UNIVERSITÉ DE MONTPELLIER

UFR STAPS Montpellier

Projet transversal
Data Science Analysis of Football GPS Tracking

Théo Lecomte, IEAP Master, 2SIA Track

Summary

I. Introduction

This report presents the analysis of a full professional football season through the use of GPS data. The goal is to transform a massive raw database into performance indicators that can be utilized by the coaching staff. The dataset includes over 166 million tracking rows, covering 42 players across 187 training sessions and 74 matches.

II. Methodology and Data Cleaning

The data processing followed a rigorous pipeline to ensure the reliability of the analyses:

Player-Sensor Mapping

Due to frequent sensor exchanges between players, a dynamic join was performed between the tracking data and a correspondence file (summary.csv) based on session dates.

Pseudonymization

To ensure confidentiality, each player was assigned a unique numeric identifier.

Outlier Data Cleaning

To handle these data, we relied on methodological principles for outlier detection detailed in the field of Mathematical Ecology (reference: [Essicolo - Valeurs et échantillons aberrants]())

Two algorithms were implemented to handle signal errors:

Position

The GPS signal sometimes shows "position jumps," where the sensor appears to teleport the player several meters from their actual location in a fraction of a second.

The function `detect_outlier_sequences_position_z` addresses trajectory anomalies. Unlike a one-dimensional check, it evaluates the player's deviation in the 2D space of the pitch.

- **Technical Logic:**
    1. **Standardization:** For each coordinate (x and y), a Z-score is calculated. $(Z = \frac{x-\mu}{\sigma})$.
    2. **Euclidean Distance of Z-Scores:** A combined distance metric is computed. $Z_{dist} = \sqrt{Z_x^2 + Z_y^2}$.
    3. **Statistical Threshold:** A threshold of 3 is applied. If `Z_dist > 3`, the position is statistically "impossible" relative to the session's mean distribution (covering 99.7% of normal positions).

- **Sequence Identification:** The algorithm does not merely flag isolated points; it identifies continuous sequences of anomalies (marking an `outlier_start` and an `outlier_end`). This allows detection of periods of "GPS drift," where the sensor loses precision for several seconds.

Speed

The function `detect_local_outliers_speed_iqr_global_iqr` addresses unrealistic speed spikes. It uses a more robust approach than the mean: the Interquartile Range (IQR).

- **Technical Logic:**
    1. **Trigger Threshold:** The analysis first focuses on spikes exceeding 35 km/h (a speed close to world-class sprint levels).
    2. **Time Window Analysis:** For each suspicious spike, the function examines a 5-second window before and after the event.
    3. **Validation via Global IQR:** Within this window, each point is compared to the session's global thresholds:
        - Q1 (25th percentile) and Q3 (75th percentile)
        - IQR = Q3 – Q1
        - Upper threshold = Q3 + 1.5 * IQR

- **Why this method?** By combining a maximum speed trigger with IQR-based validation, we avoid removing genuine sprints. If a speed is high but consistent with the overall effort, it is retained. If it exceeds the IQR limits within a sudden spike window, it is flagged as `is_outlier_speed`.

Data Processing Summary

By marking the start and end of each anomalous sequence (position or speed), we prepare the database for an imputation step. Rather than simply deleting the rows (which would break the session chronology), these segments are isolated to be corrected via linear interpolation, thereby ensuring that total distance and training load calculations are not biased by technical artifacts.

III.Individual Analysis: Athletic Profiling

The analysis is based on the calculation of advanced kinematic metrics:

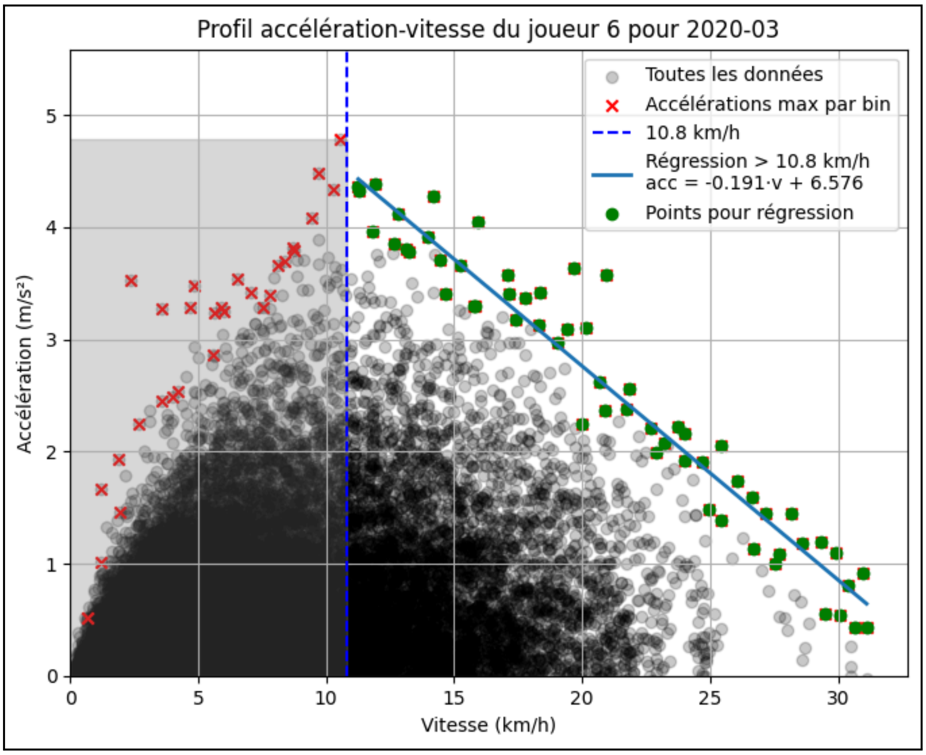
Speed Zones

Seven zones were defined according to the player's maximum speed (Vmax), ranging from walking (0–40% of Vmax) to high-intensity sprinting (90–100% of 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

Acceleration-Speed Profile

We modeled the maximum acceleration profile by speed zone using linear regression, allowing us to identify each player's power profile.



Performance Metrics

We isolate the key performance indicators:

- **Maximum Speed**
- **Maximum Acceleration and Deceleration**
- **Changes of Direction**
- **Number of Accelerations:** Moderate > 2.5 m/s^2 and High > 3.5 m/s^2
- **Number of Decelerations:** Moderate < -2.5 m/s^2 and High < -3.5 m/s^2
- **Time in Speed Zones (%)**
- **Time played**
- **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

Training Load

A load score was calculated by weighting the time spent in each speed zone, providing a synthetic overview of the effort exerted.

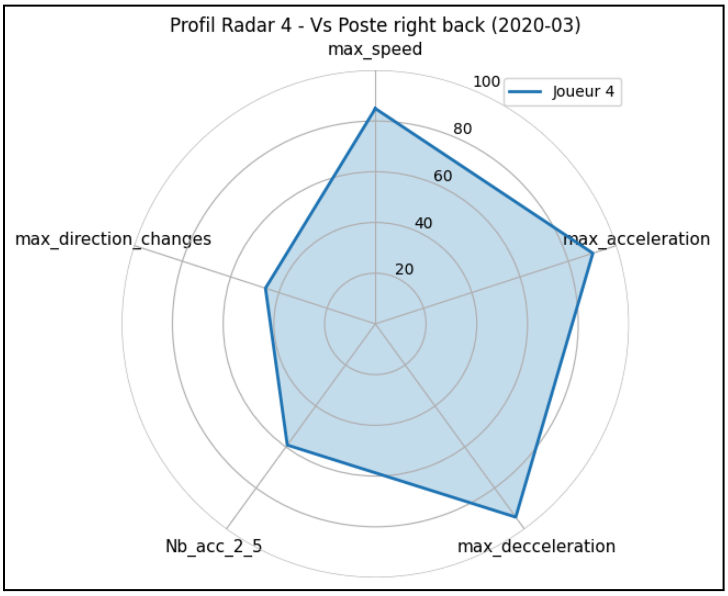
Radar

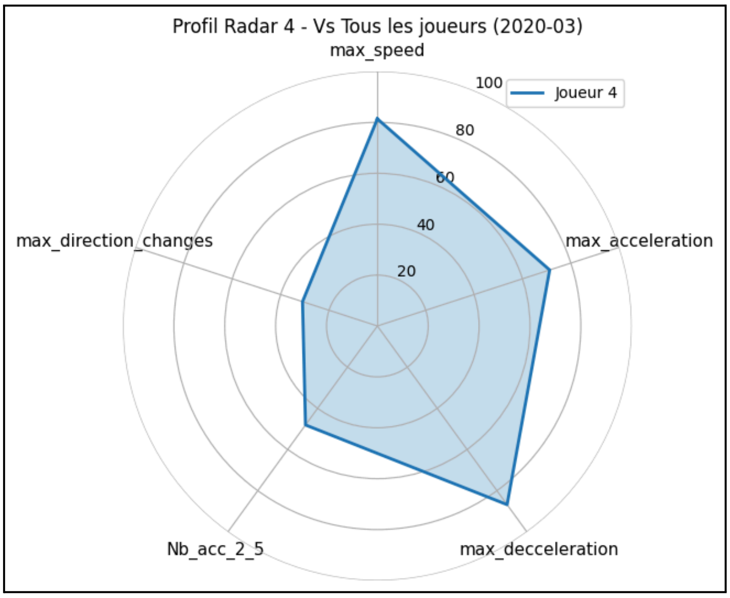
We generated radar charts that normalize key metrics. This allows visualization of the athletic "signatures" specific to each position.

Selected Metrics for the Radar

To ensure meaningful comparisons, we retained six major axes:

- **Maximum Speed:** Peak capacity.
- **Distance per Minute:** Overall running intensity.
- **High-Intensity Accelerations (>3.5 m/s^2):** Offensive or defensive explosiveness.
- **High-Intensity Decelerations (<-3.5 m/s^2):** Braking and change-of-direction ability.
- **Number of Sprints:** Ability to repeat high-intensity efforts.





IV. Collective Analysis: Match Focus

Football is a sport of spatial relationships. To analyze collective performance, we combined athletic performance data with dynamic visualization tools.

Performance Metrics

Individual performance metrics were retrieved for each player to visualize all players.

Top 5 Match Performances

Allows quick visualization of the top 5 players in terms of distance covered, highest speed, and greatest acceleration.

Dynamic Visualization

A video sequence (exported in MP4 format via `matplotlib.animation`) was generated to visualize player movements in real time over a chosen time window.

Technical Methodology: Using the cleaned (x, y) coordinates, we recreated a bird's-eye view of the pitch. Each point represents a player, identified by their sensor number.

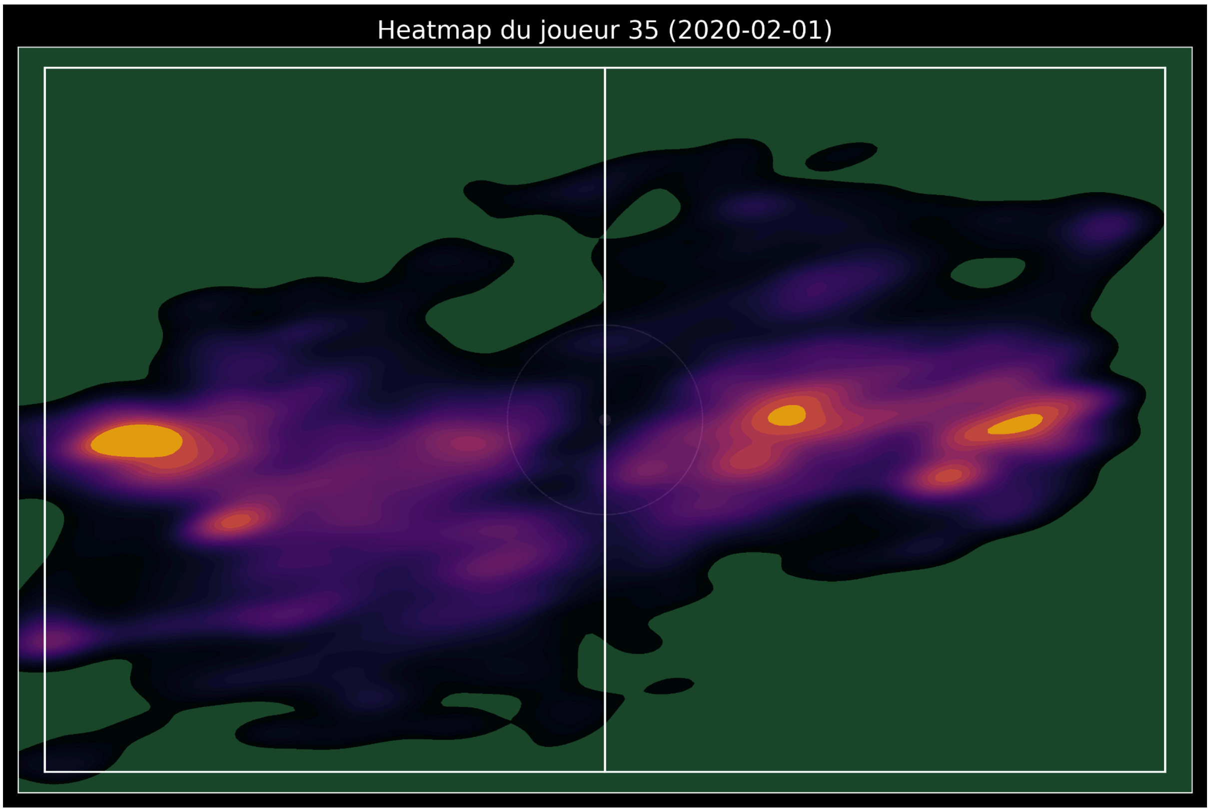
Position-Based Analysis: The video allows players to be segmented by color according to their line (Defenders, Midfielders, Forwards).

Tactical Use: This animation allows observation of:

- **Team Block:** The players' ability to remain compact during defensive phases.
- **Depth:** The synchronization of attackers' runs with the ball carrier.
- **Block Stretching:** The distances between lines (Back–Midfield–Forward) during fast transitions.

Heatmaps

In addition to the video, we generated heatmaps for a specific player.



V. Areas for Improvement and Future Perspectives

Several technical and methodological development avenues have been identified to make the tool even more operational for the coaching staff.

Optimization of Dynamic Visualization (Video)

Currently, the MP4 visualization generates a continuous stream, which remains difficult for a coach to use effectively.

- **Event Targeting:** The primary improvement involves combining GPS data with match events (Opta or StatsBomb data). By isolating the timestamps of goals conceded or scored, we could automatically generate 30-second clips to analyze defensive positioning or offensive movement during key moments.
- **Defensive Line Analysis:** Adding segments (connection lines) between defenders in the video would allow instant visualization of the distances between players.

Machine Learning and Tactical Analysis

The use of machine learning models would allow a shift from descriptive to predictive analysis:

- **Pattern Recognition:** Using clustering algorithms, we could automatically identify the team's actual tactical formation (e.g., 4-4-2 vs 4-3-3), which may differ from the match sheet.

- **Robustness vs. Volatility:** A model could measure the "maintenance time" of a structure. A robust team maintains distances between lines under pressure, while a volatile team disorganizes more quickly. This would allow quantification of tactical discipline.

Heatmap Evolution: Zone-Based Segmentation

The current "cloud" heatmap is visually appealing but can sometimes be unclear for a coach.

- **Zones of Interest:** We propose dividing the pitch into specific zones (e.g., the 18 classic playing zones or the "half-spaces").

- **Occupancy Metric:** Instead of point density, we would display the percentage of time spent in each zone. Knowing that a winger spends 60% of their time in the final 30 meters provides more direct tactical insight.

Injury Risk Score

We could develop a predictive score based on the combination of several factors:

- **Fatigue Profile:** A sudden drop in performance metrics (maximum speed or jump height) over multiple sessions is often an early warning sign of a muscle injury.

- **Mechanical Stress Indicators:** A correlation between high training load and a high volume of aggressive changes of direction (COD) could be used to define an alert score (Red/Orange/Green) for each player before training

Conclusion

This work allowed me to develop my curiosity as well as advanced data science skills. The most challenging part was finding an effective outlier detection algorithm; I would say that the detection is currently almost perfect, although the imputation could still be improved.

The lack of time was a critical factor in this report. With an extra month or two, I could have further developed the areas for improvement, particularly regarding machine learning models. I plan to continue this project on GitHub after the assignment to test my ideas and hypotheses.

VI. References and Resources

- **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*.