Analyse des statistiques des joueurs de Ligue 1 – saison 2024-2025

Artus Bleton, Guilhem Dupuy, Youssef Abida

Résumé

Dans cette étude, nous cherchons à caractériser les profils des joueurs de Ligue 1 (saison 2024–2025) à partir de leurs performances statistiques. L'objectif est double : d'une part, réduire la dimensionnalité du jeu de données tout en conservant sa structure informative ; d'autre part, identifier des groupes de joueurs similaires à l'aide de méthodes de clustering. Ces analyses s'appuient sur un jeu de données riche et détaillé, extrait de sources spécialisées.

Mots-clés : Analyse de données, Réduction de dimension, Clustering, PCA, t-SNE, UMAP, Isomap, Performance sportive, Analyse du football

Table des matières

1 Présentation des données

Les données proviennent du site `FBRef.com`, qui publie les statistiques avancées des joueurs à partir des bases de données StatsBomb. Le périmètre couvre l'ensemble des joueurs ayant participé à la saison 2024–2025 de Ligue 1.

Le jeu initial contient 666 joueurs et 146 variables numériques ou catégorielles (identifiants, pays, positions, etc.). Les principales familles de variables sont : performances offensives (tirs, buts, expected goals, passes décisives), actions défensives (tacles, interceptions, duels aériens), passes et construction du jeu (passes réussies / tentées, passes progressives), et maîtrise du ballon (dribbles, portées, pertes).

2 Préparation et nettoyage des données

2.1 Nettoyage commun à tous les jeux de données

Un ensemble d'opérations de prétraitement a été appliqué avant la construction des jeux de données d'analyse.

- **Filtrage par temps de jeu** (>500 minutes jouées) : afin de garantir une représentativité statistique suffisante.
- **Suppression des doublons, des variables redondantes ou non exploitables.**
- **Exclusion des gardiens de but** : présence susceptible de former un cluster homogène peu informatif qui déséquilibrerait les distances inter-joueurs, en augmentant la variance globale et en étirant artificiellement l'espace des représentations.
- **Suppression des variables peu informatives** (>40% de valeurs manquantes).
- **Suppression des lignes peu renseignées** (>95 % de valeurs manquantes).
- **Gestion des valeurs manquantes (NaN)** :
  - remplacées par 0 lorsque leur absence reflétait un événement non survenu (par ex. aucun tir, aucun duel...),
  - imputées par la médiane pour les ratios et pourcentages.

Ces opérations ont permis de constituer un jeu cohérent et complet, composé au final de 333 joueurs de champ, soit environ 50% du jeu initial, prêt pour les analyses de réduction de dimension et de clustering.

2.2 Constitution des quatre jeux de données d'analyse

À partir du jeu de données nettoyé, quatre sous-ensembles ont été constitués afin d'explorer différents axes d'analyse :

- **Raw** : conserve les statistiques brutes des joueurs, sans normalisation par le temps de jeu. Il permet de comparer les volumes de production totale, mais désavantage les joueurs à faible temps de jeu.
- **Per90** : reprend les mêmes variables, normalisées par 90 minutes jouées. Cette transformation permet de comparer les joueurs en termes d'efficacité, en neutralisant les effets du temps de jeu mais peut avantager les remplaçants performants

sur de courtes durées.

— **Custom** : version nettoyée normalisée du jeu de données, ne retenant que les joueurs ayant joué plus de 500 minutes et des variables exprimées par 90 minutes, en pourcentage ou en ratio. Les valeurs manquantes sont imputées de façon contextuelle (0 pour les taux d'événements, médiane pour les pourcentages). Ce jeu vise à décrire le style individuel plutôt que le volume de jeu.
— **Custom-gk** : version du jeu Custom conservant cette fois les gardiens de but, afin d'évaluer leur impact sur la structure globale des clusters. Il s'agit du seul sous-ensemble incluant cette catégorie de joueurs.

3 Réduction de Dimension

3.1 Objectif

Avant toute tentative de clustering, une réduction de la dimensionnalité a été entreprise afin de faciliter la visualisation des données, réduire leur complexité, et améliorer la stabilité des algorithmes de clustering. L'ensemble des méthodes testées projettent les données initiales ($\sim$100 variables) dans un espace de dimension plus faible.

3.2 Méthodes de réduction de dimension comparées

Quatre techniques de réduction non supervisée ont été comparées :

PCA (Analyse en Composantes Principales) – méthode linéaire classique, permettant de projeter les données sur les directions maximisant la variance. Nous l'avons choisie comme baseline, pour pouvoir évaluer l'apport d'autres méthodes plus complexes.

t-SNE (t-Distributed Stochastic Neighbor Embedding) – méthode non linéaire centrée sur la préservation locale des distances. En associant chaque joueur à ceux qui lui sont le plus similaires, nous espérons faire émerger des regroupements locaux intéressants dans l'espace réduit.

Isomap – extension non linéaire de MDS, basée également sur des distances locales. Contrairement à t-SNE, Isomap utilise des critères géométriques (plus court chemin entre nœuds), et non probabilistes (t-SNE et la distribution normale). Nous utilisons cette méthode en complément de t-SNE, réputée instable et sensible aux hyperparamètres.

UMAP (Uniform Manifold Approximation and Projection) – méthode non linéaire récente découverte dans la littérature [1, 2]. Elle s'appuie, comme Isomap, sur une construction du graphe des k plus proches voisins, mais remplace l'approche probabiliste de t-SNE par la notion de voisinage flou, qui attribue à chaque paire de points un degré d'appartenance pondéré.
Nous l'avons intégrée afin de comparer des méthodes de complexité modérée (t-SNE, Isomap) à une approche de pointe, proche de l'état de l'art. UMAP servira également de référence robuste, pour mettre en perspective t-SNE (souvent instable) et Isomap (sensible au bruit).

3.3 Métriques d'évaluation

Afin d'évaluer la qualité des projections, nous avons utilisé 4 métriques, centrées sur la préservation de la structure de voisinage :

- **Trustworthiness** : mesure à quel point les voisins proches dans l'espace réduit étaient également voisins dans l'espace d'origine. Score dans [0 ;1], avec 1 meilleur score possible.
- **Continuity** : mesure complémentaire de la trustworthiness : est-ce que les voisins dans l'espace original ont été préservés dans l'espace réduit ? Score dans [0 ;1], avec 1 meilleur résultat.
- **Distance correlation** : corrélation de Spearman entre les distances sur l'espace initial et l'espace réduit. Dans [-1 ; 1], 1 meilleur score possible.
- **Mean Relative Rank Error (MRRE)** : moyenne normalisée de l'écart de rang entre les voisins dans les deux espaces. 0 est le meilleur score possible.

À partir de ces métriques, nous avons défini un indicateur synthétique correspondant à la somme des écarts quadratiques entre les valeurs observées et leurs scores théoriquement optimaux. Cet indicateur, noté *somme des MSE*, est reporté dans nos tableaux de résultats afin de faciliter la comparaison globale des méthodes.

Le paramètre $k = 10$ a été choisi pour ces évaluations, en accord avec des recommandations de la littérature utilisant $k < 20$ pour des mesures de trustworthiness [3].

3.4 Résultats

Les résultats présentés correspondent aux meilleures performances après optimisation des hyperparamètres via une recherche en grille (*grid search*). Les hyper-paramètres optimaux sont détaillés en Annexe 1.

Ces scores évaluent uniquement la qualité de la projection. Ils ne préjugent pas de la qualité du clustering obtenu par la suite, qui dépend entre autres :

- de la distribution spatiale post-réduction,
- de la forme des groupes,
- et de la compatibilité avec les différents algorithmes de regroupement.

Il nous faut additionnellement intégrer ces composantes au choix des méthodes retenues pour le clustering, et également trouver un compromis entre qualité de projection et complexité.

Méthode	Trustworthiness	Continuity	Distance corr.	MRRE	Somme MSE
PCA-raw	0.997656	0.995897	**0.998275**	0.067003	0.0045
PCA-per90	**0.997883**	0.996068	0.997254	0.067810	0.0046
PCA-custom-GK	0.997791	0.996088	0.997459	0.065161	0.0043
PCA-custom	0.997912	**0.998422**	0.997852	**0.063516**	**0.0040**
tSNE-raw	**0.9745**	**0.9915**	0.7160	**0.4097**	0.2492
tSNE-per90	0.9711	0.9911	0.7468	0.4404	0.2589
tSNE-custom-GK	0.9717	0.9918	**0.7627**	0.4276	**0.2400**
tSNE-custom	0.9669	0.9909	0.7509	0.4514	0.2670
Isomap-raw	**0.9687**	**0.9906**	**0.9430**	**0.2978**	**0.0930**
Isomap-per90	0.9591	0.9898	0.8921	0.3565	0.1405
Isomap-custom-GK	0.9592	0.9900	0.8895	0.3885	0.1649
Isomap-custom	0.9468	0.9891	0.8929	0.3864	0.1637
UMAP-raw	0.9376	0.9900	0.5726	0.5344	0.4720
UMAP-per90	0.9391	0.9900	**0.6726**	**0.5238**	**0.3852**
UMAP-custom-GK	**0.9413**	**0.9904**	0.6239	0.5646	0.4636
UMAP-custom	0.9234	0.9887	0.6714	0.5244	0.3888

TABLE 1 – Résultats comparés des méthodes de réduction de dimension

3.5 Analyses

Les résultats montrent que PCA obtient les meilleurs scores globaux (Trustworthiness, Continuity, Distance correlation proches de 1, confirmant sa capacité à préserver la structure globale des données. Ce résultat était attendu : la PCA, méthode linéaire, conserve efficacement la variance principale des variables corrélées mais ne capture pas les relations non linéaires.

Les méthodes non linéaires (t-SNE, Isomap, UMAP) présentent des comportements contrastés, cohérents avec leurs principes : t-SNE maximise la fidélité locale (bonne Trustworthiness et Continuity) mais déforme les distances globales (moins bonne Distance Correlation globale). C'est un effet que l'on pouvait attendre de sa modélisation probabiliste. Isomap conserve mieux la géométrie globale grâce à l'apport de son approche géodésique (corrélation d'environ 0.9 entre les distances), au prix d'une sensibilité au bruit et au choix du paramètre k qui ont été plus durs à définir. UMAP, obtient en termes de métriques les pires scores. Ce résultat, surprenant à première vue, s'explique par le fonctionnement d'UMAP qui ne vise pas à conserver les distances visuelles. La méthode s'appuie sur la structure topologique du graphe et ne cherche pas à maximiser les métriques mesurées, ces mauvais scores ne reflètent donc pas nécessairement l'utilité de la méthode pour notre objectif de clustering.

Interprétations visuelles
Certaines méthodes de réduction offrent des premières pistes d'interprétation visuelle. C'est le cas d'Isomap et de la PCA, où semblent émerger parmi leurs dimensions principales les notions de "qualité" de joueurs (bons joueurs / mauvais joueurs), ainsi que d'impact offensif.

Le graphique ci-dessous, généré via Isomap, illustre cette tendance. On observe que les meilleurs attaquants (Ousmane Dembélé, Ryan Cherki, etc.) se positionnent dans la partie supérieure gauche du plan réduit. Un attaquant comme Alexandre Lacazette, ayant réalisé une saison très moyenne, se situe toujours à gauche sur l'axe horizontal mais proche de zéro sur l'axe vertical. Akor Adams, identifié comme l'un des attaquants les moins performants

de Ligue 1, se trouve également à gauche, mais dans la partie inférieure du plan. Enfin, Duje Caleta-Car, défenseur souvent critiqué pour ses erreurs et son manque de finesse technique, apparaît dans la partie inférieure droite, reflétant à la fois sa moindre efficacité et son rôle défensif.

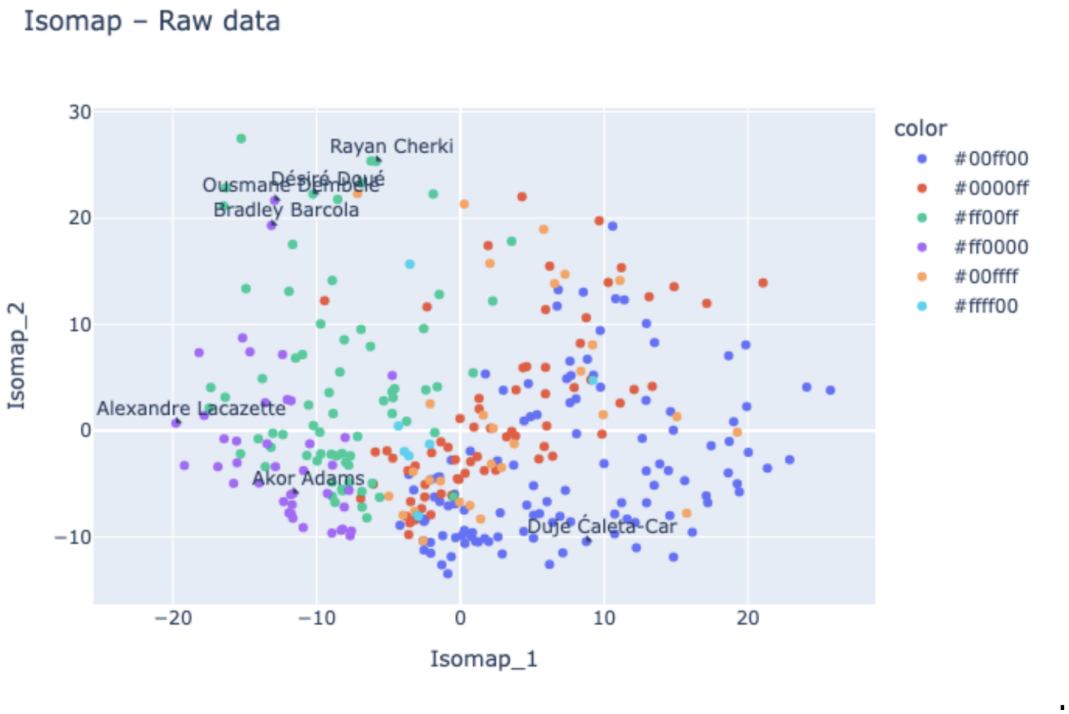

FIGURE 1 – Projection sur les 2 dimensions principales des données réduites via Isomap

Choix des meilleurs jeux de données réduites

Aucune méthode ni jeu de données initial ne domine sur tous les critères : les performances varient également selon le jeu de données, suggérant une forme de compatibilité méthode–données. Les jeux normalisés ou structurés (comme Custom) favorisent les méthodes linéaires (PCA), tandis que les jeux plus bruts ou complexes (Raw, Per90) se prêtent mieux aux approches géométriques ou de voisinage (Isomap, UMAP, t-SNE).

Nous avons émis l'hypothèse que cette logique de compatibilité se maintiendrait également lors de la phase de clustering. Nous avons donc choisi de conserver un jeu réduit par méthode, afin de conserver la diversité structurelle sur nos jeux de données durant cette phase. En s'appuyant sur nos métriques, nous conserverons pour l'étude le jeu ayant les meilleurs résultats pour chaque méthode :

- PCA-Custom
- t-SNE-Custom-GK
- Isomap-Raw
- UMAP-Per90

4 Clustering

4.1 Objectif

Une fois la dimensionnalité réduite, l'objectif est d'identifier des regroupements de joueurs aux caractéristiques statistiques similaires. Le clustering nous permettra d'explorer la structure latente du jeu de données, et de dégager des profils-types de joueurs compréhensibles par l'humain à partir d'un jeu initialement trop complexe à analyser directement.

Les meilleurs clusters seront conservés pour l'analyse globale de nos données.

4.2 Méthodes de clustering comparées

Nous avons sélectionné six algorithmes de clustering, couvrant différents types d'approche. Certains nécessitent de fixer à l'avance le nombre de clusters souhaité (noté (F) pour fixé), tandis que d'autres le déterminent automatiquement (noté (A) pour automatique).

Méthodes de partitionnement explicite :

— K-means (F) : algorithme simple qui servira de baseline aux autres méthodes de clustering.

Méthodes à base de modélisation statistique :

— Gaussian Mixture Models (GMM) - (F) : plus souple que K-Means, qui permet d'estimer des clusters elliptiques de forme variable.

Méthodes à base de densité :

— DBSCAN (A) (Density-Based Spatial Clustering of Applications with Noise) : Contrairement à GMM, il est capable d'identifier des clusters de forme arbitraire et de détecter les outliers, ce qui est intéressant dans nos jeux de données.

Méthodes de type propagation :

— Affinity Propagation (A) : repose sur la recherche de points "exemplaires" autour desquels se construisent les clusters. Recherche par itération, via la propagation de scores (Responsabilité et Disponibilité) permettant d'évaluer à quel point un point est représentatif pour un autre. Contrairement aux approches classiques basées sur la distance ou la densité, elle peut révéler des structures plus atypiques dans les données, sans faire d'hypothèses sur la forme ou la taille des clusters.

Méthodes hiérarchiques :

— Agglomerative Clustering (F) : construit une hiérarchie de clusters en fusionnant itérativement les groupes les plus similaires selon une mesure de distance et un critère de liaison (linkage). Offre une interprétabilité structurelle.

— Ward (avec linkage ward) - (F) : variante de l'approche agglomérative minimisant l'augmentation de la variance intra-cluster à chaque fusion. Elle favorise la formation de groupes compacts et de taille homogène, souvent plus adaptés à des données quantitatives continues.

Méthodes orientées scalabilité :

— BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) - (F) : conçue pour traiter efficacement de grands volumes de données. Elle construit progressivement une structure arborescente compacte (Clustering Feature Tree) qui résume la distribution des points avant l'application d'un algorithme de regroupement final.

4.3 Métriques et méthodologie d'évaluation

Pour les différentes méthodes de clustering, nous explorerons les meilleures configurations sur nos données via :

- Le choix des hyperparamètres
- L'analyse visuelle des clusters en 2D ou 3D

Les meilleurs résultats seront évalués quantitativement selon 3 métriques :

- **Silhouette Score** : mesure de compacité et de séparation dans [-1 ; 1]. 1 étant le meilleur score
- **Calinski-Harabasz Index** : ratio pondéré entre la variance inter/intra cluster, il mesure de dispersion intra et inter-cluster dans [0 ; +inf[. Ce score doit être maximisé, tout en sachant qu'il dépend de k = nombre de clusters (augmente avec k).
- **Davies-Bouldin Index** : mesure la qualité de séparation entre clusters (ratio entre dispersion intra-cluster et séparation inter-cluster) dans [0 ; +inf[. Ce score doit être minimisé.

Enfin, pour les configurations présentant les meilleurs clusterings, nous approfondirons l'évaluation en étudiant deux aspects complémentaires : la **stabilité** et la **robustesse**

4.4 Résultats

Méthode	Réduction	Nombre clusters	Silhouette	C.H. index	D.B. index
K-Means	PCA	6	0.058	7.046	3.553
	t-SNE	7	0.3717	258.809	0.960
	Isomap	6	0.313	86.910	1.090
	UMAP	5	0.355	231.095	0.9933
GMM	PCA	10	0.028	6.535	3.359
	t-SNE	5	0.344	245.057	1.056
	Isomap	6	0.249	69.796	1.289
	UMAP	6	0.335	203.952	1.084
DBSCAN	PCA	1	NaN	NaN	NaN
	t-SNE	4	0.607	420.901	0.572
	Isomap	2	0.407	51.693	0.751
	UMAP	2	0.404	200.670	0.880
Affinity Propagation	PCA	13	0.013	5.775	2.994
	t-SNE	7	0.358	250.569	0.976
	Isomap	6	0.303	84.791	1.090
	UMAP	5	0.363	189.104	0.932
Agglomerative Clustering	PCA	3	0.380	5.80	0.437
	t-SNE	3	0.355	226.451	0.836
	Isomap	3	0.272	15.119	1.05
	UMAP	5	0.333	216.042	1.086
Ward	PCA	15	0.0202	6.76	2.7084
	t-SNE	5	0.3262	230.67	1.0165
	Isomap	6	0.2972	77.96	1.1359
	UMAP	10	0.3664	279.39	0.9040
BIRCH	PCA	12	0.0245	7.08	2.8924
	t-SNE	8	0.3525	240.54	0.9979
	Isomap	6	0.3008	71.91	1.0565
	UMAP	2	0.4302	344.53	0.9068

TABLE 2 – Résultats des méthodes de clustering selon réduction et métriques

Après évaluation de plusieurs méthodes, seules celles appliquées sur les réductions UMAP et t-SNE ont été retenues pour leur structuration efficace de l'espace latent.

Une méthode de clustering a été choisie par type de projection, en considérant le métriques d'évaluation utilisées. Les méthodes produisant plus de 10 ou moins de 3 clusters ont été écartées, car elles ne permettaient pas une segmentation pertinente du jeu de données.

Les approches retenues sont Agglomerative Clustering et K-Means, appliquées sur les données réduites par UMAP, offrant un compromis entre qualité des partitions et interprétabilité.

5 Analyse

La représentation 3D ci-dessous des clusters issus de l'Agglomerative Clustering montre principalement deux ensembles : les clusters 2 et 3 d'un côté, et les trois autres de l'autre. La distinction interne à ces groupes semble donc se jouer sur d'autres dimensions.

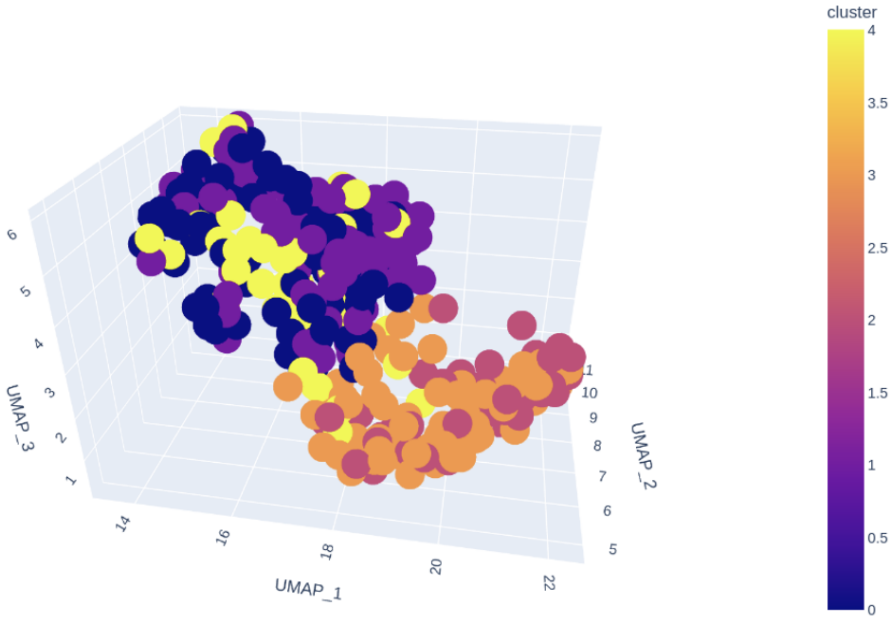

FIGURE 2 – Projection 3D des clusters - UMAP - agglomerative clustering

Il en est de même pour K-Means, qui a identifié le même nombre de clusters et effectué des regroupements analogues.

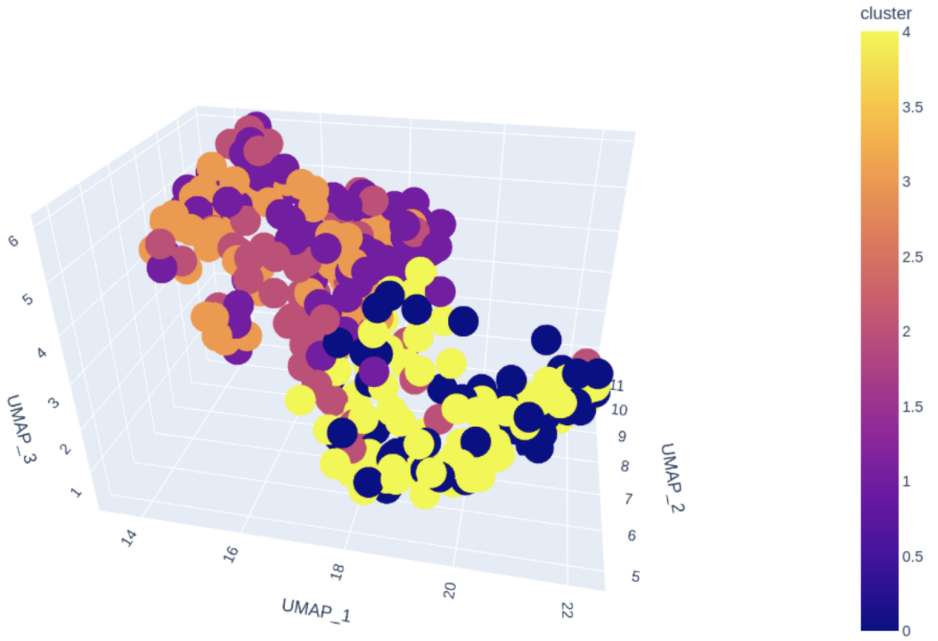

FIGURE 3 – Projection 3D des clusters - UMAP - K-means

Dans un second temps, la répartition des rôles (positions) a été examinée pour évaluer si cette dimension avait été naturellement prise en compte par le clustering. Cette analyse

s'est révélée partiellement concluante : certains postes formaient des ensembles cohérents, mais les positions hybrides restaient souvent mélangées.

Afin d'obtenir une interprétation plus pertinente, nous avons donc choisi d'aller plus loin en caractérisant les joueurs à partir de variables quantitatives décrivant leur style de jeu plutôt que leur poste :

- production offensive (per90_gls, per90_ast)
- progression du ballon (carries_prog)
- technique (pct_take_on_suc)
- récupération et activité défensive (tkl_plus_int, ball_recov)
- impact physique (pct_air_dual_won)
- expérience (âge)

L'analyse des profils moyens ainsi normalisés a permis de mieux différencier les groupes et de faire émerger les tendances dominantes (offensif, défensif, porteur, créatif...) nécessaires à l'interprétation des clusters.

Agglomerative Clustering

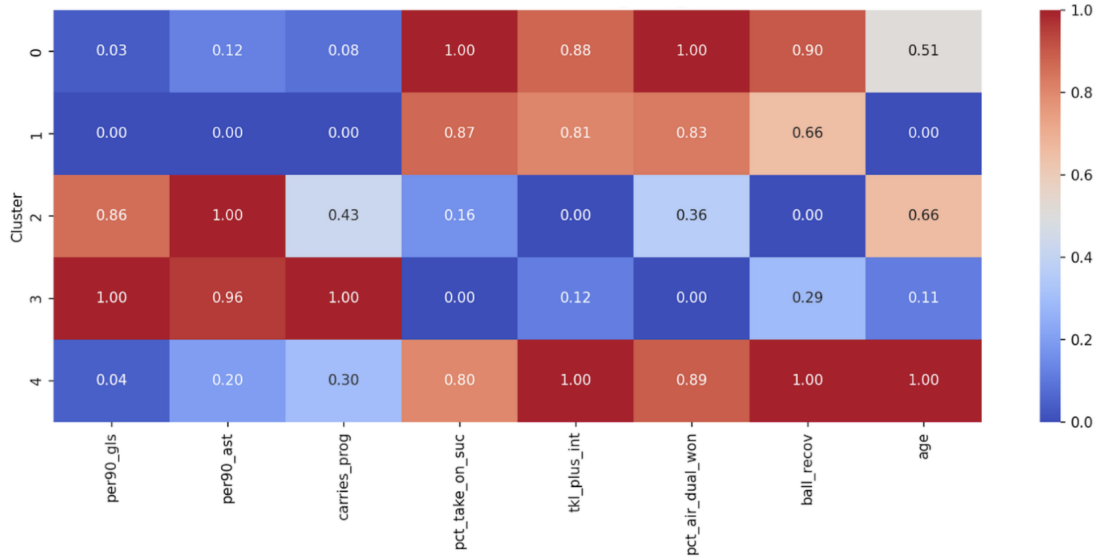

FIGURE 4 – Profils normalisés par clusters - Agglomérative clustering

Pour l'Agglomerative Clustering, les résultats confirment les tendances observées précédemment.

Les clusters 2 et 3 regroupent des joueurs au profil offensif. La variable carries_prog distingue le cluster 3, associé à des attaquants plus porteurs de balle. Le cluster 2, plus âgé, traduit un jeu davantage collectif et expérimenté.

Les clusters 0, 1 et 4 concernent principalement milieux et défenseurs. Les clusters 0 et 1 présentent des profils équilibrés sur l'ensemble des dimensions, le cluster 0 se démarquant par un âge moyen plus élevé, alliant expérience et vision de jeu. Enfin, le cluster 4, fortement corrélé aux variables tkl_plus_int et ball_recov, correspond logiquement à des profils à dominante défensive.

K-Means

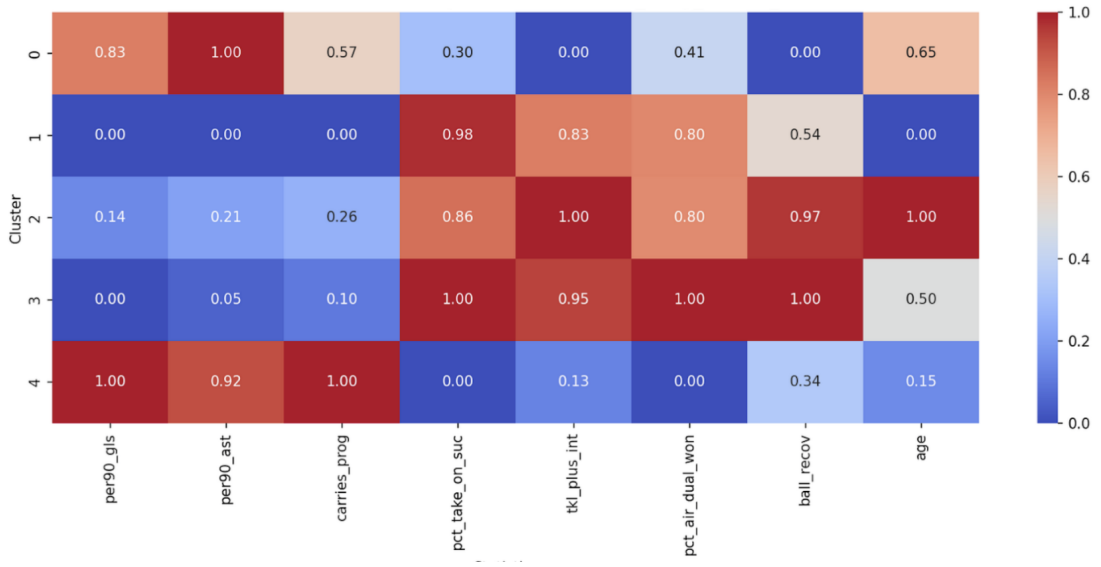

FIGURE 5 – Profils normalisés par clusters - K-means

Les mêmes tendances apparaissent parmi les clusters de K-means : les clusters 0 et 4 regroupent des joueurs offensifs, tandis que les clusters 1, 2 et 3 concernent les milieux et les défenseurs.

Parmi les offensifs, un groupe plus âgé et collectif (plus de passes décisives, moins de progression individuelle) s'oppose à un autre plus jeune, axé sur la création et le dribble.

Les clusters 2 et 3 représentent davantage la défense, très corrélés avec tkl_plus_int et ball_recov, le cluster 3 regroupant des profils plus complets, actifs en défense et en transition. Le cluster 1 correspond à des milieux plus généralistes, plus âgés.

Conclusion — Différences entre les deux méthodes

Les deux approches produisent des regroupements cohérents, mais Agglomerative Clustering offre une segmentation légèrement plus fine et nuancée, mieux adaptée à la détection de sous-profils intermédiaires.

À l'inverse, K-Means fournit une partition plus stable et homogène, mettant davantage en avant les grandes familles de rôles (offensifs, milieux, défensifs) sans forcément capturer les subtilités internes. En somme, Agglomerative se montre plus interprétable et précis, tandis que K-Means privilégie la clarté et la simplicité dans la structure des clusters.

Notons toutefois que K-Means performe remarquablement bien au regard de sa simplicité algorithmique, parvenant à capturer des structures de jeu comparables à celles issues d'une méthode hiérarchique bien plus complexe.

6 Robustesse et fiabilité

Le test de stabilité consiste à rééchantillonner aléatoirement le jeu de données (avec remise) afin de recalculer le clustering sur plusieurs sous-ensembles. Chaque nouvelle partition est ensuite comparée à la partition initiale au moyen de deux métriques :

- **ARI (Adjusted Rand Index)** : mesure la similarité entre partitions tout en corrigeant l'effet du hasard ;
- **NMI (Normalized Mutual Information)** : évalue la quantité d'information partagée entre partitions.

Des valeurs élevées de ces indices indiquent une structure de regroupement stable.

La robustesse est mesurée par ajout d'un bruit gaussien faible (1%, 3%, 5% de l'écart-type) aux coordonnées UMAP puis recalcul des clusters et évaluation d'ARI et NMI.

Ces tests permettent d'évaluer si les regroupements identifiés reposent sur une structure réelle et stable ou sur la variabilité intrinsèque des données.

	file	method	n	k	ARI_mean	ARI_std	NMI_mean	NMI_std	silhouette	calinski	davies
0	clusters_annotes_clusters_umap_agglomerative.csv	agglomerative	341	5	0.0741	0.0364	0.1637	0.0355	-0.0721	17.28	7.1092
1	clusters_annotes_clusters_umap_kmeans_5.csv	kmeans	341	5	0.0700	0.0277	0.1335	0.0349	-0.0448	13.71	8.1248

	noise_pct	ARI	NMI	method
0	1	0.041946	0.118427	agglomerative
1	3	0.137377	0.210646	agglomerative
2	5	0.114104	0.208644	agglomerative
3	1	0.049520	0.101829	kmeans
4	3	0.042615	0.098538	kmeans
5	5	0.042878	0.090091	kmeans

FIGURE 6 – Résultat des tests de stabilité / robustesse

Les résultats doivent être interprétés en tenant compte de la nature du jeu de données. Contrairement à des données artificielles présentant des classes bien séparées, les performances observées (ARI ≈ 0.07, NMI 0.13–0.16, silhouettes négatives) reflètent une continuité progressive entre styles de jeu. Dans le contexte du football, les rôles ne sont pas disjoints ; la continuité entre rôles explique la faible séparation statistique entre groupes.

Malgré cette complexité, les deux méthodes produisent des regroupements cohérents, distinguant des tendances offensives, défensives et constructrices. L'Agglomerative Clustering, légèrement plus robuste (NMI ≈ 0.21 sous bruit), capture mieux les relations hiérarchiques et progressives entre profils, tandis que K-Means, plus rigide, demeure sensible

aux variations locales. Ces résultats traduisent un espace de jeu continu où les frontières sont floues mais les tendances fonctionnelles identifiables.

7 Conclusions

L'objectif de ce travail était de caractériser les profils de joueurs de Ligue 1 à partir de données statistiques détaillées, via des techniques de réduction de dimension et de clustering non supervisé.

Les méthodes Agglomerative Clustering et K-Means, appliquées aux représentations UMAP, ont permis d'identifier cinq groupes de joueurs aux tendances distinctes : offensifs, constructeurs et défensifs.

Les heatmaps normalisées ont révélé que des variables telles que la progression de balle (*carries_prog*), la défense (*tkl_plus_int*, *ball_recov*) ou la création (*per90_ast*) structurent la répartition des joueurs. La faible séparation entre clusters reflète la continuité inhérente des styles de jeu : les transitions entre rôles (ex. ailier $\leftrightarrow$ milieu offensif) sont progressives.

Les tests de stabilité (ARI, NMI) et les indices internes (Silhouette, Calinski-Harabasz, Davies-Bouldin) confirment une structure réelle mais diffuse. L'Agglomerative Clustering se montre légèrement plus robuste, notamment face au bruit.

En conclusion, les approches non supervisées se révèlent pertinentes pour décrire la diversité des profils de joueurs. Malgré la faible séparation, les clusters offrent une lecture fonctionnelle fidèle à la dynamique des rôles dans le sport. L'intégration de données contextuelles (positionnement, intensité, contribution collective) permettrait d'affiner cette typologie.

8 Annexes

8.1 Annexe 1 : Hyper-paramètres des meilleures réductions de dimensions

Jeu de données	Nombre de dimensions	Autres hyper-paramètres
tSNE-raw	4	perplexity=40, learning_rate=60
tSNE-per90	4	perplexity=40, learning_rate=60
tSNE-custom-GK	4	perplexity=40, learning_rate=60
tSNE-custom	4	perplexity=40, learning_rate=60
Isomap-raw	5	n_neighbors=20
Isomap-per90	6	n_neighbors=20
Isomap-custom-GK	6	n_neighbors=25
Isomap-custom	6	n_neighbors=25
UMAP-raw	5	n_neighbors=15, min_dist=0.5
UMAP-per90	5	n_neighbors=10, min_dist=0.5
UMAP-custom-GK	5	n_neighbors=15, min_dist=0.3
UMAP-custom	5	n_neighbors=50, min_dist=0.5

TABLE 3 – Hyper-paramètres des méthodes de réduction de dimension

Références

[1] McInnes, L., Healy, J., & Melville, J. (2018). UMAP : Uniform Manifold Approximation and Projection. arXiv preprint. https://arxiv.org/abs/1802.03426

[2] Coenen, A., Pearce, A. Understanding UMAP. Pair-code. https://pair-code.github.io/understanding-umap/

[3] Kaski, S., Nikkilä, J., Oja, M., Venna, J., Törönen, P., & Castrén, E. (2003). Trustworthiness and metrics in visualizing similarity of gene expression. BMC Bioinformatics, 4(1), 48. https://doi.org/10.1186/1471-2105-4-48