

TP : Statistiques et Analyse de Données



MODÈLE DE PRÉVISION DE L'OCCUPATION DES STADES DE FOOTBALL AU MAROC

HFIDH Sami
NAOUADIR Moussa
KRIBICH Ayoub

Professeur : Pr. BARAKA Achraf Chakir

Table des Matières

TP : STATISTIQUES ET ANALYSE DE DONNEES	1
<i>MODÈLE DE PRÉVISION DE L'OCCUPATION DES STADES DE FOOTBALL AU MAROC</i>	<i>1</i>
HFIDH SAMI.....	1
—.....	1
<i>Professeur : Pr. BARAKA Achraf Chakir</i>	<i>1</i>
1. INTRODUCTION	3
1.1 CONTEXTE ET DOMAINE D'ETUDE.....	3
1.2 PROBLEMATIQUE	3
1.3 OBJECTIFS DU PROJET	4
1.4 METHODOLOGIE.....	4
2. DICTIONNAIRE DES VARIABLES	4
3. PRÉTRAITEMENT DES DONNÉES	6
3.1 DONNEES INITIALES.....	6
3.2 TRAITEMENT SPECIFIQUE BOTOLA PRO.....	6
3.3 GESTION DES VALEURS MANQUANTES	6
3.4 DETECTION DES DONNEES ABERRANTES.....	6
3.5 DONNEES FINALES ET STATISTIQUES CLES.....	7
4. ANALYSE EXPLORATOIRE DES DONNÉES	7
4.1 DISTRIBUTION DE LA VARIABLE CIBLE	7
4.2 TAUX D'OCCUPATION	7
4.3 ANALYSE DES CORRELATIONS.....	8
4.4 ANALYSE PAR COMPETITION ET EFFETS BOOLEENS	8
4.5 RECOMMANDATIONS POUR LA MODELISATION	8
5. MÉTHODE 1 : ANALYSE EN COMPOSANTES PRINCIPALES (ACP)	9
6. MÉTHODE 2 : RÉGRESSION LINÉAIRE MULTIVARIÉE	11
7. SYNTHÈSE COMPARATIVE DES MÉTHODES	12
8. CONCLUSION	13
9. SOURCES DE DONNEES	13

1. INTRODUCTION

1.1 Contexte et domaine d'étude

Dans la dernière décennie, le Football au Maroc transpose le simple cadre sportif pour devenir un phénomène social, culturel et économique majeur. Une révolution structurelle a eu lieu par le professionnalisme de la Botola Pro et l'organisation croissante des grands événements internationaux tels que la Coupe du monde des Clubs, la coupe d'Afrique 2025 qui se déroule actuellement étant un pays organisateur et participant et l'attribution à une coalition comprenant le Maroc, l'Espagne et le Portugal de l'organisation de la Coupe du monde de la FIFA 2030 constitue un moment historique pour le royaume chérifien. Plus qu'un événement sportif, cet engagement s'inscrit dans une stratégie globale de développement des infrastructures, de modernisation des réseaux de transport et d'accélération de la formation sportive.

Par ailleurs, la gestion des infrastructures sportives a devenu un enjeu stratégique national. Les Stades, autrefois simple lieu des rassemblements, sont aujourd'hui des actifs économiques dont la rentabilité dépend directement de leur taux d'occupation.

L'affluence est un indicateur clé de performance, impactant directement les revenus de billetterie, les recettes annexes, la sécurité et l'expérience des supporters.

1.2 Problématique

L'occupation des stades au Maroc présente une très forte variabilité. Cependant, l'affluence reste susceptible un éclat en créant une ambiance électrique reconnue mondialement, tandis que d'autres peinent à attirer quelques milliers de spectateurs.

Les données observées montrent des taux d'occupation balançant entre 0% (matchs à huis clos ou très faible affluence) et 99.7% (matchs complets). Cette volatilité pose une problématique centrale pour les organisateurs et les décideurs publics :

Comment prédire l'affluence pour optimiser la gestion des infrastructures sportives et maximiser les revenus ?

Une sous-estimation de l'affluence peut entraîner des risques sécuritaires majeurs, en revanche, une surestimation conduit à des pertes financières et une image dégradée d'un stade vide.

1.3 Objectifs du projet

Ce rapport relate cette problématique à travers une approche rigoureuse de Statistiques et Analyse de Données. Nos objectifs sont triples :

- ✓ **Appliquer l'Analyse en Composantes Principales (ACP)** afin de réduire la dimensionnalité des données et identifier les structures latentes qui régissent l'affluence.
- ✓ **Développer un modèle de régression linéaire multiple** pour prédire l'affluence future à base des caractéristiques connues avant le match.
- ✓ **Comparer les approches** (variables brutes vs composantes principales) pour présenter la stratégie de modélisation la plus efficace.

1.4 Méthodologie

Notre démarche se compose de cinq phases principales : la collecte (Scraping Data Réelle) et le nettoyage d'un jeu de données de 363 observations ensuite une analyse exploratoire approfondie (EDA) pour comprendre les distributions et corrélations puis l'application de l'ACP sur 11 variables quantitatives et la construction, l'évaluation de trois modèles de régression distincts et enfin, une synthèse comparative pour en tirer des recommandations opérationnelles.

2. DICTIONNAIRE DES VARIABLES

L'ensemble de données (dataset) collectées à partir des sources réelles comprend initialement **19** variables décrivant les caractéristiques des matchs, enrichies par la suite de **11** variables calculées pour les besoins de l'analyse.

Variable	Type	Description
Variables Initiales		
Competition	Catégorielle	Type de compétition (Botola Pro, CAF, FIFA CWC, etc.)
saison	Catégorielle	Année de la saison sportive
Stade	Catégorielle	Nom du stade accueillant le match

Ville	Catégorielle	Ville où se situe le stade
Capacite	Numérique	Capacité maximale officielle du stade
Equipe_Home	Catégorielle	Nom de l'équipe jouant à domicile
Equipe_Away	Catégorielle	Nom de l'équipe visiteuse
Score	Texte	Résultat final du match
isDerby	Booléen	Indique si le match oppose deux rivaux historiques (1/0)
PhaseImportance	Catégorielle	Niveau de la phase (poule, éliminatoire, finale)
is_important_match	Booléen	Match à fort enjeu (demi-finale, finale) (1/0)
is_top_team_home	Booléen	Présence d'une équipe majeure à domicile (Raja, Wydad, AS FAR...)
affluence_moyenne	Numérique	Variable cible : Nombre de spectateurs présents
date_Match	Date	Date calendaire du match
heure_match	Temps	Heure du coup d'envoi
prix_billet_moyen	Numérique	Prix moyen estimé du billet (MAD)
temperature	Numérique	Température enregistrée lors du match (°C)
Recette_Moyenne	Numérique	Recette totale générée (Estimation : Affluence × Prix)
source_file	Texte	Fichier source de l'observation
Variables Créées		

Tableau 1 : Dictionnaire des variables

3. PRÉTRAITEMENT DES DONNÉES

3.1 Données initiales

Le jeu de données brut contient 363 observations et 19 variables. Ces données proviennent de sources diverses (FRMF, FIFA, CAF, Wikipédia, Transfermarkt) et couvrent plusieurs saisons sportives.

3.2 Traitement spécifique Botola Pro

Un traitement particulier a été portée aux données de la **Botola Pro** vue l'insuffisance de l'affluence des matches selon les 30 journée, qui représentent 240 lignes, soit environ 66% du dataset. Les sources originales présentaient parfois des données agrégées (Affluence Moyenne).

Nous avons procédé à une désagrégation de l'affluence en utilisant des coefficients multiplicateurs basés sur l'historique.

Pour refléter la réalité de l'incertitude de mesure, une variance aléatoire de $\pm 5\%$ a été introduite artificiellement sur ces estimations désagrégées.

3.3 Gestion des valeurs manquantes

L'intégrité des données a été assurée par des méthodes d'imputation robustes :

- ✓ **Température** : Les valeurs manquantes ont été remplacées par la médiane globale (méthode robuste aux valeurs extrêmes).
- ✓ **Prix du billet** : Imputation par la médiane calculée spécifiquement pour chaque type de compétition, afin de respecter les disparités tarifaires entre championnats locaux et internationaux.
- ✓ **Recette** : Lorsque manquante, elle a été recalculée systématiquement par la formule : $\text{Affluence} \times \text{Prix moyen du billet}$.

3.4 Détection des données aberrantes

La méthode de l'écart interquartile (Interquartile Range - IQR) a été utilisée. Elle a permis d'identifier :

- 8 outliers sur la variable **affluence**.
- 0 outlier sur la variable **taux_occupation**.

Décision : Ces observations ont été repérées via des variables booléennes mais conservées dans le jeu de données. En effet, dans le contexte sportif, les affluences exceptionnelles (derbys, finales) sont des événements réels et significatifs, non des erreurs de mesure.

3.5 Données finales et statistiques clés

Après nettoyage, le dataset final ne comporte aucune suppression de ligne (363 observations conservées) et s'est enrichi pour atteindre 30 variables. Le fichier a été exporté sous le nom **data_cleaned.csv**.

Statistiques Descriptives Clés :

Affluence moyenne : 15 025 - 11 974 spectateurs.

Taux d'occupation moyen : 47.3%.

Capacité moyenne des stades : 35 406 places.

4. ANALYSE EXPLORATOIRE DES DONNÉES

4.1 Distribution de la variable cible

La variable **affluence_moyenne** présente une distribution asymétrique à droite (positive skewness). La moyenne est de **15 025** clairement supérieure à la médiane qui est de **9 796**, indiquant la présence de quelques matchs à très forte affluence qui poussent la moyenne vers le haut. L'étendue va de matchs quasi-vides à des stades combles de 50 000 spectateurs.

Cette non-normalité conduit à des transformations (comme le logarithme) pourraient être bénéfiques pour certains modèles linéaires.

4.2 Taux d'occupation

Le taux d'occupation montre une très forte variabilité, couvrant tout le spectre possible de 0.0% à 99.7%. Avec une moyenne inférieure à 50% (47.3%), on constate que les stades marocains opèrent majoritairement en sous-capacité, soulignant le potentiel d'optimisation commerciale.

4.3 Analyse des corrélations

L'analyse bivariée avec la variable cible révèle les associations suivantes :

Recette_Moyenne (0.691) : Très forte corrélation positive, logique puisque la recette dépend directement de l'affluence.

Capacité (0.603) : Forte corrélation. Les grands stades attirent plus de monde, ou sont choisis pour les grands matchs. C'est une contrainte physique structurante.

taux_occupation (0.469) : Corrélation modérée, indiquant que l'affluence absolue et le taux de remplissage sont liés.

prix_billet_moyen (0.278) : Corrélation positive mais plus faible. Des prix plus élevés sont souvent associés à des matchs plus prestigieux qui attirent plus de monde.

4.4 Analyse par compétition et effets booléens

Le prestige de la compétition est un facteur déterminant. La FIFA Club World Cup enregistre les affluences maximales, suivie par les compétitions africaines (CAF). La Botola Pro présente une moyenne plus faible, typique d'un championnat régulier.

Les variables contextuelles ont un impact quantifié majeur :

- ✓ Un **Derby** (isDerby=1) attire en moyenne +8 000 spectateurs supplémentaires.
- ✓ Un **Match important** (is_important_match=1) génère un surplus moyen de +12 000 spectateurs.
- ✓ La présence d'une **Top Team à domicile** apporte environ +6 000 spectateurs.

4.5 Recommandations pour la modélisation

L'EDA suggère de prioriser les variables de structure (Capacité) et de contexte (Derby, Top Team). Les variables temporelles (mois, jour) semblent pertinentes pour capturer la saisonnalité. L'interaction **derby_top_team** mérite d'être testée pour capturer l'effet multiplicateur des grands chocs.

5. MÉTHODE 1 : ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

5.1 Cadre théorique

L'Analyse en Composantes Principales (ACP) est une méthode factorielle qui vise à réduire la dimensionnalité d'un jeu de données quantitatives corrélées. Elle transforme les variables initiales en nouvelles variables non-corrélées, les composantes principales, tout en préservant le maximum d'information (variance).

$$\text{Formule : } \mathbf{Z} = \mathbf{XW}$$

Où X est la matrice des données centrées-réduites et W la matrice des vecteurs propres. La standardisation est cruciale ici car nos variables ont des unités très différentes (capacité en dizaines de milliers, température en degrés).

5.2 Résultats : Variance expliquée

L'application de l'ACP sur 11 variables quantitatives a donné la décomposition de variance suivante :

Composante	Valeur propre (λ)	Variance expliquée	Variance cumulée
PC1	3.663	33.30%	33.30%
PC2	2.264	20.58%	53.88%
PC3	1.208	10.98%	64.87%
PC4	1.055	9.59%	74.46%
PC5	0.791	7.19%	81.65%

5.3 Sélection du nombre de composantes

Selon le critère de Kaiser (valeurs propres > 1), nous retenons 4 composantes. Ce choix permet de conserver 74.46% de l'information initiale tout en réduisant le nombre de variables de 11 à 4, réalisant ainsi une compression efficace des données.

5.4 Interprétation des axes

PC1 (33.30%) - Dimension Économique et Attractivité

Cette première dimension est fortement corrélée positivement avec la Recette Moyenne (+0.75), le Prix du billet (+0.71), et l'indicateur isDerby (+0.70). Elle oppose les matchs "premium" (derbys, chers, lucratifs) aux matchs ordinaires. C'est l'axe de la valeur événementielle.

PC2 (20.58%) - Dimension Temporelle

Le deuxième axe est dominé par le jour de la semaine (+0.88) et l'indicateur weekend (-0.79). Il est orthogonal à l'attractivité, ce qui signifie que la programmation temporelle (semaine vs weekend) est un facteur indépendant de la qualité de l'affiche.

PC3 (10.98%) - Dimension Compétition Locale

Cet axe capture une nuance spécifique liée aux Derbys (+0.55) joués dans des stades de grande Capacité (+0.42), distinguant l'aspect purement populaire et massif de l'aspect économique.

5.5 Synthèse ACP

L'ACP a permis de structurer l'information en trois dimensions: l'attractivité économico-sportive, la programmation temporelle et l'ampleur de l'infrastructure.

Ces nouvelles variables synthétiques (PC1 à PC4) seront utilisées comme inputs dans nos modèles de régression.

6. MÉTHODE 2 : RÉGRESSION LINÉAIRE MULTIVARIÉE

6.1 Cadre théorique

Nous visons à représenter l'affluence moyenne par le modèle linéaire $Y = X\beta + \varepsilon$.

Les paramètres β sont estimés via la méthode des **moindres carrés ordinaires**, qui consiste à minimiser la somme des carrés des différences entre les valeurs observées et celles prédites par le modèle.

6.2 Stratégie de modélisation

Le dataset a été divisé en un ensemble d'entraînement (80%, $n=290$) et de test (20%, $n=73$). Trois approches ont été comparées :

1. **Modèle 1 (Variables Originales)** : Utilise les 10 variables explicatives brutes.
2. **Modèle 2 (Composantes Principales)** : Utilise les 5 premières composantes issues de l'ACP.
3. **Modèle 3 (Approche Hybride)** : Combine les variables originales avec les deux premières composantes (PC1 et PC2).

6.3 Résultats - Modèle 3 : Approche Hybride (Meilleur Modèle)

L'approche hybride s'est révélée explicitement supérieure aux deux autres.

Métrique	Train Set	Test Set
R^2 (Coeff. détermination)	0.7400	0.8519
RMSE (Erreur quadratique)	5946.67	5027.37
MAE (Erreur absolue)	4118.66	3696.36

6.4 Prédiction du taux d'occupation

En concluant le taux d'occupation à partir de l'affluence prédite et de la capacité connue, nous obtenons un modèle capable de prédire le remplissage avec une erreur moyenne de **12.11** points de pourcentage.

NB : C'est un outil précieux pour anticiper la densité de foule.

7. SYNTHÈSE COMPARATIVE DES MÉTHODES

7.1 Complémentarité des approches

Ce projet illustre parfaitement la synergie entre l'exploration (ACP) et la prédiction (Régression). L'ACP a permis de comprendre que l'affluence est structurée par des dimensions cachées (attractivité, temps) que les variables brutes peinent parfois à capturer individuellement.

7.2 Performance comparative

Modèle	R ²	RMSE	MAE
Modèle 1 : Variables Originales	0.674	7458	5270
Modèle 2 : ACP seule	0.628	7963	5727
Modèle 3 : Hybride	0.852	5027	3696

Comparatif final des modèles (Test Set)

L'ajout des composantes PC1 et PC2 aux variables originales (Modèle 3) a permis une performance élevée (+18 points de R²). Cela suggère que les composantes principales ont capturé une information synthétique "pure" qui a aidé le modèle à mieux stabiliser ses prédictions.

8. Conclusion

Ce projet a permis de développer un outil fiable pour projeter l'affluence des États marocains. L'utilisation combinée de l'analyse en composantes principales et de la régression linéaire multivariée a non seulement permis de comprendre les facteurs profonds influençant la fréquence des étapes (l'importance cruciale de la dimension première), mais elle a également produit un modèle prédictif qui explique **85 %** de la variance observée.

9. Sources de données

- ✓ Fédération Royale Marocaine de Football (FRMF) - Données Botola Pro.
- ✓ FIFA.com - Statistiques Coupe du Monde des Clubs.
- ✓ CAFonline.com - Rapports compétitions africaines.
- ✓ Wikipédia – Archive des compétitions.
- ✓ TransferMarket – Résultat des matchs – Statistiques variées.