UNIVERSITÉ DE MONTPELLIER

UFR STAPS Montpellier

Projet transversal Data Science Analysis of Football GPS Tracking

Théo Lecomte, IEAP Master, 2SIA Track

Sommaire

I. Introduction

Ce rapport présente l'analyse d'une saison complète de football professionnel à travers l'exploitation de données GPS. L'objectif est de transformer une base de données brute massive en indicateurs de performance exploitables par le staff technique. Le jeu de données comprend plus de **166 millions de lignes** de tracking, couvrant **42 joueurs** sur **187 séances d'entraînement** et **74 matchs**.

II. Méthodologie et Nettoyage

Le traitement des données a suivi un pipeline rigoureux pour garantir la fiabilité des analyses :

Mapping Joueur-Capteur

En raison des échanges fréquents de capteurs entre les joueurs, une jointure dynamique a été réalisée entre les données de tracking et un fichier de correspondance (summary.csv) basé sur les dates de sessions.

Pseudonymisation

Pour respecter la confidentialité, chaque joueur a été transformé en identifiant numérique unique.

Nettoyage des données aberrantes

Pour traiter ces données, nous nous sommes appuyés sur les principes méthodologiques de détection des valeurs aberrantes détaillés dans les travaux d'**Écologie Mathématique** (référence : Essicolo - Valeurs et échantillons aberrants)

Deux algorithmes ont été implémentés pour traiter les erreurs de signal:

Position

Le signal GPS présenter des "sauts" de position, où le capteur semble téléporter le joueur à plusieurs mètres de sa position réelle en une fraction de seconde.

La fonction detect_outlier_sequences_position_z traite les anomalies de trajectoire. Contrairement à une vérification unidimensionnelle, elle évalue l'écart du joueur dans l'espace 2D du terrain.

- **Logique Technique :**
    1. **Standardisation :** Pour chaque coordonnée (x et y), on calcule le score Z ($Z = \frac{x-\mu}{\sigma}$).
    2. **Distance Euclidienne des Z-Scores :** On calcule une métrique de distance combinée : $Z_{dist} = \sqrt{Z_x^2 + Z_y^2}$.
    3. **Seuil Statistique :** Un seuil de 3 est appliqué. Si Z_dist > 3, la position est statistiquement "impossible" par rapport à la distribution moyenne de la session (couvrant 99,7% des positions normales).
- **Identification de Séquences :** L'algorithme ne se contente pas de marquer des points isolés ; il identifie des **séquences continues** d'aberrations (en marquant un outlier_start et un outlier_end). Cela permet de détecter des périodes de "dérive GPS" où le capteur perd sa précision pendant plusieurs secondes.

Vitesse

La fonction detect_local_outliers_speed_iqr_global_iqr s'attaque aux pics de vitesse irréalistes. Elle utilise une approche plus robuste que la moyenne : l'Écart Interquartile (IQR).

- **Logique Technique :**
    1. **Seuil de Déclenchement (Trigger) :** L'analyse se concentre d'abord sur les pics dépassant **35 km/h** (vitesse limite proche du sprint de classe mondiale).
    2. **Analyse de Fenêtre Temporelle :** Pour chaque pic suspect, la fonction examine une fenêtre de **5 secondes** avant et après l'événement.
    3. **Validation par l'IQR Global :** À l'intérieur de cette fenêtre, chaque point est comparé aux barrières globales de la session :
        - Q1 (25ème percentile) et Q3 (75ème percentile).
        - IQR = Q3 - Q1.
        - Barrière supérieure = Q3 + 1.5 * IQR
- **Pourquoi cette méthode ?** En couplant un déclencheur de vitesse maximale et une validation par IQR, on évite de supprimer de vrais sprints. Si une vitesse est haute mais cohérente avec l'effort global, elle est conservée. Si elle dépasse les limites de l'IQR dans une fenêtre de pic soudain, elle est marquée comme is_outlier_speed.

Synthèse du Traitement

En marquant le début et la fin de chaque séquence aberrante (position ou vitesse), nous préparons la base de données pour une étape d'**imputation**. Plutôt que de simplement supprimer les lignes (ce qui briserait la chronologie de la session), ces segments sont isolés pour être corrigés par interpolation linéaire, garantissant ainsi des calculs de distance totale et de charge de travail (Training Load) non biaisés par des artefacts techniques.

III. Analyse Individuelle : Profilage Athlétique

L'analyse repose sur le calcul de métriques cinématiques avancées :

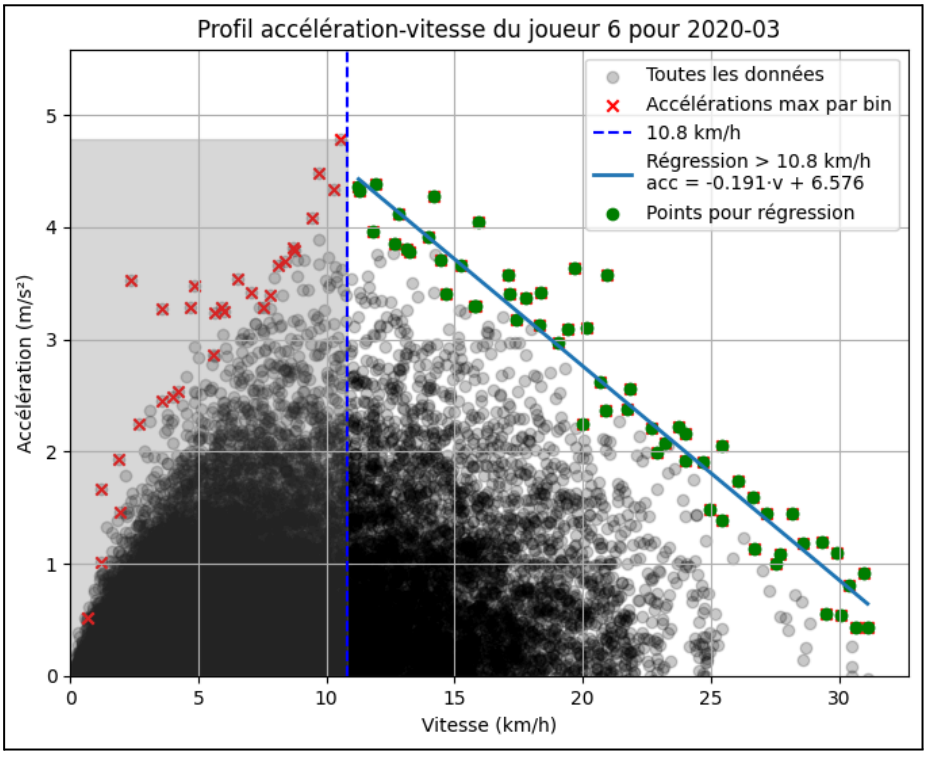
Zones de Vitesse

Sept zones ont été définies en fonction de la vitesse maximale (Vmax) du joueur, allant de la marche (0-40% de Vmax) au sprint de haute intensité (90-100% de Vmax).

	zone	min_speed_kmh	max_speed_kmh
0	0-40%	0.000000	12.444048
1	40-50%	12.444048	15.555060
2	50-60%	15.555060	18.666072
3	60-70%	18.666072	21.777084
4	70-80%	21.777084	24.888096
5	80-90%	24.888096	27.999108
6	90-100%	27.999108	31.110120

Profil Accélération-Vitesse

Nous avons modélisé le profile d'accélération maximale par zone de vitesse via une régression linéaire, permettant d'identifier le profil de puissance de chaque joueur.



Métriques de Performance

Nous isolons les indicateurs de performance clés.

- Vitesse Maximale
- Accélération et Décélération Maximale
- Changements de Direction
- Nombre d'accélération : Modérées de sup à 2.5 m/s^2 et Intense sup à 3.5 m/s^2.
- Nombre de décélération : Modérées de inf à -2.5 m/s^2 et Intense inf à -3.5 m/s^2.
- Temps passé dans zone de vitesse (%)
- Temps joué
- Training Load

	max_speed	max_distance	max_acceleration	max_decceleration	max_direction_changes
0	24.23520	6072.959619	4.411	-5.010	2254
1	25.52328	5250.231637	4.807	-6.172	4234
2	27.19692	3145.992339	4.920	-5.454	2943
3	28.58580	8107.300847	4.782	-5.522	11597

Nb_acc_2_5	Nb_acc_3_5	Nb_dcc_2_5	Nb_dcc_3_5	time_played_hms	training_load
57	6	93	9	00:45:25	1.68
91	13	120	30	00:45:18	1.40
59	9	96	18	00:26:59	1.39
115	19	136	36	01:27:16	1.24

time_in_zone_0-40%	time_in_zone_40-50%	time_in_zone_50-60%	time_in_zone_60-70%	time_in_zone_70-80%	time_in_zone_80-90%	time_in_zone_90-100%
65.39	15.11	10.47	5.50	2.41	0.81	0.32
79.84	9.00	5.83	3.26	1.32	0.60	0.16
80.90	8.61	5.20	2.38	1.94	0.75	0.22
86.51	6.63	4.00	1.88	0.73	0.21	0.03

Charge d'entraînement (Training Load)

Un score de charge a été calculé en pondérant le temps passé dans chaque zone de vitesse, offrant une vision synthétique de l'effort fourni.

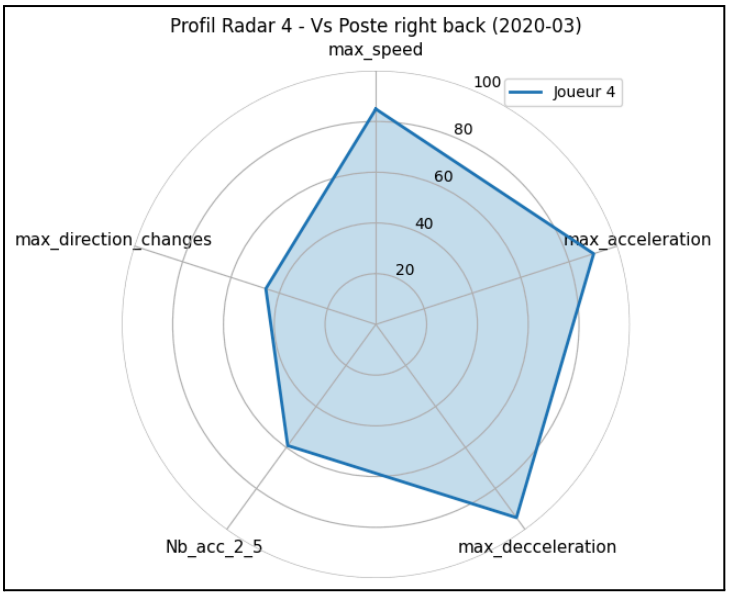
Radar

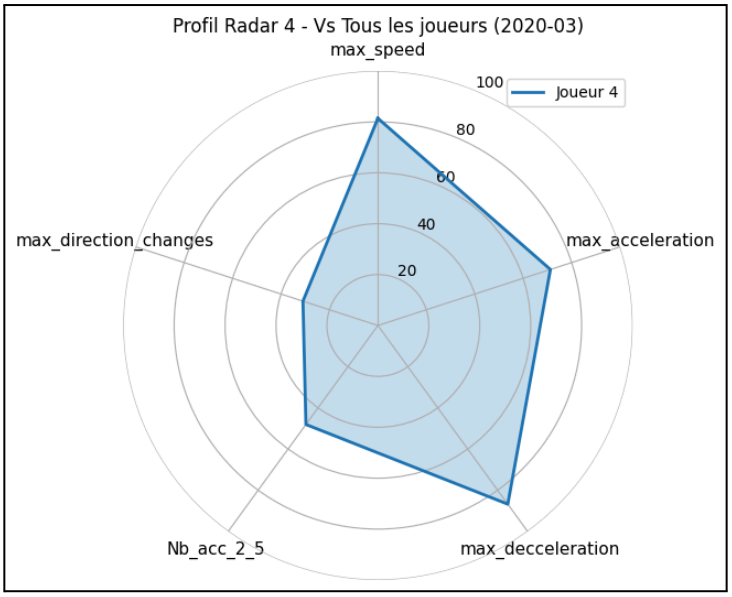
Nous avons généré des graphiques en radar qui normalisent les métriques clés. Cela permet de visualiser les "signatures" athlétiques propres à chaque poste.

Les Métriques Sélectionnées pour le Radar

Pour que la comparaison soit pertinente, nous avons retenu 6 axes majeurs :

1. Vitesse Max : Capacité de pointe.
2. Distance / Minute : Intensité globale du volume de course.
3. Accélérations High-Int (>3.5 m/s^2) : Explosivité offensive ou défensive.
4. Décélérations High-Int (< -3.5 m/s^2) : Capacité de freinage et de changement de direction.
5. Nb de Sprints : Capacité à répéter les efforts de haute intensité.





IV. Analyse Collective : Focus sur le Match

Le football est un sport de relations spatiales. Pour analyser la performance collective, nous avons croisé les données de performance athlétique avec des outils de visualisation dynamique.

Métriques de Performance

Récupération des métriques de performance indivduelle pour chaque joueur afin de visualisez tous les joueurs

Top 5 des Performances Match

Permet de visualiser rapidement le top 5 des joueurs ayant fait le plus de distance, la plus haute vitesse et la plus grande accélération.

Visualisation Dynamique

La génération d'une séquence vidéo (exportée en format MP4 via matplotlib.animation) permettant de visualiser les déplacements des joueurs en temps réel sur une fenêtre temporelle choisie.

Méthodologie technique : En utilisant les coordonnées (x, y) nettoyées, nous avons recréé une vue "bird-eye" (vue de dessus) du terrain. Chaque point représente un joueur, identifié par son numéro de capteur.

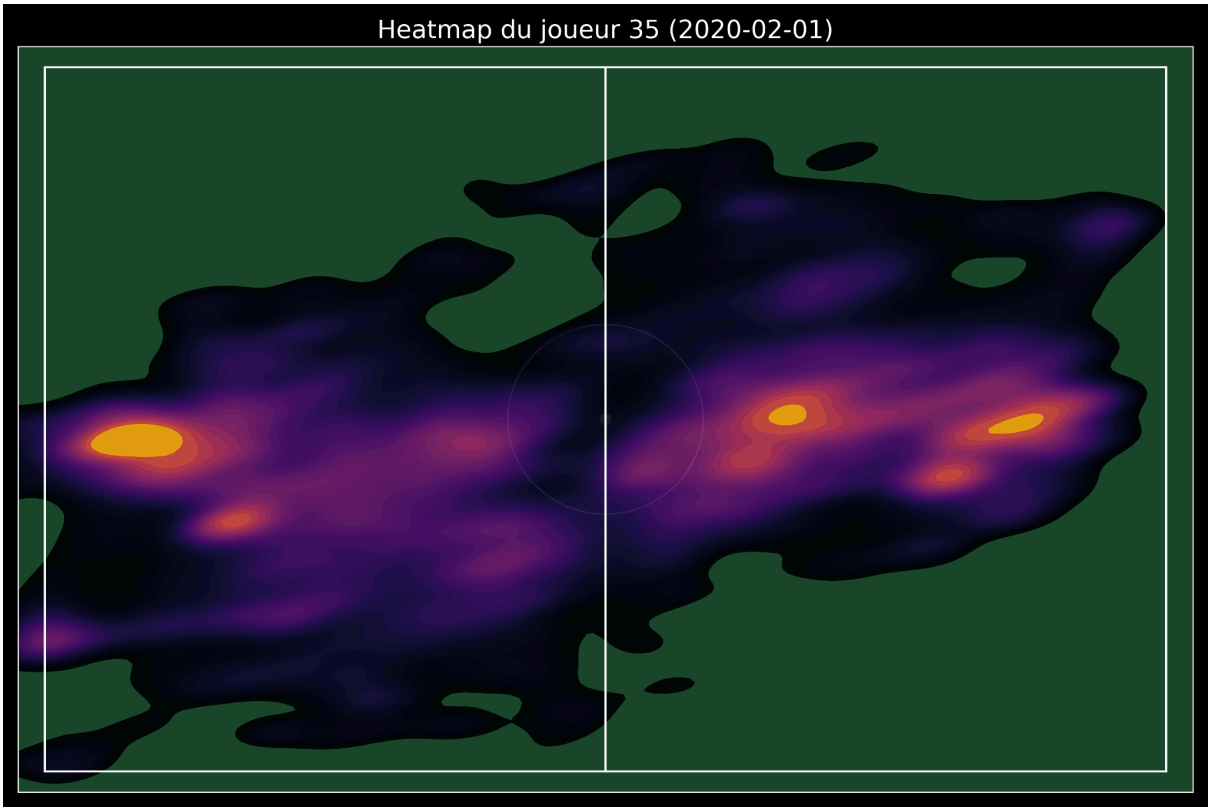
Analyse par poste : La vidéo permet de segmenter les joueurs par couleur selon leur ligne (Défenseurs, Milieux, Attaquants).

Utilité tactique : Cette animation permet d'observer :

- **Le bloc équipe :** La capacité des joueurs à rester compacts lors des phases défensives.
- **La profondeur :** La synchronisation des appels des attaquants avec le porteur de balle.
- **L'étirement du bloc :** Les distances entre les lignes (Back-Mid-Forward) lors des transitions rapides.

Heatmaps

En complément de la vidéo, nous avons généré des **Heatmaps** pour un joueur spécifique.



V. Axes d'amélioration et Perspectives

Plusieurs pistes d'évolution technique et méthodologique ont été identifiées pour rendre l'outil encore plus opérationnel pour le staff technique.

Optimisation de la Visualisation Dynamique (Vidéo)

Actuellement, la visualisation MP4 génère un flux continu qui reste difficile à exploiter pour un entraîneur.

- **Ciblage événementiel :** L'amélioration prioritaire consiste à coupler les données GPS avec les événements du match (données d'Opta ou StatsBomb). En isolant les "timestamps" des buts encaissés ou marqués, nous pourrions générer automatiquement des clips de 30 secondes pour analyser le placement défensif ou l'animation offensive lors de moments clés.

- **Analyse de la ligne défensive :** L'ajout de segments (lignes de connexion) entre les défenseurs sur la vidéo permettrait de visualiser instantanément la distance entre les joueurs.

Machine Learning et Analyse Tactique

L'utilisation de modèles d'apprentissage automatique permettrait de passer d'une analyse descriptive à une analyse prédictive :

- **Reconnaissance de formes (Pattern Recognition) :** En utilisant des algorithmes de *Clustering*, nous pourrions identifier automatiquement la formation tactique réelle (ex: 4-4-2 vs 4-3-3) qui peut différer de la feuille de match.
- **Robustesse vs Volatilité :** Un modèle pourrait mesurer le "temps de maintien" d'une structure. Une équipe robuste maintient ses distances entre les lignes malgré la pression, tandis qu'une équipe volatile se désorganise plus vite. Cela permettrait de quantifier la discipline tactique.

Évolution de la Heatmap : Le Découpage par Zone

La heatmap "nuage" actuelle est esthétique mais parfois floue pour un coach.

- **Zones d'intérêt :** Nous proposons de découper le terrain en zones spécifiques (ex: les 18 zones de jeu classiques ou les "half-spaces").
- **Métrique d'occupation :** Au lieu d'une densité de points, nous afficherions un pourcentage de temps passé dans chaque zone. Savoir qu'un ailier passe 60% de son temps dans les 30 derniers mètres est une information tactique plus directe.

Score de Risque de Blessure (Injury Risk Score)

Nous pourrions développer un score prédictif basé sur le croisement de plusieurs facteurs :

- **Profil de fatigue :** Une chute soudaine des métriques de performance (vitesse max ou hauteur de saut) sur plusieurs sessions est souvent un signe avant-coureur d'une lésion musculaire.
- **Indicateurs de stress mécanique :** Une corrélation entre un *Training Load* élevé et un volume de changements de direction (COD) agressifs

permettrait de définir un score d'alerte (Rouge/Orange/Vert) pour chaque joueur avant l'entraînement.

Conclusion

Ce travail m'a permis de développer ma curiosité ainsi que des compétences approfondies en data science. Le plus difficile a été de trouver un bon algorithme de détection des outliers ; je dirais qu'actuellement la détection est quasi parfaite, même si l'imputation pourrait encore être améliorée.

Le manque de temps a été crucial dans ce rapport. Avec un ou deux mois de plus, j'aurais pu développer davantage les axes d'amélioration, notamment sur les modèles de machine learning. Je compte continuer ce projet sur GitHub après le devoir afin de tester mes idées et mes hypothèses.

VI. Bibliographie et Ressources

- **Cédric Noël**, *Profils Force-Vitesse et Accélération-Vitesse : deux outils de mesures de la performance*.
- **Essicolo**, *Valeurs et échantillons aberrants : méthodes de détection et traitement*.
- Documentation technique de **DuckDB** et **Matplotlib Animation**.