Analyse d'une compétition mondiale de football féminin par Process Mining

Laly Lacroix, Julie Treilhou, Emmanuelle Claeys, Sébastien Dejean

Analyse d'une compétition mondiale de football féminin par Process Mining.

Laly Lacroix [1] & Julie Treilhou [1] & Emmanuelle Claeys [2] & Sébastien Déjean [3]

[1] *INSA Toulouse, France*
[2] *Unviversité Paul Sabatier, IRIT, Toulouse, France*
[3] *Institut de Mathématiques de Toulouse, Toulouse, France*

Résumé. Cet article présente un retour d'expérimentation d'analyse statistique sur des données issues de la *FIFA Women's World Cup 2023*, coupe du monde de football. Contrairement aux indicateurs traditionnels développés en statistique sportive, notre approche intègre l'utilisation d'outils de *process mining*, une méthodologie rarement explorée dans le domaine du sport jusqu'à présent. Cette démarche nous permet d'analyser les processus sous-jacents de manière approfondie, offrant ainsi une perspective nouvelle pour analyser la performance d'une équipe. Les données présentées ici proviennent de la Coupe du monde féminine 2023 représentant des équipes de haut niveau et donnant une visibilité au football féminin. L'intégralité des données sont disponibles sur le site StatsBomb[1], l'une des principales source dans l'*open data* sportif. Les résultats obtenus ont permis une cartographie des trajectoires des tirs ainsi qu'une mise en évidence des schémas tactiques, grâce aux méthodes de *process mining*. Ils offrent également la possibilité d'observer l'évolution tactique de l'équipe victorieuse, en l'occurrence l'Espagne, depuis son premier match jusqu'au dernier.

Mots-clés. statistique sportive, event log, sport féminin.

Abstract. This article presents the results of a statistical analysis experiment carried out on the Women's World Cup 2023 competition. Unlike traditional indicators developed in sports statistics, our approach are based on *process mining* tools. This approach enables us to analyze underlying processes in plays, offering a new point of view on the analysis of sports performance. The data presented here representing both top-level teams and giving visibility to women's soccer. All the data is available on the StatsBomb website, one of the biggest open-source data platforms. The results provide map shot trajectories and highlight tactical patterns, using *process mining* methods. It also highlights the tactical evolution of the winning team, in this case Spain, from its first to its last match.

Keywords. Sports Statistics, event log, women's sport

[1] https://statsbomb.com

1 Introduction

L'analyse statistique appliquée au football représente une discipline évolutive et cruciale dans la compréhension approfondie des performances sportives [3]. L'intérêt grandissant pour l'analyse de données dans de nombreuses disciplines permet d'adapter au mieux l'entraînement [4]. Dans le cadre du football, les indicateurs doivent permettre notamment d'extraire les dynamiques tactiques des équipes. Les données sportives présentent des caractéristiques particulièrement intéressantes :

- D'une part les observations utilisées sont parfois issues d'un seul match, le contexte (joueurs présents ce jour là, équipe et période considérée) étant fortement différent d'un match a un autre
- D'autre part les observations peuvent se présenter sous la forme d'un *event log*, c'est-à-dire un jeu de données enregistrant un ensemble d'événements temporels.

Ces propriétés peuvent être en contradiction avec les outils nécessitant des hypothèses telles que des variables indépendantes et identiquement distribuées (i.i.d.) ou un volume substantiel de données. Parmi les métriques fréquemment étudiées, on retrouve la possession de balle, les tirs au but, la précision des passes, ainsi que les statistiques individuelles des joueuses, telles que les dribbles réussis ou encore les interceptions. Bien que ces indicateurs permettent d'évaluer la performance globale d'une équipe, pour enrichir davantage cette analyse, nous proposons d'utiliser une approche novatrice telle que le *process mining* pour obtenir une compréhension plus fine des schémas de jeu, de transitions entre phases de jeu, et des variations tactiques au sein d'une équipe tout au long de la compétition. La première section présente la modélisation de l'*event log* par rapport aux données utilisées, la seconde partie présente les résultats obtenus suite à la comparaison entre deux matchs (ouverture : Costa - Rica vs Espagne et finale : Angleterre vs Espagne). La dernière section conclut sur nos résultats. Nous utilisons principalement la librairie BupaR [2] (codée en `R`), l'intégralité du code pour reproduire les expériences est disponible sur un dépôt GitHub[2].

2 Modélisation

L'exploration de processus implique des méthodes d'analyse de *process* représentés par des modèles à partir de *event logs* (c'est-à-dire les données réelles émises, supposées suivre le processus observé) [1]. Chaque événement de l'*event logs* est composé de trois informations : un identifiant de cas, un horodatage et une activité. Grâce à ces informations, un événement (cas) rapporte plusieurs actions réalisées (activités) différenciées par un marqueur temporel (horodatage). Les activités sont rattachées à un même évènement pour

[2]https://github.com/julietrlh/StatbombR

former une séquence, ainsi une séquence individuelle est appelée *trace*. Toutes les traces possibles sont représentées sous forme de DFG (*Direct Follower Graph*). Ce graphique décrivant tous les parcours possibles est appelé *process map*.

Un exemple d'*event log* traditionnellement utilisé pour le *process mining*, serait par exemple un listing de patients (évènements) dont les interventions représenteraient les activités et leur enregistrements l'horodatage. La problématique pour appliquer le *process mining* aux données issues d'un match de football est de définir ce que seraient un évènement et une activité (l'horodatage étant naturellement l'enregistrement du temps d'une action observée pendant le match). Ce découpage est une question nécessitant une attention particulière : un découpage trop long engendre une perte de granularité, risquant ainsi d'occulter des nuances significatives dans le déroulement du jeu. D'autre part, un découpage excessivement court peut conduire à une fragmentation excessive des données, rendant difficile la capture des schémas de jeu plus larges et des dynamiques stratégiques. Ainsi, la détermination d'une unité de temps appropriée pour le découpage des événements demeure un enjeu essentiel dans l'analyse statistique du football. Plusieurs découpages ont été réalisés dans notre travail : (1) Découpage chronologique : on considère qu'un événement est délimité par un intervalle de temps, tel que, par exemple, toutes les 5 minutes. (2) Découpage post-activité : on considère qu'un événement est composé des X activités précédant un évènement spécifique, tel qu'un but. Si l'analyse se focalise sur un évènement spécifique, c'est plutôt le second découpage qui sera utilisé.

Dans les exemples qui sont traités par la suite, des représentations graphiques différentes sont proposées pour mettre en évidence et caractériser les types d'actions qui suivent ou précèdent un évènement spécifique. Ces représentations sont obtenues par des analyses de graphes ou des analyses exploratoires classiques sur les évènements et leurs durées respectives.

3 Comparaison de deux matchs pour l'équipe d'Espagne

3.1 Process map des actions précédant un but

Nous rapportons ici, par souci d'analyse rapide et d'interprétation pour un lecteur non familiarisé au football, un découpage post-activités composé des deux activités précédant un but. Pour le match d'ouverture 3-0, seulement deux buts ont été étudiés, l'autre but étant marqué par le Costa Rica contre son camp en faveur de l'Espagne. Pour le match terminal un seul but a été marqué en faveur de l'Espagne. Par conséquent, l'objectif de l'analyse ici présentée n'est pas d'inférer le déroulement d'un match futur mais plutôt de comparer deux matchs dont le niveau de l'équipe adversaire diffère fortement. Nous rappelons ici que la *process map* représente l'ensemble des séquences menant à un but sous forme d'un *Direct Follower Graph*. Nos résultats ont montré qu'il n'y avait pas de différence significative entre les séquences d'actions des deux matchs.

Lors du match d'ouverture, la complexité de la *process map* de gauche de la figure 1

Figure 1: *Process map* illustrant la séquence des deux dernières actions menant à un tir réussi pour l'Espagne lors de son premier match (à gauche) et de la finale (à droite).

indique plusieurs opportunités pour marquer un but. Deux sous-ensembles d'actions principaux conduisant à un but se dégagent. Pour la combinaison Dribble-Carry-Shot, il s'agit d'une opportunité de tir direct, comme cela a été observé dans d'autres matchs. La combinaison Ball Recovery-Shot, représente plusieurs tentatives précédant un tir réussi. Plus précisément, 16,67% des activités répertoriées impliquent des tentatives de reprise de possession du ballon par une joueuse adverse. Dans tous les cas où cette tentative a abouti, elle est suivie d'un contrôle du ballon au niveau du pied de la joueuse en mouvement, ce qui conduit systématiquement à un tir. En outre, un cas isolé montre qu'un tir sur trois est suivi d'une tentative de reprise de possession, ce qui indique que, lorsqu'une joueuse adverse intercepte le ballon après un tir dans la zone adverse, cela conduit toujours à un autre tir. En revanche, le *process map* de droite est linéaire en raison de l'unique but marqué lors de la finale. Le schéma de jeu gagnant, qui est très simple, indique qu'une passe bien reçue a été suivie par un tir réussi. Cette comparaison met en évidence :

- une diminution des occasions de but entre les deux matchs, en raison du niveau de l'équipe adverse.
- la variété des schémas d'actions dans le premier match qui reflète la diversité des situations de but potentielles, alors que les schémas d'actions (en incluant les buts manqués) sont quasi identiques pour le dernier match. Cela peut s'expliquer par une efficacité du jeu dans un contexte de pression intense et de compétition décisive.

Ces résultats, évidents du point de vue du football, illustrent l'intérêt d'une démarche de *process mining* qui permet d'étudier des séquences de jeu menant à des évènements spécifiques au cours d'un match.

3.2 Actions menant le ballon hors du terrain

Nos analyses sur les sorties de terrain (Fig. 2) ont montré qu'une plus grande variété de types d'actions conduit le ballon hors des limites du terrain pendant la finale. Les passes représentent 17% des actions précédant les sorties de balle dans le match terminal contre 11% dans le match d'ouverture. Par ailleurs, la stratégie de l'équipe diffère entre l'Angleterre qui incite l'Espagne à commettre davantage de fautes, devenant la deuxième

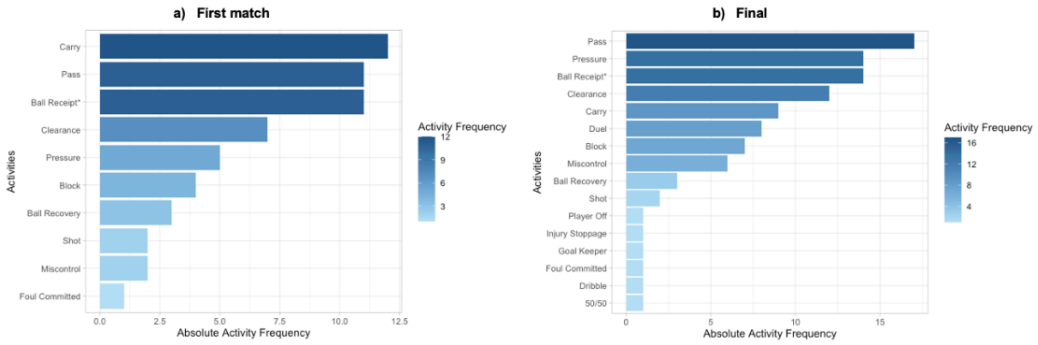

Figure 2: Bar Chart illustrant la distribution de fréquence des actions de l'Espagne menant à une sortie de terrain lors du premier match de l'Espagne (à gauche) et de la finale (à droite).

cause des sorties de ballons, soit 14% des actions précédentes, contre seulement 5% lors du match contre le Costa Rica. Cette évolution suggère une intensification de la défense anglaise, mettant en évidence une pression accrue sur l'équipe espagnole et un ajustement stratégique de sa part pour contrer cette pression accrue.

3.3 Actions suivant la remise en jeu par la gardienne de but

L'analyse comparative des deux graphiques de la figure 3 se concentre sur les deux actions qui suivent un coup de pied par la gardienne. On s'intéresse donc ici aux types d'évènements ainsi qu'à leur durée. Cette analyse révèle des différences significatives entre le premier match et la finale.

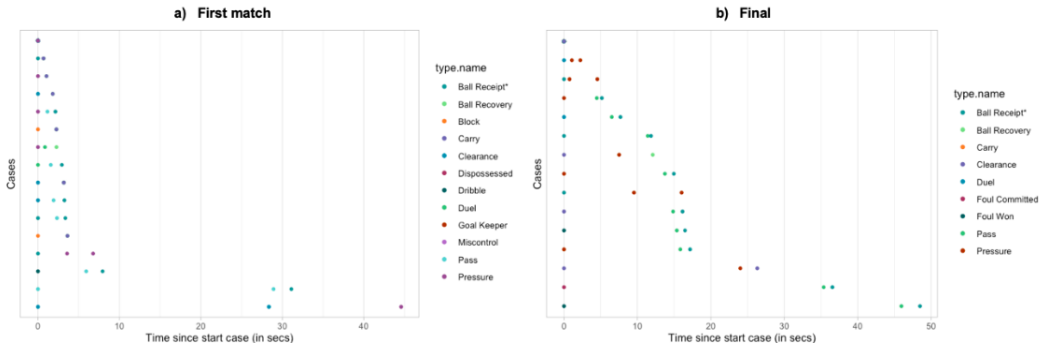

Figure 3: Graphique illustrant les quatre types d'actions consécutives à un coup de pied de la gardienne espagnole lors du premier match (à gauche) et de la finale (à droite). Chaque case est un évènement spécifique.

Contrairement à d'autres études d'évènements, les durées d'évènements sont similaires pour les deux matches (globalement inférieurs à 15 sec). Ici 75% des évènements des actions se produisent dans les 5 secondes qui suivent le coup de pied de la gardienne, avec une extension jusqu'à 40 secondes pour les cas les plus longs. Les outliers du graphique associé au premier match (4 points les plus à droite sur le graphique de gauche) correspondent aux premières actions se produisant près de 30 secondes après le début de la remise en jeu. A l'inverse, dans le match terminal (à droite) environ un tiers des évènements voient leur première action se produire dans les quinze secondes, tandis que les deux tiers restants ont leur première action entre 35 et 45 secondes après le coup de pied. Cette analyse met en évidence des dynamiques temporelles distinctes entre les deux matches, suggérant des variations dans les schémas de jeu et l'évolution des situations après les remises en jeu de la gardienne, notamment pour la finale suggérant un changement de stratégie à des moments clés spécifiques du match.

4 Conclusion

En conclusion, l'analyse des performances de l'Espagne lors de la Coupe du monde de football féminin à travers l'exploration des process, notamment entre les matchs initial et final, permet de contribuer à comprendre le parcours de l'équipe. L'examen des résultats des tirs met en évidence les variations dans les stratégies offensives, en soulignant la réaction face à la défense ainsi que les actions de la gardienne de but. La finale est marquée par une défense agressive de l'Angleterre. En réaction, l'équipe d'Espagne a mis en place des ajustements tactiques et des schémas de jeu potentiellement plus prévisibles en finale par rapport au jeu variable contre le Costa Rica. Cette dynamique contrastée du premier match contre le Costa Rica, caractérisée par des tentatives de tir différées, met en évidence la capacité d'adaptation de l'Espagne. La diminution des occasions de but en finale se présente sous la forme d'un *processus simplifié*. Le changement des temps d'évènements, en particulier après les coups de pied de but, suggère des ajustements stratégiques ou des moments clés dans le match final. Malgré les variations dans le temps de possession, les joueuses espagnoles adaptent très rapidement leurs positions. Cependant, l'interprétation des données reste complexe, soulignant la nécessité de prendre en compte de multiples facteurs contribuant à la dynamique du match. L'analyse des passes espagnoles tout au long de la compétition indique une amélioration générale de la précision, avec une diminution significative des passes hors limites. En conclusion, cet article fournit des informations sur les performances de l'Espagne à travers une analyse statistique, illustrant sa capacité d'adaptation par des ajustements tactiques. Notre travail montre également que l'approche par l'exploration des processus permet d'analyser des séries d'actions sous forme de séquence, et se révèle utile pour comprendre les complexités inhérentes liées au football. Comme travaux futurs nous envisageons de détecter des similarités entre différents couples équipe/match à l'aide d'une classification de *process*.

References

[1] van der Aalst, W.: Data Science in Action, pp. 3–23. Springer Berlin Heidelberg, Berlin, Heidelberg (2016). https://doi.org/10.1007/978-3-662-49851-4˙1https://doi.org/\detokenize{10.1007/978-3-662-49851-4_1}

[2] Janssenswillen, G., Depaire, B., Swennen, M., Jans, M.J., Vanhoof, K.: bupaR: Enabling reproducible business process analysis. Knowledge-Based Systems **163**, 1857 (2019). https://doi.org/10.1016/j.knosys.2018.10.018https://doi.org/\detokenize{10.1016/j.knosys.2018.10.018}

[3] Peter O'Donoghue, Katerina Papadimitriou, V.G., Haralambis, K.: Statistical methods in performance analysis: an example from international soccer. International Journal of Performance Analysis in Sport **12**(1), 144–155 (2012). https://doi.org/10.1080/24748668.2012.11868590https://doi.org/\detokenize{10.1080/24748668.2012.11868590}

[4] Rico-González, M., Pino-Ortega, J., Praça, G.M., Clemente, F.M.: Practical applications for designing soccer' training tasks from multivariate data analysis: A systematic review emphasizing tactical training. Perceptual and Motor Skills **129**(3), 892–931 (2022). https://doi.org/10.1177/00315125211073404https://doi.org/\detokenize{10.1177/00315125211073404}, pMID: 35084256