La Statistique et le Sport

Que ce soit pour agrémenter les retransmissions télévisées de rencontres sportives ou pour améliorer les performances d'un athlète ou d'une équipe, l'analyse de données est une aide précieuse dans le monde du sport.

L'ACCÈS SUR LE PODIUM EN PARA TIR À L'ARC

https://fr.wikipedia.org/wiki/Tir_à_l'arc_aux_Jeux_paralympiques

Dans une compétition sportive, **est-ce que les athlètes arrivent au podium par hasard ?** Heureusement, non ! Les athlètes doivent posséder différentes qualités compétitives, qui sont les principaux facteurs de la performance sportive : des qualités physiques, techniques, mentales et tactiques. Les coachs visent à améliorer la performance globale en exploitant des données externes, comme les capteurs sur les joueurs, les vidéos des matchs, les résultats des compétitions sportives.

Prenons l'exemple du **tir à l'arc aux Jeux paralympiques 2024** : existe-il des facteurs de performance individuelle permettant de monter sur le podium ? Le coach sportif dispose de l'ensemble des résultats de tirs de flèches des para-athlètes dans différentes compétitions depuis 2009, récupérés sur internet. La figure suivante explicite la démarche : à partir du **tableau de données**, des **facteurs de performance individuelle** sont construits pour chaque para-athlète (comme le nombre de tirs en plein centre de la cible, la somme des tirs lors des phases de qualification et en finale). Ensuite, les facteurs de performance des para-athlètes qui accèdent au podium sont identifiés à l'aide d'une **méthode statistique** (CART). Les coachs déterminent ainsi les domaines sur lesquels leurs para-athlètes doivent travailler pour accéder au podium.

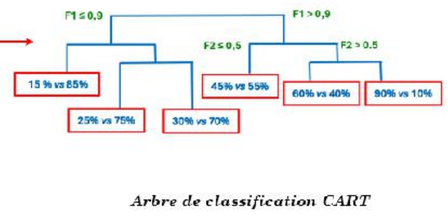
Rang	Score	Nbre de 10	Centre cible	...	Nation	Catégorie	Num Compet
1	1382	101	52	...	KOR	CMO	7776
2	1369	81	56	...	SUI	CMO	7776
3	1364	78	29	...	GER	CMO	7776
.	.	.	.	...	.	.	.
.	.	.	.	...	.	.	.
.	.	.	.	...	.	.	.
.	.	.	.	...	.	.	.
.	.	.	.	...	.	.	.
.	.	.	.	...	.	.	.
.	.	.	.	...	.	.	.
.	.	.	.	...	.	.	.

Tableau de données

F1 = g_1(données)
F2 = g_2(données)
...
F10 = g_{10}(données)

F1	F2	...	F10
0,2	0,4	...	0,8
0,5	0,8	...	0,25
0,8	0,1	...	0,3
.	.	...	.
.	.	...	.
.	.	...	.
.	.	...	.
.	.	...	.
.	.	...	.
.	.	...	.

Facteurs de performance



Arbre de classification CART

La **méthode CART** (Classification And Regression Tree) construit un arbre pour étudier la population d'archers.

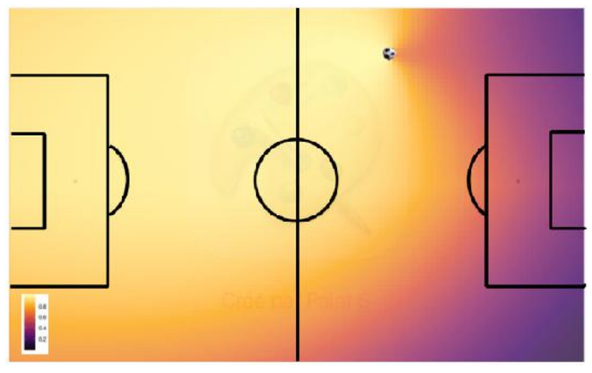
Au sommet de l'arbre, la racine contient tous les para-athlètes, puis l'algorithme sépare au mieux les accédants au podium des non-accédants.
La première séparation utilise le facteur F1 (score moyen), avec un seuil de 0,9 points. Chacune des deux branches est scindée en deux sous-branches, en fonction d'un autre facteur : la première branche est divisée grâce au facteur F2 (centre de la cible), avec un seuil de 0,5.
L'arbre déploie de nouvelles sous-branches, jusqu'aux feuilles, qui représentent des groupes d'athlètes partageant des caractéristiques similaires. Par exemple, la feuille la plus à droite contient 90% des accédants au podium.

PERFORMANCE COLLECTIVE EN FOOTBALL

Dans n'importe quel sport collectif, les **passes entre coéquipiers** sont des éléments-clés qui traduisent la stratégie d'une équipe pour faire face à son adversaire. Comprendre ce qui rend une passe réussie est crucial pour améliorer la performance d'un collectif.

Considérons le football. On peut obtenir des données sur les passes des joueurs ou joueuses (d'où elles partent, leur longueur, l'angle par rapport au but, si un adversaire est proche …). On sait aussi si la passe réussit (un coéquipier la récupère) ou échoue (le ballon sort des limites du terrain ou est intercepté par un adversaire). Avec ces informations, on peut modéliser la probabilité de réussite d'une passe.

L'image représente un terrain de football, et on considère une passe avec comme point de départ le ballon. La couleur correspond à la probabilité que la passe soit réussie allant du noir (pour une probabilité proche de 0, la passe est ratée) au jaune (pour une probabilité proche de 1, la passe est réussie).
Comme attendu, une passe en retrait courte a plus de chance de réussir qu'une passe longue vers l'avant.



D'un point de vue statistique, on cherche à calculer **la probabilité de réussir une passe** en fonction de différentes variables comme la position de départ sur le terrain, la longueur et l'angle de la passe.

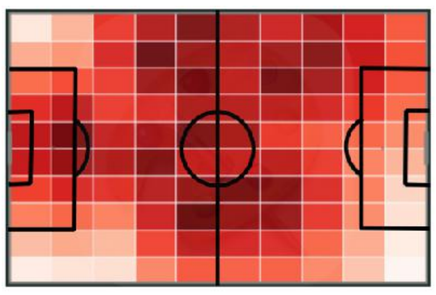
Cela peut s'effectuer à l'aide d'un modèle de **régression logistique**, éventuellement associé à une technique de sélection de variables, pour identifier les facteurs de réussite d'une passe et mesurer leur importance.

Cette première approche est assez simple car elle considère chaque passe comme un événement isolé. Des approches plus avancées (telles que le **process mining**) permettent d'étudier la passe en tant qu'élément dans un processus. Il est ainsi possible d'identifier les successions d'actions les plus susceptibles de mener à un but. En ajoutant les données de localisation de plusieurs joueurs ou joueuses simultanément, on peut même modéliser les dynamiques collectives !

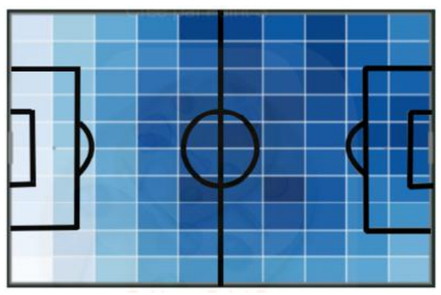
CLASSIFICATION DE SITUATIONS DE JEU AU FOOTBALL POUR ALIMENTER UN ENVIRONNEMENT DE RÉALITÉ VIRTUELLE

En football, chaque joueur perçoit les **mouvements collectifs** autour de lui, ce qui guide sa décision sur ses actions futures. Pour comprendre et analyser les interactions entre la perception et l'action, les athlètes sont immergés dans des **environnements virtuels** simulant des situations de jeu réalistes. On peut en déduire des **classes de situations de jeux,** dans lesquelles les entraîneurs peuvent piocher des exemples pour adapter les entraînements.

Imaginons une situation de jeu à un instant donné. Cette situation peut être représentée par une paire de cartes de chaleur, dépeignant la pression exercée par les deux équipes sur différentes zones du terrain : dans la figure, en rouge la pression exercée sur tout le terrain par l'équipe qui possède le ballon, et en bleue la pression exercée par l'équipe qui n'a pas le ballon. Plus la couleur est foncée, plus la pression est forte.



Pression exercée par l'équipe qui possède le ballon.



Pression exercée par l'équipe qui ne possède pas le ballon.

Pour mesurer cette pression, on utilise un **descripteur** traduisant la capacité des joueurs à se déplacer rapidement vers différents points du terrain, en tenant compte de leur vitesse et de leur orientation. Les données à analyser sont des **données spatio-temporelles.** Pour les regrouper, nous avons opté pour un clustering en deux étapes : d'abord un clustering spatial, puis un clustering temporel.

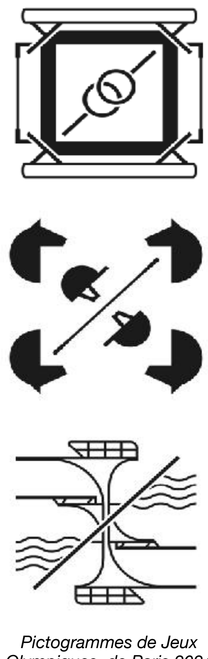
Clustering spatial : on classe ensemble les paires de cartes de chaleurs qui se ressemblent. La ressemblance se traduit sur le plan mathématique par une faible distance. Ici, la distance utilisée est appelée distance de *Sinkhorn*. On obtient des classes d'états de jeu.
Clustering temporel : on classe ensemble les successions d'états de jeux qui se ressemblent. Ici, la distance utilisée est celle de *l'Optimal Matching*. On obtient des classes de situations de jeu.

ÉVALUER LA PERFORMANCE DES JUGES INTERNATIONAUX ET IDENTIFIER LEUR BIAIS

Dans beaucoup de sports, les performances sont évaluées par des juges qui attribuent une note aux compétiteurs et compétitrices. Parmi les sports olympiques ou paralympiques, c'est le cas par exemple de la gymnastique, du plongeon, du dressage équestre, du patinage sur glace, du saut à ski ou du ski/snowboard freestyle.

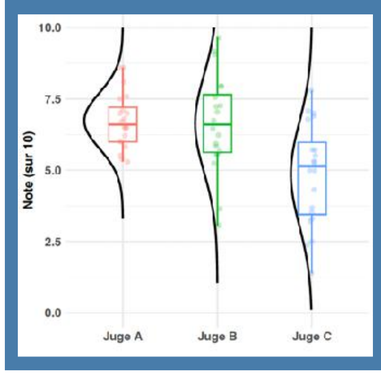
De nos jours, avec les possibilités de regarder les prestations au ralenti et de faire des arrêts sur image, **les notes des juges peuvent être facilement remises en question.**
Avant de critiquer les juges ou de parler de corruption, il est indispensable de prendre en compte un certains nombres d'éléments. Par exemple, la **variabilité des notes** peut être due à des facteurs comme la température lors de la compétition, la durée de la compétition ou encore la fatigue des juges. Mais parfois, elle est propre au sport considéré : par exemple, en gymnastique, les notes peuvent varier en fonction de l'agrès utilisé. De plus, sur les performances des meilleurs athlètes, les juges semblent accorder des notes similaires.

Les statisticiens et statisticiennes du sport peuvent alors modéliser cette variabilité des notes en tenant compte de tous ces facteurs, détecter les notes qui sortent de l'ordinaire (les valeurs aberrantes) et éventuellement **corriger les biais identifiés.**



Pictogrammes de Jeux Olympiques de Paris 2024

En statistique, la **variance** mesure la dispersion des valeurs prises par une variable aléatoire. Lorsque la variabilité diffère selon certaines sous-populations (ici, les jurys), on parle alors d'hétéroscédasticité. Sinon, on parle d'homoscédasticité.



Voici les résultats pour 3 juges sur 20 sportifs. On a représenté les notes (les points), une boite à moustache et la distribution pour chaque juge.
Les juges A et B ont la même moyenne, mais les notes sont hétéroscedastiques : le juge B a une plus grande variabilité, avec des notes plus basses et plus hautes que le juge A.
Le juge B et le juge C sont homoscédastiques, mais les moyennes sont très différentes : le juge C est plus dur que les juges A et B !

SFdS

La Société Française de Statistique assure la promotion de la statistique dans toutes ses composantes : RECHERCHE, ENSEIGNEMENT, APPLICATIONS.

Retrouvez sur notre site nos activités **: organisation d'évènements (conférences, soirées débats,..) , publications scientifiques, formation.**

Scannez moi pour télécharger ce poster