{jtc}}{s_{tc}}|$, s_{tc} l'écart absolu médian de x_{tc}.

4.2.4 Variables de position

→ **Variables de position 1, Y_{15} (voir section 3.2) :** La figure 3.a (page 8) illustre les 15 positions prises en compte par ces variables. Pour évaluer la distance de positionnement entre deux joueurs, deux aspects seront pris en compte : le côté (L pour gauche, R pour droit et C pour centre) et la position (D pour défenseur, DM pour milieu défensif, M pour milieu de terrain, AM pour milieu offensif ou ailier selon les côtés et FW pour attaquant). Pour construire une distance interprétable, nous opterons pour une représentation en trois dimensions (3D) plutôt qu'une en deux dimensions (2D) (voir figure 6 page 25). L'intérêt de cette représentation 3D réside dans sa meilleure interprétabilité. En effet, dans une représentation en 2D, deux joueurs jouant l'un à droite et l'autre à gauche sont plus éloignés que le sont deux joueurs dont l'un joue à droite et l'autre au centre. Cela impliquerait par exemple que deux latéraux soient moins similaires entre eux qu'un latéral et un défenseur central, ou que deux ailiers soient moins similaires entre eux que le sont un ailier droit et un milieu offensif, ce qui est contradictoire avec l'interprétation du football.

Ainsi, le positionnement d'un joueur sur le terrain de football sera représenté par un point (x, y, z) d'un espace euclidien à trois dimensions, (x, y) pour le côté et z pour la position. Les valeurs de (x, y) sont fixées pour chaque côté, quelque soit la position, de sorte que les distances entre deux côtés quelconques soient la même. Ainsi, les trois côtés L, R et C seront vus comme les sommets d'un triangle équilatéral, de coordonnées respectives $(0, 0)$, $(1, 0)$ et $(\frac{1}{2}, \frac{\sqrt{3}}{2})$, et donc de longueur 1. Les valeurs de z sont les entiers de 0 à 4 respectivement pour les positions D, DM, M, AM et FW. La figure 7 est une représentation en perspective de la figure 6.(b) et illustre mieux le principe.

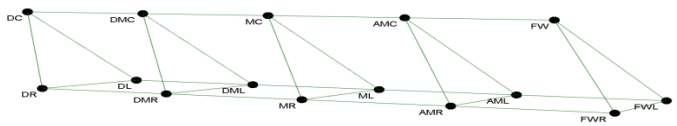


FIGURE 7 – Représentation en perspective des positions avec les trois côtés en 3D

Par exemple, les coordonnées du milieu offensif (AMC) sont $(\frac{1}{2}, \frac{\sqrt{3}}{2}, 3)$, ceux de l'avant-droit (FWR) sont $(1, 0, 4)$.

Maintenant que l'on sait représenter chaque position en 3D, il faut trouver la position moyenne d'un joueur ayant joué à plusieurs positions avec les proportions données par Y_{15}. Cela se fera en deux étapes :

• Dans un premier temps, on déterminera le côté moyen occupé par le joueur pour chaque position grâce à la formule suivante :

$$c(q_{ik}) = \frac{\sum_{j=1}^{3} \omega_{ijk} q_{ijk}}{\sum_{j=1}^{3} \omega_{ijk}}$$

où $q_{ijk} \in \mathbb{R}^2$ représente les coordonnées du j^{eme} côté occupé par le joueur i à la k^{eme} position (on suppose que les trois côtés sont numérotés de 1 à 3 et les cinq positions de 1 à 5) ; ω_{ijk} le pourcentage de fois où le joueur i a joué au côté j de la position k (valeur donnée par Y_{15}).

• Dans un second temps, on détermine la position finale du joueur en moyennant ses côtés moyens pour chaque position par les pourcentages de fois où il a joué à chacune de ces positions. En effet, soit $\omega_{ik} = \sum_{j=1}^{3} \omega_{ijk}$ le pourcentage de fois où le joueur i a joué à la position k, tous côtés confondus. Alors la position finale est donnée par les formules suivantes :

$$c(q_i) = \sum_{k=1}^{5} c(q_{ik})\omega_{ik} \qquad c(z_i) = \sum_{k=1}^{5} c(z_{ik})\omega_{ik}$$

où $z_{ik} \in \{0, ..., 4\}$ est la troisième dimension qui représente la k^{eme} position du joueur i.

Au final, la distance de Manhattan sera utilisée pour évaluer la dissimilarité entre deux joueurs i et j pour la variable position 1 représentée comme décrit ci-haut :

$$d_{pos1}(i, j) = d_M(c(q_i), c(q_j)) + d_M(c(z_i), c(z_j).)$$

→ **Variables de position 2, Y_{11} (voir section 3.2) :** La figure 3.b (page 8) illustre les 11 positions prises en compte par ces variables. Contrairement au cas précédent où des points à trois dimensions ont été moyennés par des coefficients donnés par Y_{15} pour représenter le positionnement d'un joueur, la variable Y_{11} donne une information binaire et la mesure de dissimilarité à construire doit prendre en compte et les coordonnées en 3D des positions sur la figure 3.(b) et l'information binaire donnée par Y_{11}.

Christian Henning et Bernhard Hausdorf dans leur article publié en 2006 : "Design of dissimilarity measures : A new dissimilarity between species distribution areas" (voir référence 4) ont proposé une nouvelle mesure de dissimilarité entre les aires de répartition d'espèces en travaillant sur la présence et l'absence de ces espèces dans des régions données. Ils ont déclaré que si deux espèces A et B sont présentes sur deux petites zones disjointes, elles sont très dissimilaires, mais les deux devraient être traitées comme similaires à une espèce C couvrant une plus grande superficie qui comprend à la fois A et B si les zones doivent être interprétées comme des regroupements d'espèces. Dans cet article, les données de répartition des espèces sont présentées dans une certaine région géographique, et le "coefficient geco" (le nom vient de "distance géographique et congruence"), qui est la distance géographique entre deux régions, est également introduit comme nouvelle mesure de dissimilarité.

En fait, l'idée vient du coefficient de Kulczynski (Kulczynski, 1927a, voir référence 5), donné par la formule

$$d_k(A_1, A_2) = 1 - \frac{1}{2}(\frac{|A_1 \cap A_2|}{|A_1|} + \frac{|A_1 \cap A_2|}{|A_2|})$$

où A_i est la région géographique de l'objet i et $|A_i|$ le nombre d'éléments dans la région géographique de l'objet i. Ensuite, Hennig et Hausdorf ont conçu le coefficient geco dans lequel le coefficient de Kulczynski et l'information géographique sont incorporés. La définition générale est donnée par

$$d_G(A_1, A_2) = \frac{1}{2}(\frac{\sum_{a \in A_1} \min\limits_{b \in A_2} \quad u(d_R(a,b))}{|A_1|} + \frac{\sum_{b \in A_2} \min\limits_{a \in A_1} \quad u(d_R(a,b))}{|A_2|})$$

où u est une transformation croissante avec $u(0) = 0$ et $d_R(a, b)$ est la distance entre les objets a et b. Dans la formule précédente, lors du calcul de la distance globale, l'idée est de ne pas incorporer le nombre d'absences communs à A_1 et A_2.

Dans notre cas, nous définissons le coefficient d_G en termes de Y_{11}, de sorte que les espèces sont remplacées par les joueurs, et les emplacements géographiques par les 11 positions prises en compte par Y_{11}. Le concept est très similaire ; c'est-à-dire que les absences communes de position entre deux joueurs seront ignorées, car les prendre en compte rendrait ces joueurs plus similaires, même s'ils n'apparaissent pas dans ces positions. Nous utiliserons le coefficient d_G pour mesurer les dissimilarités entre les joueurs pour la variable Y_{11}. A_i représentera l'ensemble des positions occupées par le joueur i, la distance d_R utilisée sera la distance euclidienne (d_E), les éléments de A_i seront les positions représentées par des points en 3D comme décrit précédemment et la fonction u sera l'identité :

$$d_{pos2}(A_1, A_2) = \frac{1}{2}(\frac{\sum_{a \in A_1} \min\limits_{b \in A_2} \quad d_E(a,b)}{|A_1|} + \frac{\sum_{b \in A_2} \min\limits_{a \in A_1} \quad d_E(a,b)}{|A_2|}).$$

Ci-desssous, nous illustrons le calcul de cette dissimilarité pour les joueurs Cristiano Ronaldo (joueur 1) et Lionel Messi (joueur 2).

Joueurs	Liste des positions										
	DC	DL	DR	DMC	MC	ML	MR	AMC	AML	AMR	FW
Joueur 1	0	0	0	0	0	0	0	0	1	1	1
Joueur 2	0	0	0	0	0	0	0	1	0	1	1

TABLE 12 – variable Y_{11} : Cristiano Ronaldo et Lionel Messi

Ainsi, $A_1 = \{AML, AMR, FW\}$ et $A_2 = \{AMC, AMR, FW\}$, $|A_1| = |A_2| = 3$. Le tableau suivant permet de calculer chacun des termes des deux sommes dans l'expression de d_{pos2} :

Jou1↓/Jou2→	AMC	AMR	FW	$\min\limits_{a \in A_1}$
AML	1	1	$\sqrt{2}$	1
AMR	1	0	$\sqrt{2}$	0
FW	1	$\sqrt{2}$	0	0
$\min\limits_{a \in A_2}$	1	0	0	

TABLE 13 – Distances de positions : Cristiano Ronaldo et Lionel Messi

En conclusion, la dissimilarité entre Cristiano Ronaldo (C.R) et Lionel Messi (L.M) pour la variable Y_{11} est

$$d_{pos2}(C.R, L.M) = \frac{1}{2}(\frac{1+0+0}{3} + \frac{1+0+0}{3}) = \frac{1}{3}.$$

4.3 Agrégation des mesures construites dans la section précédente

Dans la section 4.2, nous avons discuté de différentes mesures de dissimilarité pour des variables de différents types. La distance de Manhattan a été sélectionnée pour l'agrégation des variables de comptage de niveau supérieur, des compositions de niveau inférieur et de certaines autres variables (âge, poids, taille, apps, mins). Toutes ces variables de performance seront représentées dans une mesure de dissimilarité, que nous appellerons la dissimilarité de performance. Pour les variables d'équipe et de ligue, la distance de Manhattan a été adoptée dans le sens où x_l et x_{tp} sont combinées avant de prendre la valeur absolue. Pour les deux variables de position, nous avons construit une nouvelle mesure de dissimilarité entre les positions des joueurs pour Y_{15}, et la mesure de dissimilarité pour Y_{11} est conçue sur la base du "coefficient geco" de Hennig et Hausdorf (voir référence 4), où les espèces sont remplacées par des joueurs.

Afin de trouver une matrice de dissimilarité unique sur les données, ces mesures doivent être agrégées. Deux aspects devraient être discutés : d'une part comment choisir une technique de standardisation appropriée pour rendre les différentes valeurs comparables, d'autre part quels poids subjectifs devraient être attribués à ces dissimilarités une fois la méthode de standardisation choisie. Dans les applications, nous nous laisserons le choix entre trois constantes de standardisation afin de visualiser leurs différents impacts sur les résultats : l'écart absolu médian (voir section 4.1.3), la racine carrée de la variance empirique non biaisée ($\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2$), et l'étendue ($\max\limits_{1\leq i\leq N} x_i - \min\limits_{1\leq i\leq N} x_i$), où x désigne le vecteur (partie triangulaire supérieure stricte de la matrice de dissimilarité) de toutes les dissimilarités évaluées entre les joueurs 2 à 2, pour chacune des variables (performance, équipe et ligue, position 1 et position 2), et $N = \frac{n(n-1)}{2}$ avec n le nombre de joueurs dans le jeu de données.

Une fois que la standardisation a été appliquée à chaque matrice de dissimilarité, des poids appropriés doivent être attribués à chacun d'eux pour avoir la matrice finale. Le poids attribué aux dissimilarités pour les variables de niveau supérieur, les compositions de niveau inférieur et certaines autres variables est le total des pondérations du tableau 11 (page 23), qui est de 43.5, parce que nous voulons préserver les pondérations de ces variables. Pour les données de position, Y_{15} et Y_{11} ont respectivement 15 et 11 variables. Les pondérations de ces dissimilarités peuvent être les nombres totaux de ces variables, mais nous voulons que les deux types de variables de position aient le même impact sur la matrice de dissimilarité finale, de sorte que les mesures de dissimilarités d_{pos1} et d_{pos2} se verront attribuer chacun un poids de $13 = \frac{15+11}{2}$. L'attribution du nombre total (3) de variables d'équipe et de ligue comme poids peut également être adoptée pour les dissimilarités de ces variables, mais un joueur peut avoir des performances très différentes, s'il passe d'une équipe forte à une équipe moins forte, ou vice versa ; par conséquent, nous devons accorder un poids qui permet de discriminer les joueurs en fonction de leurs équipes et de leurs ligues. Nous pensons que le poids devrait être supérieur à 3, mais il ne devrait pas être beaucoup plus élevé, car sinon les dissimilarités des joueurs peuvent être dominées par ces variables. Ainsi, nous avons décidé d'attribuer un poids égal au double du nombre total de variables d'équipe et de ligue.

Encore une fois, toutes les pondérations dans ce document sont faites en fonction de notre interprétation du football, et peuvent varier dans les applications en fonction des utilisateurs (par exemple les managers, les recruteurs, etc.) et selon les profils ou les qualités des joueurs qui les intéressent le plus.

En conclusion, la mesure de dissimilarité globale que nous utiliserons dans la suite est donnée par

$$d_{fin}(x,y) = \frac{\sum_{k=1}^{4}\omega_k d_k(x,y)}{s_k}$$

avec les définitions des paramètres données dans le tableau suivant :

	Performance	Position 1	Position 2	Ligue et équipe
Dissimilarité	$d_1(x, y)$	$d_2(x, y)$	$d_3(x, y)$	$d_4(x, y)$
Poids	43.5	13	13	4
Constante de standardisation	s_1	s_2	s_3	s_4

Table 14

4.4 Une première application : Des requêtes de dissimilarité

A présent, la matrice de dissimilarité à disposition, nous pouvons nous amuser à déterminer pour un joueur i, les n joueurs les plus similaires à lui. C'est ce que nous appelons une requête de dissimilarité. Bien évidemment, il sera possible de varier les pondérations des variables en fonction des requêtes. Par exemple, si le joueur qui nous intéresse est un défenseur, nous pourrions augmenter les poids des variables défensives (tacle, interception, dégagement, etc.) pour calculer la matrice de dissimilarité et trouver les joueurs les plus similaires à notre défenseur.

Cette idée de requête de dissimilarité a été la consécration de ce projet. En effet, nous avons développé une application web avec *R shiny* pour ce faire. Voici comment elle fonctionne : Sur la première page de navigation, vous calculerez une matrice de dissimilarité avec des pondérations et une technique de standardisation de votre choix, et vous la sauvegarderez sur votre ordinateur. La deuxième page est consacrée aux requêtes de distance. Après avoir chargé la matrice calculée précédemment, vous choisissez le joueur qui vous intéresse, le nombre n de joueurs similaires que vous voulez et l'application vous retourne par ordre croissant de dissimilarité la liste des joueurs demandés. Notez que vous pourriez appliquer des filtres sur l'âge des joueurs, le poids, la taille, le nombre d'apparitions ou même le nombre de minutes jouées.

Les données utilisées sont également visualisables et directement téléchargeables depuis l'application. Il y a même une page consacrée à la détermination de la bonne constante pour la transformation des variables (voir section 4.1.2).

Des photos des pages de l'application sont disponibles en annexe, les codes sources ".R" sont disponibles sur mon github via le lien fourni en référence 6.

5 Remarques et Conclusion

Durant environ trois mois, j'ai travaillé avec passion sur un sujet d'analyse sportive. Cela pouvait faire peur au départ car je me suis attaqué à un travail volumineux, avec un jeu de données assez grand de part le nombre de variables (plus d'une centaine), qui plus ont des types variés dont chacun nécessitait un traitement spécifique. En analyse sportive, l'aspect le plus important est de faire en sorte que les calculs reflètent au mieux la réalité des athlètes sur le terrain et mon projet n'a pas été une exception à la règle. Dans le cas du recrutement de joueurs en football, quand il s'agit de trouver des joueurs correspondant au mieux à un profil donné, il n'y a pas vraiment de critère de validation formelle à moins de s'en tenir à l'avis globale des experts et aussi des fans. Cependant, il n'y a pas besoin d'être un grand observateur du football pour savoir qu'une mesure qui évalue un défenseur rugueux comme joueur le plus similaire à un attaquant fin et dribbleur au détriment d'autres attaquants connus pour être du même style est une catastrophe à tous les niveaux. Heureusement, cela n'est pas le cas ici comme on peut le voir sur les quelques requêtes de dissimilarité présentées en annexe (les joueurs choisis sont célèbres et connus d'à peu près tout le monde). Comme dit au début de ce projet, l'intérêt d'un tel outil de comparaison réside essentiellement dans le recrutement de joueurs, mais d'autres analyses peuvent être effectuées sur ces données pour d'autres intérêts. Je pense notamment aux méthodes de clustering pour regrouper des joueurs où même des clubs en fonction de certains critères, et dont les intérêts peuvent être nombreuses pour les journalistes, les managers, les organismes du football mondial ou même les fans. D'ailleurs, cela devrait être la prochaine étape de ce projet après construction de la matrice de dissimilarité, mais toutes les contraintes et difficultés rencontrées, pour l'accès aux données et surtout leur nettoyage, la prise en main de R shiny, les efforts de déploiement en vain de l'application (qui marche pourtant bien en local), ajouté à cela le manque de motivation quelques fois du fait de la période estivale, ont eu raison de moi. Néanmoins, réaliser ce projet m'a apporté énormément de connaissances, mathématiques sur les méthodes de traitement de données, et surtout informatique, que ce soit pour la collecte de données par la technique du web scraping ou dans le développement d'application web, et ce fut une vraie satisfaction. Cela constitue pour moi un bon point de départ en analyse footballistique, un domaine dans lequel j'espère réaliser de grandes choses dans un avenir proche.

6 Remerciements

Je voudrais remercier le docteur **Serhat AKHANLI** pour sa thèse qui m'a permis d'avoir de quoi travailler pour la première fois sur un sujet d'analyse footballistique, un domaine qui me passionne tellement.

Je remercie ma responsable de formation, et superviseuse de ce TER, **Christine KERIBIN**, de son assistance, ses corrections, et particulièrement de m'avoir obtenu le jeu de données qui m'a servit durant tout le projet, auprès du professeur Christian HENNIG, superviseur de la thèse de Serhat.

Merci à toi, Mouhite ADEBO, élève-ingénieur en informatique et ami de longue date, pour ton aide sur les questions liées à la programmation informatique durant tout mon projet.

7 Annexe

7.1 Quelques images de l'application des requêtes de dissimilarité

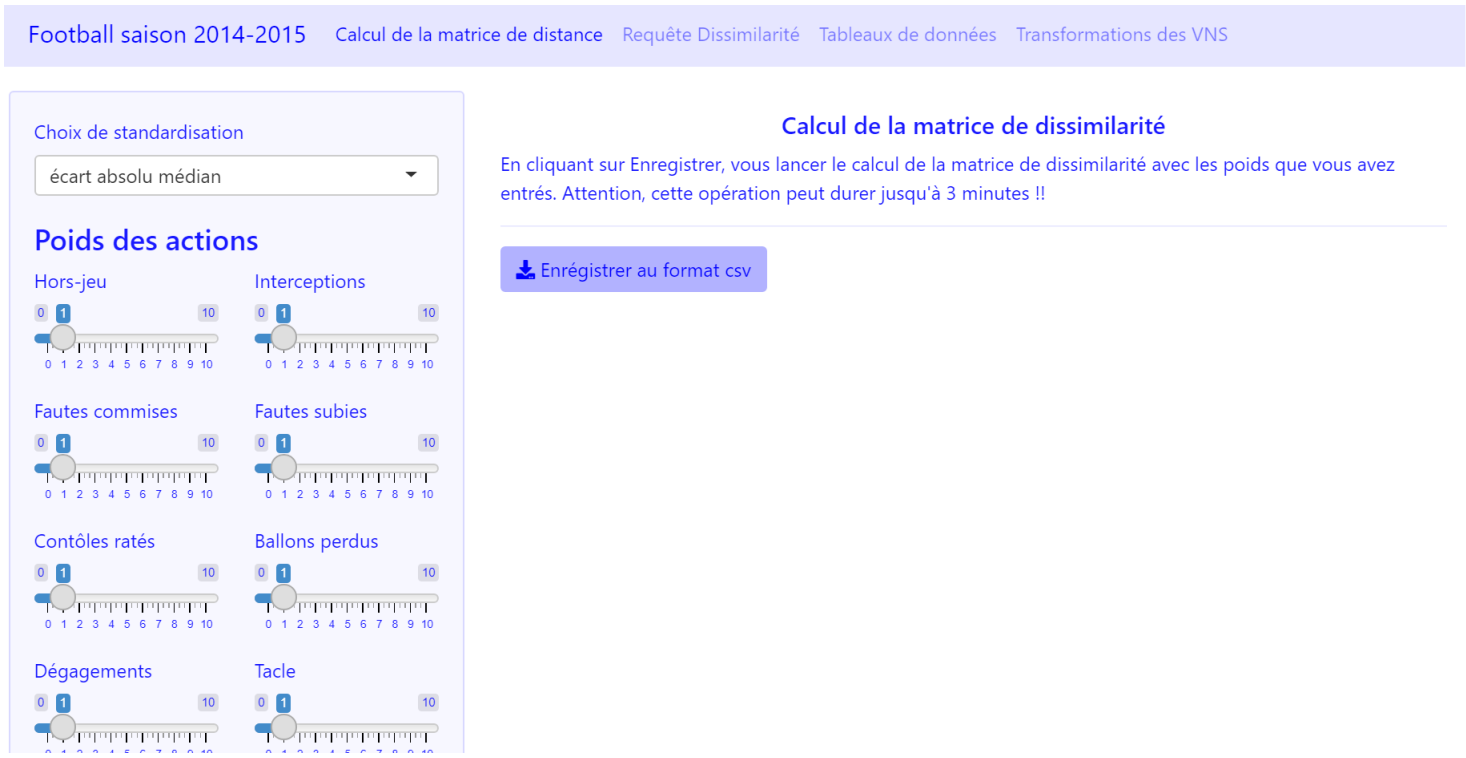


FIGURE 8 – Appli web : Page de calcul de la matrice de distance

FIGURE 9 – Requête distance : Lionel Messi

Figure 10 – Requête distance : Cristiano Ronaldo

Figure 11 – Requête distance : Eden Hazard


```
[1] "Name : Andrés Iniesta"
[1] "Team : Barcelona"
[1] "League : La Liga"
[1] "Height : 170"
[1] "Weight : 65"
[1] "Age : 30"
[1] "Position : Midfielder"
[1] "Apps : 27"
[1] "Mins : 1886"
[1] "Dissimilarity : 0"
[1] "-----------------------------------"
[1] "Name : Rafinha Alcantara"
[1] "Team : Barcelona"
[1] "League : La Liga"
[1] "Height : 178"
[1] "Weight : 67"
[1] "Age : 22"
[1] "Position : Midfielder"
[1] "Apps : 24"
[1] "Mins : 933"
[1] "Dissimilarity : 2.81"
[1] "-----------------------------------"
[1] "Name : Isco"
[1] "Team : Real Madrid"
[1] "League : La Liga"
[1] "Height : 176"
[1] "Weight : 74"
[1] "Age : 23"
```

Figure 12 – Requête distance : Andrés Iniseta


```
[1] "Name : Sergio Ramos"
[1] "Team : Real Madrid"
[1] "League : La Liga"
[1] "Height : 183"
[1] "Weight : 75"
[1] "Age : 29"
[1] "Position : Defender"
[1] "Apps : 29"
[1] "Mins : 2410"
[1] "Dissimilarity : 0"
[1] "-----------------------------------"
[1] "Name : Gerard Piqué"
[1] "Team : Barcelona"
[1] "League : La Liga"
[1] "Height : 192"
[1] "Weight : 85"
[1] "Age : 28"
[1] "Position : Defender"
[1] "Apps : 31"
[1] "Mins : 2745"
[1] "Dissimilarity : 2.64"
[1] "-----------------------------------"
[1] "Name : Shkodran Mustafi"
[1] "Team : Valence"
[1] "League : La Liga"
[1] "Height : 184"
[1] "Weight : 82"
[1] "Age : 23"
```

Figure 13 – Requête distance : Sergio Ramos

7.2 Liste des figures et pages

7.3 Liste des tables et pages

8 Référence

1. Daunis-i Estadella, J., Martin-Fernandez, J., and Palarea-Albaladejo, J. Bayesian tools for count zeros in compositional data. *Proceedings of CODAWORK*, 8 :8, 2008
2. Martin-Fernandez, J. A., Barcelo-Vidal, C., and Pawlowsky-Glahn, V. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35(3) :253–278, 2003.
3. Aitchison, J. The Statistical Analysis of Compositional Data. *Chapman and Hall, Ltd., London, UK*, 1986. ISBN 0-412-28060-4.
4. Hennig, C. and Hausdorf, B. Design of dissimilarity measures : A new dissimilarity between species distribution areas. *In Data Science and Classification*, pages 29–37. Springer, 2006.
5. Kulczynski, S. Die pflanzenassoziationen der pieninen. *Bulletin International de l'Academie Polonaise des Sciences et des Lettres, Classe des Sciences Mathematiques et Naturelles, B (Sciences Naturelles)*, II :57–203, 1927a.
6. Lien vers le Github de l'application : https ://github.com/akedjouadj/football-app.git