

ANALYSE DE LA PERFORMANCE CYCLISTE

Modélisation prédictive pour la sélection nationale

KIGALI 2025

*Optimisation de la sélection par une approche multi-critères :
Topographie, Morphologie et Dynamique de Performance Élite*

Étudiants :

Lilou MALFOY

Mathias LE BOUEDEC

Siloé VELTZ

Encadrant :

M. Atiq

Résumé Exécutif : Orientations Stratégiques

Ce rapport présente une méthodologie d'aide à la décision entièrement quantitative pour la sélection de l'Équipe de France de cyclisme sur route aux Championnats du Monde 2025 à Kigali. Face à un parcours exceptionnel par sa sévérité (270 km et 5475 m de dénivelé positif), reposant sur la répétition de côtes courtes et très raides, l'objectif est de dépasser une sélection fondée sur l'intuition ou la réputation, au profit d'un processus analytique objectif, traçable et reproductible.

I. Le défi de Kigali : traduire un parcours extrême en données

Le parcours rwandais impose une contrainte physiologique singulière, combinant endurance prolongée, accumulation de fatigue et répétition d'efforts explosifs à forte intensité. Afin de rendre ces exigences comparables aux performances passées des coureurs, nous avons transformé le parcours en un vecteur de difficulté multidimensionnel reposant sur trois variables clés :

- la distance totale, représentant l'exigence d'endurance globale,
- le dénivelé positif cumulé, traduisant la charge de travail et la fatigue accumulée,
- la pente moyenne des ascensions, indicateur central de l'intensité relative des efforts.

Cette formalisation constitue la référence objective sur laquelle repose l'ensemble des comparaisons ultérieures.

II. Une architecture de données exhaustive et rigoureuse

La méthodologie s'appuie sur la construction d'une base de données complète concernant les 63 premiers coureurs français du classement UCI 2024–2025. Celle-ci intègre :

- les caractéristiques morphologiques (taille, poids),
- l'ensemble des performances historiques récentes (classements, points UCI, prestige et récence des courses),
- les profils altimétriques détaillés des courses, incluant les côtes et leur intensité.

Toutes les données ont été collectées à partir de sources publiques fiables, automatisées lorsque possible, puis structurées dans des fichiers exploitables garantissant cohérence, traçabilité et reproductibilité.

III. Comparaison multi-niveaux des courses avec Kigali

L'évaluation de l'adéquation des coureurs au parcours repose sur une double analyse complémentaire :

- **Analyse macroscopique** : comparaison globale des courses passées avec Kigali à l'aide d'une distance euclidienne normalisée sur la distance et le dénivelé, permettant d'identifier les épreuves les plus comparables en termes d'exigence globale.
- **Analyse microscopique** : appariement précis des côtes de Kigali avec celles des courses passées via l'algorithme hongrois, mesurant la capacité des coureurs à reproduire des efforts courts, raides et répétés.

Cette approche permet de capturer à la fois la résistance à la fatigue sur la durée et la spécificité technique des efforts exigés par le parcours.

IV. Profilage physiologique et scoring multi-critères

Au-delà des résultats bruts, la méthodologie intègre la dimension physiologique des coureurs. Un clustering K-Means, associé à une analyse en composantes principales (PCA), permet de segmenter les athlètes en profils cohérents (grimpeurs, puncheurs, rouleurs) et de vérifier leur compatibilité réelle avec les exigences de Kigali. Le score final synthétise l'ensemble des dimensions analysées :

- performance historique pondérée (classement, prestige et récence),
- similarité macro et micro avec le parcours cible,
- adéquation physiologique aux efforts explosifs répétés.

V. Conclusion : une sélection optimisée et justifiable

L'application de cette méthodologie conduit à l'identification d'un **Top 8 de coureurs** dont la sélection est intégralement justifiée par des critères mesurables et transparents. Ce modèle fournit une base décisionnelle robuste, auditable et réplicable, limitant les biais subjectifs tout en maximisant l'adéquation entre les profils sélectionnés et l'un des parcours les plus exigeants de l'histoire récente des Championnats du Monde.

Table des matières

| | |
|--|-----------|
| Résumé Exécutif | 1 |
| 1 Introduction | 6 |
| 2 Analyse et formalisation du parcours des Championnats du Monde 2025 | 8 |
| 2.1 Présentation stratégique du tracé de Kigali | 8 |
| 2.2 Identification des contraintes physiologiques majeures | 8 |
| 2.3 Choix et justification des variables descriptives | 9 |
| 2.4 Modélisation vectorielle et distance euclidienne | 9 |
| 2.5 Méthodologie de structuration du référentiel CSV | 10 |
| 2.6 Limites du modèle de formalisation | 10 |
| 3 Collecte et structuration des données des coureurs | 11 |
| 3.1 Sélection du panel : Objectivité et représentativité du classement UCI | 11 |
| 3.2 Architecture du système d'acquisition de données | 12 |
| 3.2.1 Web Scraping et Ingénierie Morphologique (Python) | 12 |
| 3.2.2 Extraction et Normalisation des Performances (Python) | 13 |
| 3.2.3 Enrichissement Altimétrique et Traitement R | 13 |
| 3.3 Traitement et Nettoyage des données (Data Quality Assurance) | 14 |
| 3.3.1 Homogénéisation et Standardisation | 14 |
| 3.3.2 Vérification de cohérence et détection d'anomalies (<i>Outliers</i>) | 14 |
| 3.3.3 Stratégie d'Imputation et de Complétude | 14 |
| 3.4 Résultat de l'étape : Le socle de données certifié | 15 |
| 4 Comparaison multi-niveaux : De la charge globale à l'effort spécifique | 17 |
| 4.1 Analyse macroscopique : Similarité de charge globale | 17 |
| 4.1.1 Normalisation et Vecteur d'État | 17 |
| 4.1.2 Calcul du score de similarité et pondération stratégique | 17 |
| 4.2 Analyse microscopique : L'Algorithme Hongrois pour l'appariement des côtes | 19 |
| 4.2.1 Formalisation du problème d'affectation | 19 |
| 4.2.2 Application de l'Algorithme de Kuhn-Munkres (Hongrois) | 19 |
| 4.2.3 Validation par Analyse en Composantes Principales (PCA) | 20 |
| 4.2.4 Justification de la réduction dimensionnelle (PCA) | 20 |
| 4.3 Justification de la double approche : Pourquoi le "Score Global" ne suffit pas | 22 |
| 4.3.1 Visualisation de la dissociation des similarités | 22 |

| | |
|---|-----------|
| 5 Profilage Physiologique et Segmentation des Coureurs | 24 |
| 5.1 Qualification du terrain : La Bibliothèque d'Étapes | 24 |
| 5.2 Attribution des profils individuels aux coureurs | 25 |
| 5.2.1 Analyse de Dominance : La signature de performance | 25 |
| 5.2.2 Intégration Morphologique : Le filtre de validation physiologique | 26 |
| 5.2.3 Exemple d'application du filtre morphologique | 26 |
| 5.3 Validation par Clustering K-Means : Segmentation objective du panel | 27 |
| 5.3.1 Logique de l'algorithme et variables d'entrée | 27 |
| 5.3.2 Optimisation du nombre de groupes (Méthode de l'Elbow) | 27 |
| 5.3.3 Projection PCA et Justification de la dimensionnalité | 28 |
| 5.3.4 Justification de la dimensionnalité | 29 |
| 5.4 Synthèse : Cohérence des profils et validation de l'expertise | 30 |
| 6 Construction du Score Final et Sélection | 32 |
| 6.1 Modélisation du Score Macro-Physiologique | 32 |
| 6.1.1 Formalisation mathématique | 32 |
| 6.1.2 Signification Physiologique : La validation de la résilience métabolique | 33 |
| 6.1.3 Visualisation de l'espace de performance Macro | 33 |
| 6.1.4 Analyse des résultats intermédiaires | 33 |
| 6.2 Modélisation du Score Micro-Spécifique : L'expertise des côtes | 34 |
| 6.2.1 Méthodologie : Un changement de référentiel | 34 |
| 6.2.2 Résultats : L'émergence des spécialistes de l'explosivité | 35 |
| 6.3 Calcul du Score Final et Pondération Hybride | 36 |
| 6.3.1 La formule de synthèse | 36 |
| 6.3.2 Le Bonus de Cluster : L'arbitrage physiologique | 36 |
| 6.3.3 Synthèse et arbitrage : Le Score Final | 37 |
| 6.3.4 Hiérarchie globale et arbitrage de sélection | 38 |
| 6.3.5 La sélection officielle (Top 8) | 38 |
| 7 Difficultés rencontrées et gestion des aléas | 40 |
| 7.1 Défis de l'acquisition et de l'ingénierie des données | 40 |
| 7.2 Complexité algorithmique et optimisation | 40 |
| 7.3 Arbitrages et instabilité des modèles | 40 |
| 7.4 Biais statistiques et limites intrinsèques | 41 |
| 8 Conclusion | 42 |
| 8.1 Apports de la méthodologie | 42 |

| | | |
|-----|---|----|
| 8.2 | Limites du modèle : Entre statistiques et réalités du terrain | 42 |
| 8.3 | Perspectives de développement | 43 |

Introduction

La sélection d'une équipe nationale pour un championnat du monde constitue un enjeu stratégique majeur, tant sur le plan sportif qu'institutionnel. En cyclisme sur route, cette décision repose traditionnellement sur l'expertise des sélectionneurs, l'expérience accumulée et la réputation des coureurs. Bien que cette approche empirique ait démontré son efficacité, elle demeure fondamentalement subjective et ne garantit pas une adéquation optimale entre les profils sélectionnés et les exigences spécifiques d'un parcours donné.

Les Championnats du Monde de cyclisme sur route 2025, organisés à Kigali (Rwanda), s'inscrivent dans un contexte particulièrement singulier. Le tracé se distingue par une sévérité exceptionnelle, combinant une distance de 270 kilomètres et un dénivelé positif cumulé de 5475 mètres, ponctués par la répétition de côtes courtes et très raides. Ce type de parcours impose des contraintes physiologiques rares, associant endurance prolongée, accumulation de fatigue et répétition d'efforts explosifs à forte intensité. Dans ce cadre, l'analyse des seuls résultats passés apparaît insuffisante pour garantir une sélection pleinement adaptée.

Ce projet vise à dépasser les limites des approches traditionnelles en proposant une méthodologie de sélection fondée exclusivement sur l'analyse de données objectives. L'objectif est de construire un outil d'aide à la décision rigoureux, transparent et reproducible, capable de mesurer quantitativement l'adéquation entre les performances passées des coureurs français et les exigences spécifiques du parcours de Kigali. Cette approche ne se substitue pas à l'expertise humaine, mais a vocation à la compléter en fournissant un cadre analytique justifiant chaque décision à partir de critères mesurables.

La méthodologie développée repose sur une formalisation quantitative du parcours de Kigali, la constitution d'une base de données exhaustive des performances récentes des coureurs français, ainsi que sur une comparaison multi-niveaux entre les courses passées et le parcours cible. Elle intègre à la fois une analyse globale des exigences de la course et une prise en compte fine des efforts spécifiques imposés par les côtes, complétée par une analyse des profils physiologiques des coureurs. L'ensemble de ces éléments est synthétisé au sein d'un score final permettant d'identifier les profils les plus adaptés.

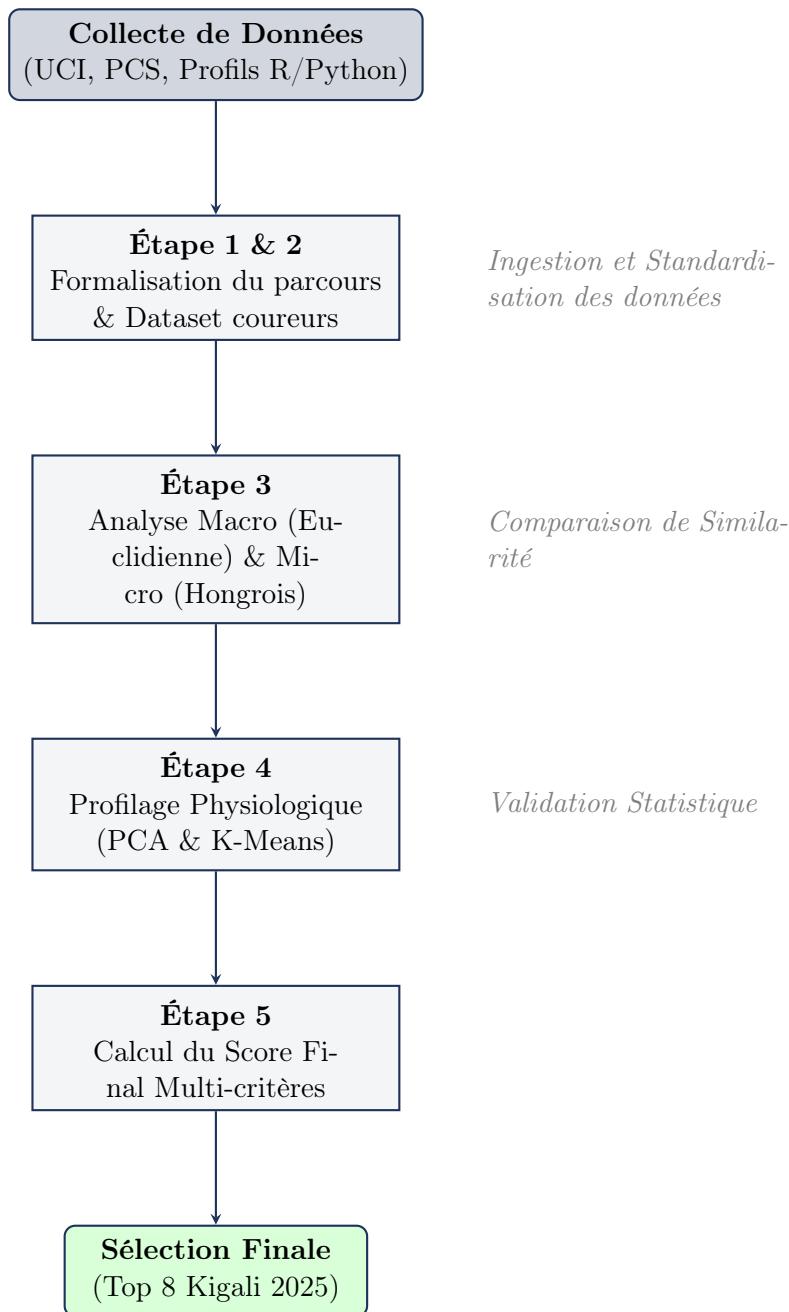


FIGURE 1 – Architecture globale de la méthodologie : de la collecte multi-sources à la sélection finale optimisée.

Analyse et formalisation du parcours des Championnats du Monde 2025

Présentation stratégique du tracé de Kigali

Les Championnats du Monde 2025, organisés à Kigali (Rwanda), marquent un tournant historique par la sévérité inédite de leur tracé. Le parcours masculin élite présente des caractéristiques biotopographiques extrêmes : une **distance totale de 270 km** associée à un **dénivelé positif cumulé de 5475 m**.

La spécificité majeure de ce tracé ne réside pas dans la présence de cols de haute montagne, mais dans la **répétition systémique de côtes courtes et très raides**. Cette configuration impose une alternance continue entre des phases d'efforts explosifs supra-physiologiques et des périodes de récupération incomplète, rendant la course particulièrement sélective.

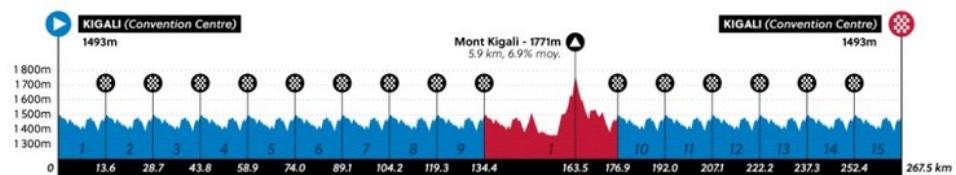


FIGURE 2 – Profil altimétrique officiel du parcours de Kigali (Source : UCI 2025).

Identification des contraintes physiologiques majeures

La réussite sur un parcours aussi atypique que celui de Kigali nécessite une synergie de capacités athlétiques spécifiques que notre modèle doit être capable de discriminer pour identifier les profils les plus performants :

- **Résilience métabolique et endurance critique** : La distance exceptionnelle de 270 km impose une capacité à maintenir une efficience énergétique élevée malgré l'épuisement des stocks de glycogène après plus de 6 heures d'effort. Le modèle doit valoriser les coureurs capables de préserver leur potentiel de puissance dans la phase finale de la course.
- **Capacité anaérobie lactique répétée** : La succession de côtes courtes et raides exige des fibres musculaires capables de produire des efforts explosifs répétés à très haute intensité. Cette contrainte nécessite une cinétique de récupération rapide entre chaque ascension pour éviter l'accumulation paralysante de lactate.
- **Optimisation du rapport poids/puissance** : Avec un dénivelé positif cumulé dépassant les 5000 m, la gravité constitue le principal facteur de résistance à

l'avancement. Le parcours favorise donc intrinsèquement les coureurs présentant un excellent rapport poids/puissance, caractéristiques des profils de "puncheurs-grimpeurs".

Ces exigences biotopographiques confirment que la performance à Kigali ne repose pas uniquement sur l'endurance pure, mais sur une combinaison rare de résistance à la fatigue et d'explosivité répétée.

Choix et justification des variables descriptives

Afin de transformer ce parcours en un objet mathématique traitable, nous avons isolé trois variables explicatives fondamentales :

1. **Distance totale (D_{tot})** : Variable pivot pour mesurer la robustesse métabolique sur le long terme.
2. **Dénivelé positif cumulé (H_{tot})** : Indicateur de la charge verticale globale et de la dépense énergétique totale de l'épreuve.
3. **Pente moyenne des ascensions (\bar{p})** : Variable discriminante permettant de distinguer les parcours roulants des parcours "cassants". Une pente moyenne élevée à Kigali indique une exigence de puissance explosive répétée.

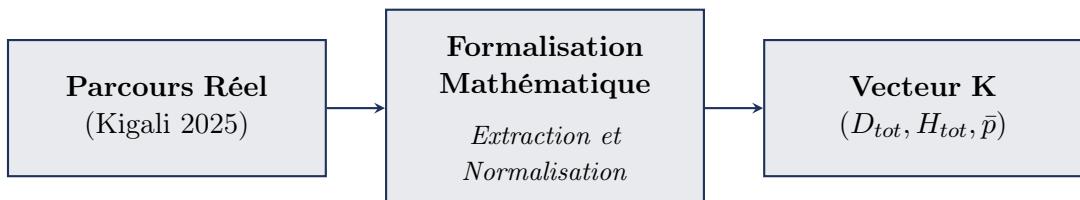


FIGURE 3 – Processus de transformation des caractéristiques physiques du parcours en variables quantitatives.

Modélisation vectorielle et distance euclidienne

Nous modélisons le parcours de Kigali par un vecteur de référence **K** :

$$\mathbf{K} = \begin{pmatrix} D_{tot} \\ H_{tot} \\ \bar{p} \end{pmatrix} = \begin{pmatrix} 270 \text{ km} \\ 5475 \text{ m} \\ \bar{p}_K \% \end{pmatrix}$$

Pour chaque course historique C_i présente dans notre dataset, un vecteur similaire \mathbf{C}_i est généré. Afin de calculer la similarité macroscopique entre le profil de Kigali et les performances passées, nous utilisons une **distance euclidienne normalisée** :

$$d(\mathbf{K}, \mathbf{C}_i) = \sqrt{\omega_D \left(\frac{D_K - D_i}{\sigma_D} \right)^2 + \omega_H \left(\frac{H_K - H_i}{\sigma_H} \right)^2 + \omega_p \left(\frac{\bar{p}_K - \bar{p}_i}{\sigma_p} \right)^2}$$

Où ω représente les poids de pondération (notamment une surpondération du dénivelé à 60% pour refléter la sévérité verticale du tracé) et σ représente l'écart-type de la variable sur l'ensemble du dataset pour neutraliser les effets d'échelle.

Méthodologie de structuration du référentiel CSV

La collecte des données topographiques a suivi un protocole de rigueur analytique :

- **Sources** : Exploitation des profils altimétriques officiels de l'UCI et digitalisation des données publiques de terrain.
- **Granularité** : Chaque ascension a été isolée pour extraire son kilométrage, sa pente moyenne et son dénivelé propre, permettant une analyse ultérieure à l'échelle "micro".
- **Stockage** : Les données sont structurées dans un référentiel CSV garantissant la traçabilité et permettant une réPLICATION de la méthodologie sur d'autres épreuves internationales.

Limites du modèle de formalisation

Bien que robuste, cette formalisation admet des hypothèses simplificatrices nécessaires à la modélisation :

- Le modèle se concentre sur les variables physiques et topographiques, excluant les facteurs exogènes (météo, altitude, tactique d'équipe).
- La linéarité de la distance euclidienne suppose une indépendance des variables, bien que le dénivelé et la pente soient physiologiquement liés.

Cette étape de formalisation constitue néanmoins une base objective indispensable pour sortir de la subjectivité décisionnelle et engager les analyses comparatives de performance.

Collecte et structuration des données des coureurs

L'objectif de cette étape est de constituer une base de données multidimensionnelle, fiable et réplicable. La pertinence de notre modèle de sélection dépend directement de l'intégrité des données sources, car elles alimenteront l'ensemble des analyses macro, micro et physiologiques ultérieures.

Sélection du panel : Objectivité et représentativité du classement UCI

Pour garantir un niveau de performance "Élite" et une neutralité totale dans la constitution de notre échantillon, nous avons retenu les **63 premiers coureurs français du classement mondial UCI 2024-2025**. Ce seuil n'est pas arbitraire ; il répond à plusieurs impératifs méthodologiques cruciaux pour la suite de l'analyse :

- **Élimination des biais subjectifs** : En utilisant le classement UCI, la sélection initiale ne repose sur aucune intuition humaine ou réputation passée, mais uniquement sur des points acquis en compétition officielle lors de la période de référence. Cela garantit une transparence totale, nécessaire pour une approche purement analytique.
- **Qualité et densité du signal** : Filtrer les 63 meilleurs athlètes permet de travailler sur des profils "actifs". Les coureurs au-delà de ce classement présentent souvent des données fragmentaires (moins de jours de course, participations à des épreuves de moindre catégorie). Ce panel garantit donc des données de performance régulières, denses et exploitables pour nos modèles de calcul.
- **Seuil de performance "Championnats du Monde"** : L'objectif est de sélectionner une équipe capable de performer sur un parcours extrême. Les coureurs retenus évoluent majoritairement dans les divisions *WorldTour* et *ProSeries*, ce qui assure que l'échantillon possède déjà les prérequis physiques minimaux pour terminer une course de 270 km.
- **Masse critique pour le clustering** : D'un point de vue statistique, un échantillon de 63 individus est idéal pour l'application d'algorithmes de classification non-supervisée comme le *K-Means*. Il est assez large pour faire émerger des tendances significatives (distinction nette entre grimpeurs, puncheurs et sprinteurs) sans être pollué par des "bruits" statistiques provenant de coureurs ayant des profils trop hétérogènes ou accidentellement classés.

En résumé, ce panel constitue le "socle de performance" du cyclisme français, offrant une base de données fiable et répllicable sur laquelle peut s'appuyer notre modèle de sélection objective.

TABLE 1 – Synthèse des performances pour le Top 10 des coureurs du panel (Saisons 2024-2025).

| Nom du Coureur | Total Points UCI | Nombre de courses |
|--------------------|------------------|-------------------|
| Julian Alaphilippe | 3170.7 | 167 |
| Lenny Martinez | 3149.0 | 177 |
| Romain Gregoire | 3093.0 | 186 |
| Emilien Jeanniere | 2637.0 | 160 |
| Paul Magnier | 2276.0 | 154 |
| Pavel Sivakov | 2258.9 | 184 |
| Valentin Madouas | 2144.0 | 161 |
| Dorian Godon | 2138.0 | 176 |
| Benoit Cosnefroy | 2006.0 | 78 |
| Christophe Laporte | 1903.0 | 79 |

Architecture du système d'acquisition de données

La fiabilité de notre modèle repose sur la mise en place d'un pipeline d'acquisition automatisé hybride, exploitant la complémentarité des environnements Python et R. Cette approche garantit la traçabilité des données brutes et la reproductibilité de l'étude à travers le temps.

3.2.1 Web Scraping et Ingénierie Morphologique (Python)

L'acquisition des variables biométriques a été automatisée via un script Python utilisant les bibliothèques `BeautifulSoup` pour le parsing HTML et `Requests` pour la gestion des requêtes réseau. Le script cible les pages individuelles des coureurs sur les plateformes *ProCyclingStats* et l'UCI.

Pour chaque athlète, nous extrayons le poids (W) et la taille (T) afin de calculer un indicateur de densité morphologique, utilisé comme variable d'entrée pour nos modèles de classification :

$$R_{\text{poids/taille}} = \frac{W}{T} \approx \text{Indice d'aptitude gravitationnelle} \quad (1)$$

Justification physiologique : Ce ratio constitue un proxy robuste du rapport puissance/poids (W/T). Un ratio faible (morphologie légère) identifie les profils adaptés aux forts pourcentages de Kigali (grimpeurs), tandis qu'un ratio élevé pointe vers des profils

dotés d'une masse musculaire importante, plus adaptés aux sections planes et venteuses (sprinteurs).

3.2.2 Extraction et Normalisation des Performances (Python)

Nous avons également collecté l'historique complet des compétitions disputées en 2024 et 2025. L'algorithme de scraping parcourt les tables de résultats pour extraire des variables structurées :

- **Indexation par catégorie** : Chaque course est classée (WorldTour, ProSeries, .1, .2) pour appliquer un coefficient de pondération reflétant le niveau d'adversité réel.
- **Variables de charge biotopographique** : Extraction systématique de la distance totale () et du dénivelé positif () pour chaque épreuve.
- **Normalisation temporelle** : Les dates sont converties au standard ISO 8601 pour permettre l'application d'une fonction de pondération exponentielle favorisant la forme récente (récence des performances).



FIGURE 4 – Pipeline d'acquisition : de la donnée web non-structurée au socle analytique.

TABLE 2 – Extrait de la base de données des performances historiques (5 premières lignes).

| Coureur | Date | Épreuve | Pos. | km | Déniv. (m) | Cat. | Pts UCI |
|--------------------|------------|--|------|-------|------------|------------|---------|
| Julian Alaphilippe | 2024-09-29 | World Championships ME - Road Race | DNF | 273.9 | 4291 | Étape | 0.0 |
| Julian Alaphilippe | 2024-09-21 | SUPER 8 Classic | 69 | 197.6 | 1872 | Étape | 0.0 |
| Julian Alaphilippe | 2024-09-15 | Grand Prix Cycliste de Montréal | 3 | 209.1 | 3899 | Étape | 325.0 |
| Julian Alaphilippe | 2024-09-13 | Grand Prix Cycliste de Québec | 81 | 201.6 | 2508 | Étape | 0.0 |
| Julian Alaphilippe | 2024-09-08 | Tour of Britain Mountains classification | 7 | - | - | Classement | 0.0 |

3.2.3 Enrichissement Altimétrique et Traitement R

Pour pallier l'absence de données altimétriques détaillées sur certains sites, nous avons utilisé l'environnement **R** et la bibliothèque spécifique **cyclingstats**. Cette étape permet de valider les dénivélés extraits en Python par une source tierce et de générer les variables nécessaires à l'analyse microscopique des côtes.

Traitement et Nettoyage des données (Data Quality Assurance)

La phase de *Data Cleaning* est l'étape charnière garantissant que les algorithmes de *Machine Learning* (Clustering K-Means) et d'optimisation (Algorithme Hongrois) ne soient pas biaisés par des données erronées. Un nettoyage rigoureux a été appliqué pour éliminer le "bruit" statistique et assurer l'intégrité du dataset.

3.3.1 Homogénéisation et Standardisation

Le scraping multi-sources engendre naturellement des hétérogénéités syntaxiques. Nous avons procédé à :

- **Normalisation nominale** : Alignement des noms de coureurs (gestion des accents, des traits d'union et de la casse) pour assurer la jointure parfaite entre les fichiers morphologiques et les fichiers de résultats.
- **Standardisation des unités** : Conversion systématique de toutes les variables métriques : distances en kilomètres (km), altitudes en mètres (m) et pentes en pourcentages (%).
- **Formatage temporel** : Conversion des dates de courses au format ISO 8601 (*AAAA – MM – JJ*) pour permettre le calcul précis des scores de récence.

3.3.2 Vérification de cohérence et détection d'anomalies (*Outliers*)

Pour garantir la fiabilité physiologique, nous avons instauré des filtres de plausibilité :

- **Contrôle biométrique** : Identification des valeurs aberrantes de poids ou de taille (ex : erreurs de saisie sur les sites sources). Toute valeur située à plus de 3 écarts-types de la moyenne du panel a été vérifiée manuellement par rapport aux profils officiels de l'UCI.
- **Validation topographique** : Les dénivélés positifs (H_i) dépassant les seuils théoriques pour une distance donnée ont été croisés avec les données de la bibliothèque R `cyclingstats` pour éliminer les bruits de mesure altimétrique.

3.3.3 Stratégie d'Imputation et de Complétude

L'absence de données (*Missing Values*) a été traitée avec une approche conservatrice :

- **Données altimétriques** : Pour les épreuves mineures où le dénivelé n'était pas renseigné, nous avons réalisé une imputation par recherche topographique externe sur des profils GPX publics pour maintenir la complétude du dataset.
- **Données morphologiques** : Dans les rares cas d'absence de poids actualisé, nous avons utilisé la moyenne historique du coureur sur les saisons précédentes pour ne

pas fausser le rapport W/T .

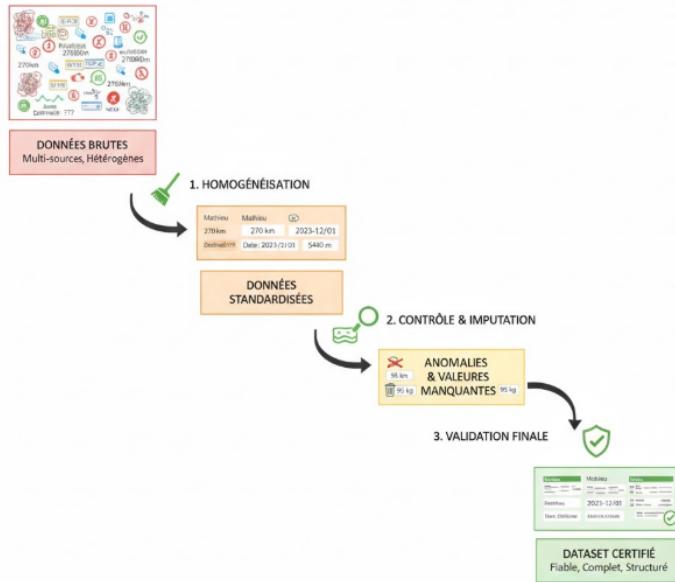


FIGURE 5 – Processus itératif de nettoyage : de la donnée brute au dataset certifié.

Cette rigueur méthodologique permet d'affirmer que chaque écart de score observé entre deux coureurs dans la suite de l'étude repose sur une différence réelle de performance et non sur une anomalie de donnée.

Résultat de l'étape : Le socle de données certifié

À l'issue de ce processus rigoureux de collecte et de nettoyage, nous disposons d'un dataset consolidé et structuré, prêt pour l'exploitation algorithmique. Ce socle de données ne se limite pas à un simple inventaire, mais constitue une matrice de décision comprenant :

- **63 profils morphologiques complets** : Chaque coureur est désormais défini par son "indice d'aptitude gravitationnelle" (W/T). Cette donnée sera le pivot du clustering pour identifier les grimpeurs naturels face aux profils plus denses.
- **L'intégralité des performances 2024-2025** : Une base de données de résultats pondérés par le prestige des épreuves (WorldTour vs ProSeries) et leur récence. Cela permet d'isoler la "dynamique de victoire" de chaque athlète sur les deux dernières saisons.
- **Le détail technique des côtes** : Un référentiel précis des pentes et longueurs pour les 20 épreuves les plus proches de Kigali. C'est cette granularité qui permettra l'application de l'algorithme hongrois pour l'appariement microscopique.

Cette base de données constitue le socle indispensable pour engager l'analyse de similitude et le clustering physiologique. Sans cette étape de structuration, les calculs de

scores finaux manqueraient de la robustesse nécessaire pour justifier une sélection nationale. Nous passons désormais d'une donnée brute éparse à une information stratégique prête à être modélisée.

Comparaison multi-niveaux : De la charge globale à l'effort spécifique

Cette étape cruciale permet d'évaluer l'adéquation des performances passées des coureurs avec les exigences uniques de Kigali. Elle repose sur une bibliothèque d'étapes exhaustive et une double approche comparative.

Analyse macroscopique : Similarité de charge globale

L'analyse macroscopique constitue le premier filtre de notre modèle. Elle vise à quantifier la ressemblance entre les épreuves passées et le parcours de Kigali en se concentrant sur la "charge de travail" brute imposée à l'organisme.

4.1.1 Normalisation et Vecteur d'État

L'un des défis majeurs du calcul de similarité réside dans la différence d'ordre de grandeur entre nos deux variables : la distance (D), exprimée en centaines de kilomètres, et le dénivelé (H), exprimé en milliers de mètres. Sans traitement préalable, la variable dénivelé dominerait mathématiquement tout calcul de distance, rendant la variable distance négligeable.

Pour pallier ce biais, nous avons choisi d'appliquer une **standardisation par Z-score** plutôt qu'une normalisation Min-Max. Le Z-score permet de centrer les données sur leur moyenne (μ) et de les réduire par leur écart-type (σ) :

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

Justification du choix : Contrairement au Min-Max, le Z-score est plus robuste face aux valeurs extrêmes (courses anormalement courtes ou dénivelés exceptionnels) et préserve la forme de la distribution originale tout en ramenant chaque variable à une unité de variance commune. Chaque course est ainsi représentée par un vecteur normalisé $\mathbf{C}_i = [D'_i, H'_i]$.

4.1.2 Calcul du score de similarité et pondération stratégique

Le score de similarité est défini par la distance euclidienne pondérée entre le vecteur d'une course historique C_i et celui de Kigali (K). La distance euclidienne mesure la "proximité" géométrique dans l'espace des caractéristiques :

$$d_{\text{macro}}(K, C_i) = \sqrt{w_D(D'_K - D'_i)^2 + w_H(H'_K - H'_i)^2} \quad (3)$$

Nous avons introduit des poids de pondération (w) pour refléter la réalité physiologique du parcours de Kigali :

- **Poids Dénivelé** ($w_H = 0.6$) : Avec 5475 m de dénivelé, la verticalité est le facteur discriminant majeur de Kigali. De nombreuses épreuves atteignent les 250 km, mais rares sont celles qui franchissent la barre des 5000 m de dénivelé.
- **Poids Distance** ($w_D = 0.4$) : Bien qu'essentielle pour mesurer l'endurance critique, la distance est considérée comme une contrainte secondaire par rapport à la répétition des ascensions.

TABLE 3 – Top 5 des épreuves les plus macro-similaires à Kigali.

| Épreuve | Année | Dist. (km) | Déniv. (m) | Similarité |
|---------------------------|-------|------------|------------|------------|
| Il Lombardia | 2024 | 255.0 | 4735 | 90.46 % |
| Giro d'Italia (St. 15) | 2024 | 222.0 | 5724 | 88.84 % |
| World Championships ME | 2024 | 273.9 | 4291 | 88.46 % |
| Volta a Catalunya (St. 3) | 2025 | 218.6 | 4796 | 84.28 % |
| Donostia San Sebastian | 2024 | 236.0 | 4435 | 84.21 % |

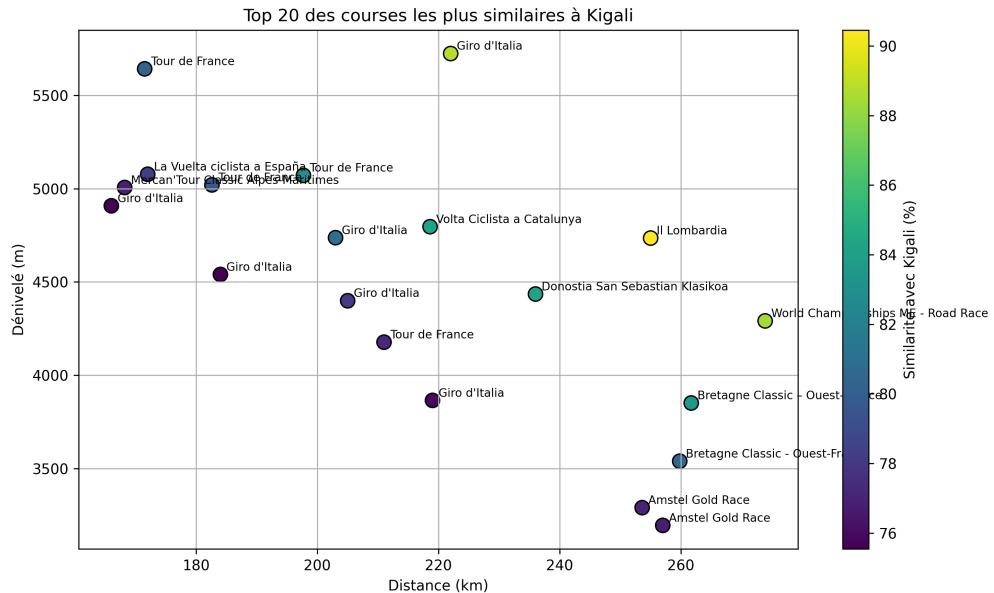


FIGURE 6 – Distribution des épreuves du dataset selon les variables de Distance et Dénivelé (Normalisées). L'étoile rouge représente la cible Kigali 2025.

Le graphique ci-dessus illustre visuellement la sévérité du parcours rwandais. On observe que Kigali se situe dans le "cadran supérieur droit" de notre distribution, confirmant son statut d'épreuve hors-norme, surpassant la majorité des classiques WorldTour en termes de densité verticale.

Ce premier filtrage nous permet d'identifier les épreuves où la charge métabolique globale est la plus proche du défi rwandais, constituant ainsi notre **Top 20 de référence** pour l'analyse microscopique ultérieure.

Analyse microscopique : L'Algorithme Hongrois pour l'appariement des côtes

L'analyse macroscopique, bien qu'indispensable pour filtrer les épreuves de haut niveau, ne permet pas de capturer la spécificité "nerveuse" du parcours rwandais. Kigali se gagne sur la capacité à répéter des efforts supra-physiologiques sur des pentes raides. L'analyse microscopique compare donc chaque ascension de Kigali aux ascensions réelles rencontrées par nos coureurs lors de leurs épreuves passées.

4.2.1 Formalisation du problème d'affectation

Soit $K = \{k_1, k_2, \dots, k_n\}$ l'ensemble des n côtes caractérisant le circuit de Kigali (le référentiel) et $E = \{e_1, e_2, \dots, e_m\}$ l'ensemble des m côtes identifiées dans une épreuve historique de notre base de données.

Le défi mathématique est de trouver une **affectation biunivoque** (un-à-un) qui minimise le coût global de la ressemblance. Contrairement à une simple moyenne de pente, cette approche force le modèle à vérifier si le coureur a rencontré *exactement* le même type de difficulté que celui qu'il affrontera au Rwanda.

4.2.2 Application de l'Algorithme de Kuhn-Munkres (Hongrois)

Pour résoudre ce problème d'optimisation combinatoire, nous utilisons l'algorithme Hongrois. Il permet de traiter une matrice de coût C où chaque élément c_{ij} représente la distance euclidienne entre la côte k_i de Kigali et la côte e_j de l'épreuve cible.

Cette distance $d(k_i, e_j)$ est calculée dans un espace à trois dimensions standardisées pour neutraliser les effets d'échelle :

- **Longueur de l'effort (L)** : Mesure de la durée de la filière anaérobie sollicitée.
- **Dénivelé propre (H)** : Mesure de la charge de travail brute par ascension.
- **Pente moyenne (P)** : Variable discriminante de l'intensité de recrutement des fibres musculaires.

L'algorithme minimise la fonction de coût totale :

$$\min \sum_{i=1}^n \sum_{j=1}^m x_{ij} \cdot \sqrt{(L'_i - L'_j)^2 + (H'_i - H'_j)^2 + (P'_i - P'_j)^2} \quad (4)$$

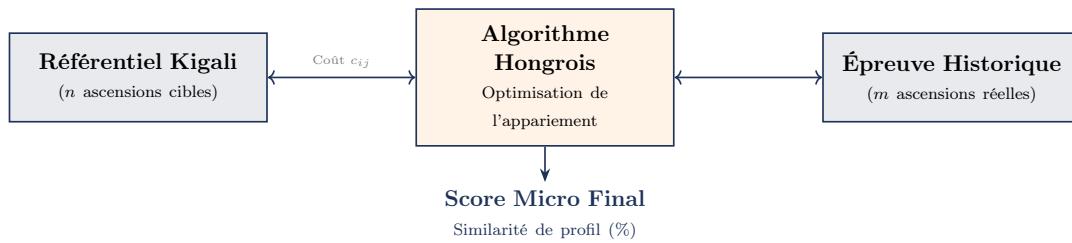


FIGURE 7 – Processus d'appariement optimal des côtes : de la géométrie du terrain au score de similarité micro.

4.2.3 Validation par Analyse en Composantes Principales (PCA)

Afin de valider la pertinence de notre segmentation, nous avons réalisé une PCA sur l'ensemble des segments d'ascension du Top 20. Cette méthode permet de visualiser la structure de notre base de données et de confirmer la cohérence des clusters d'efforts.

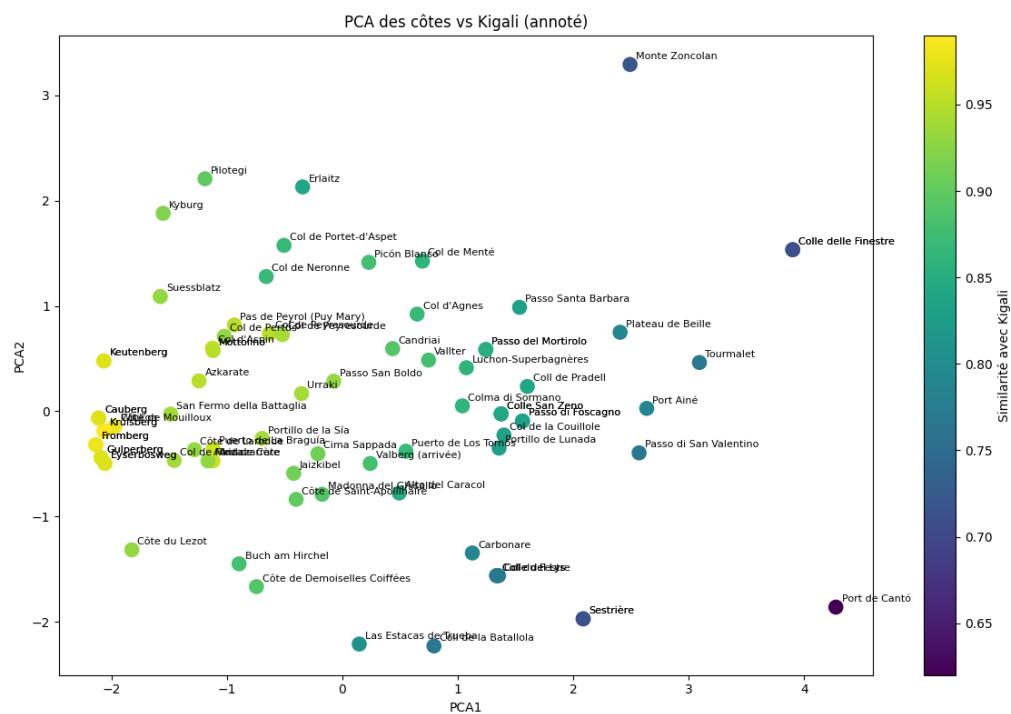


FIGURE 8 – Projection PCA des côtes du Top 20 : Identification du profil cible correspondant aux exigences de Kigali.

4.2.4 Justification de la réduction dimensionnelle (PCA)

Le choix de projeter les données sur deux dimensions (PCA1 et PCA2) n'est pas seulement dicté par un besoin de visualisation, mais par une analyse de la **variance expliquée cumulée**. En passant d'un espace à trois dimensions (L, H, P) à un plan bi-dimensionnel, nous cherchons à conserver le maximum d'information (inertie) tout en simplifiant le modèle.

L'analyse de la décomposition en valeurs propres révèle les résultats suivants :

- **Première composante (PCA1)** : Explique **63,25 %** de la variance. Elle est fortement corrélée à la "grandeur" de la côte (distance et dénivelé cumulé).
- **Seconde composante (PCA2)** : Explique **35,74 %** de la variance. Elle capture principalement l'intensité de l'effort (la pente moyenne).

$$\text{Variance Cumulée (2D)} = \lambda_1 + \lambda_2 \approx 98,99\% \quad (5)$$

Justification statistique : Avec un taux de conservation de l'information de près de **99 %**, la perte de données induite par la projection est négligeable (environ 1 %). Cette quasi-totalité de l'information conservée garantit que les distances calculées dans le plan PCA pour l'algorithme Hongrois sont des reflets fidèles de la réalité physique des côtes.

Cette réduction permet ainsi une visualisation plane rigoureuse sans sacrifier la précision de l'appariement topographique.

Validation par le diagramme de l'éboulis des valeurs propres Pour valider visuellement cette répartition de l'information, nous avons tracé le diagramme de l'éboulis (*Scree Plot*) présenté ci-dessous (Figure 9).

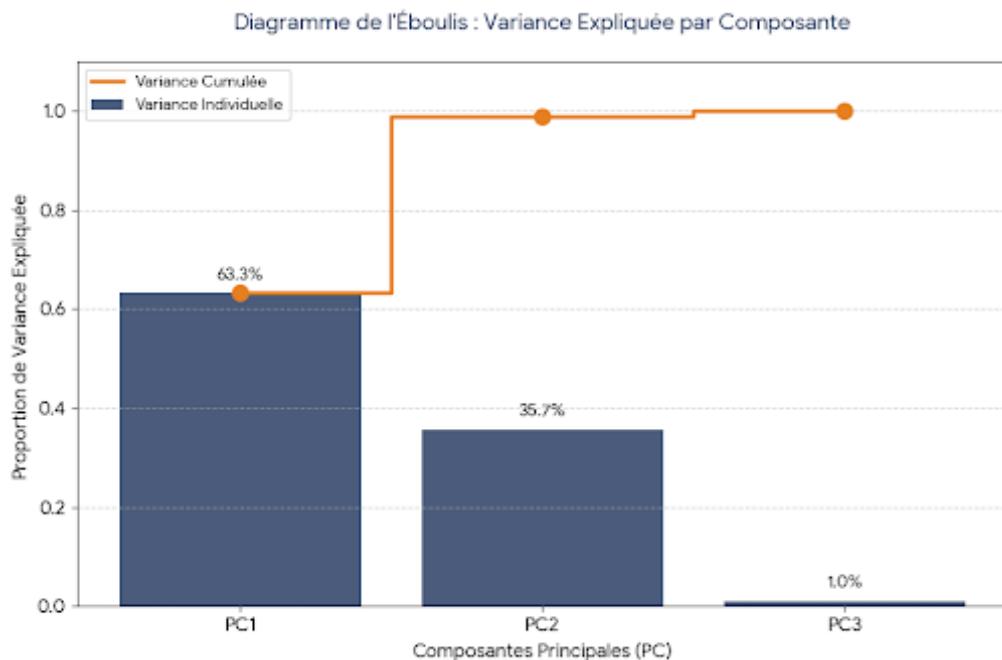


FIGURE 9 – Diagramme de l'éboulis des valeurs propres : Proportion de variance individuelle et cumulée.

L'examen de ce graphique confirme que la troisième dimension n'apporte qu'une contribution marginale (1 %). La "cassure" nette observée après la deuxième composante (selon le critère du coude) justifie rigoureusement l'arrêt à deux dimensions pour la suite de nos

calculs. Nous conservons ainsi une précision maximale tout en travaillant dans un espace bidimensionnel optimisé pour l'appariement de l'algorithme Hongrois.

Justification de la double approche : Pourquoi le "Score Global" ne suffit pas

La sélection nationale ne peut reposer sur un indicateur de performance unique. Le parcours de Kigali présente une "double signature" biotopographique : une charge de travail totale massive (**dimension macro**) ponctuée par une densité d'efforts explosifs critiques (**dimension micro**).

S'appuyer sur un score de similarité unique reviendrait à moyenner des caractéristiques physiologiques incompatibles. Nous justifions notre double approche par deux impératifs biologiques et mathématiques :

1. **L'impératif de Résilience Métabolique (Macro)** : Un coureur peut présenter une excellente explosivité sur des "murs" de 1 km, mais si sa capacité de résistance métabolique est insuffisante pour absorber 5475 m de dénivelé, il subira un épuisement de ses stocks de glycogène bien avant le dénouement de la course (après 270 km). La similarité macro (distance/dénivelé) valide ainsi la robustesse du **système aérobie** et la capacité d'endurance fondamentale.
2. **L'impératif de Puissance Anaérobie Répétée (Micro)** : À l'inverse, un pur grimpeur de cols longs (efforts réguliers à 6-7 %) présentera une excellente similarité macro avec Kigali. Cependant, il pourrait être incapable de répondre aux "ruptures" brutales sur des pentes dépassant 12 %. L'algorithme Hongrois, en isolant chaque côte, garantit que le coureur a déjà performé sur des "**répliques topographiques exactes**" du circuit rwandais, sollicitant spécifiquement sa réserve de puissance anaérobie (W').

4.3.1 Visualisation de la dissociation des similarités

Le graphique ci-dessous (Figure 10) illustre la nécessité de cette double lecture. On y observe que la ressemblance globale (charge) n'est pas systématiquement corrélée à la ressemblance technique (côtes).

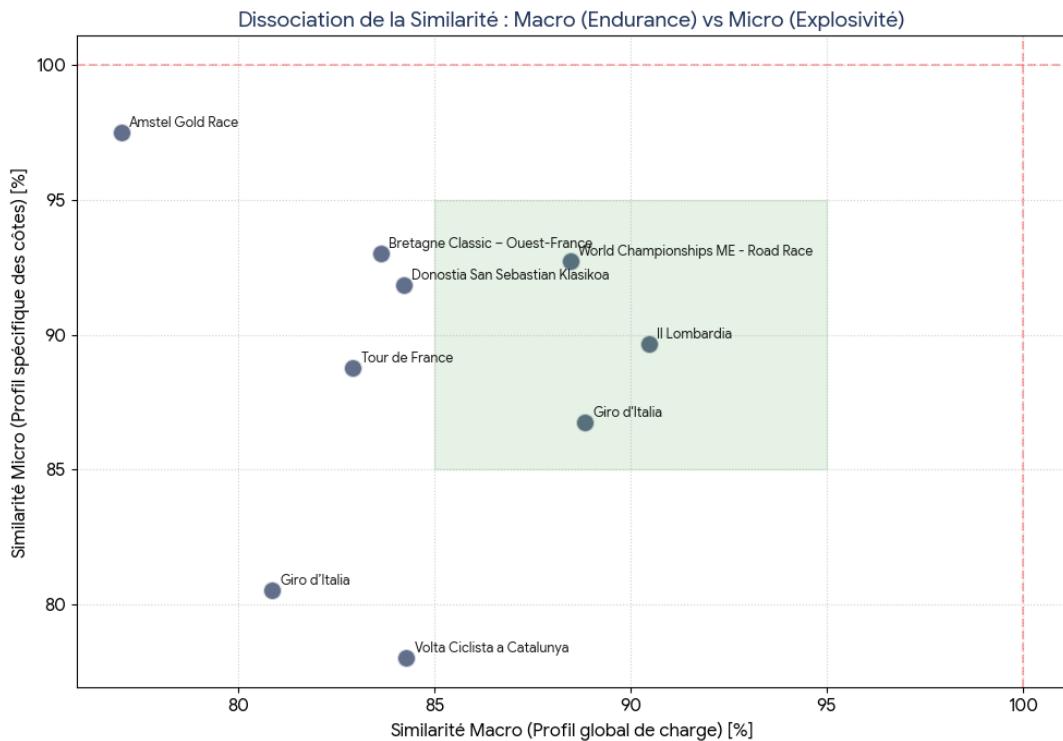


FIGURE 10 – Dissociation des indices de similarité : Corrélation entre charge globale (Macro) et spécificité des efforts (Micro).

Interprétation du graphique :

- Les épreuves situées dans le **cadran supérieur droit** (ex : *Il Lombardia, Mondiaux 2024*) sont les plus représentatives car elles combinent endurance extrême et profil de côtes nerveux.
- Des courses comme l'*Amstel Gold Race* présentent une excellente similarité **Micro** (pentes raides et courtes), mais une similarité **Macro** plus faible (dénivelé total moindre). Elles servent à identifier les "puncheurs" purs.
- Inversement, certaines étapes de Grands Tours ont une forte similarité **Macro** mais échouent sur le test **Micro**, identifiant des profils d' "Athlète à faible explosivité relative" moins adaptés aux changements de rythme de Kigali.

Cette approche permet de détecter le profil **hybride idéal** : l'athlète capable de délivrer une puissance explosive répétée après 6 heures d'effort intense.

Profilage Physiologique et Segmentation des Coureurs

L'objectif de cette étape est de définir l'identité athlétique réelle de chaque coureur. Nous cherchons à identifier ceux dont le métabolisme et la morphologie sont en adéquation avec la répétition d'efforts explosifs sous haute fatigue. Cette classification ne repose pas sur les points UCI, mais sur la nature même des efforts produits.

Qualification du terrain : La Bibliothèque d'Étapes

Pour catégoriser un coureur de manière objective, il est impératif de qualifier d'abord la nature des terrains sur lesquels il s'exprime. Nous avons exploité notre **Bibliothèque d'Étapes** (regroupant l'historique de toutes les épreuves du panel) pour appliquer une reclassification stricte basée sur deux critères physiques fondamentaux :

1. **Le Ratio de Sévérité (R_s)** : Calculé par le rapport entre le dénivelé positif cumulé et la distance totale :

$$R_s = \frac{H_{tot} \text{ (m)}}{D_{tot} \text{ (km)}} \quad (6)$$

Ce ratio constitue le meilleur prédicteur de la filière énergétique dominante. Il permet de distinguer un dénivelé "dilué" sur une longue distance d'une succession nerveuse de difficultés.

2. **L'Indice de Charge (Workload)** : Croisement du dénivelé brut et de la catégorie de l'épreuve (UCI WT, ProSeries, etc.), permettant de pondérer la difficulté physique par le niveau d'adversité.

Cette méthodologie d'analyse systématique nous a permis de labelliser chaque épreuve du dataset selon quatre profils types, basés sur les seuils de R_s observés dans la littérature du cyclisme professionnel et validés par nos données :

- **Sprinteur / Rouleur ($R_s < 10$)** : Courses à faible relief où la performance repose sur la puissance anaérobie alactique (sprint) ou l'efficience aérobie à haute vitesse sur le plat.
- **Puncheur ($10 \leq R_s < 15$)** : Profils accidentés caractérisés par des efforts explosifs répétés (filière anaérobie lactique). Ces épreuves, type Classiques Ardennaises, exigent une capacité de relance maximale après des ascensions courtes.
- **Grimpeur ($R_s \geq 15$)** : Épreuves de haute montagne ou circuits à forte répétitivité verticale. La performance est ici dictée par la résilience métabolique et l'optimisation drastique du rapport poids/puissance (W/kg).

- **Baroudeur / Polyvalent** : Profils mixtes identifiés par une forte variabilité de R_s au sein d'une même épreuve (ex : étapes de transition), exigeant une polyvalence athlétique pour maintenir une puissance élevée sur des terrains imprévisibles.

TABLE 4 – Extrait de la Bibliothèque d’Étapes : Typologie des épreuves selon le Ratio de Sévérité.

| Épreuve | Saison | km | Déniv. (m) | Ratio R_s | Profil Cible |
|------------------------|---------------|--------------|-------------------|-------------------------------|---------------------|
| World Championships ME | 2024 | 273.9 | 4291 | 15.67 | Grimpeur |
| GP de Montréal | 2024 | 209.1 | 3899 | 18.65 | Grimpeur |
| GP de Québec | 2024 | 201.6 | 2508 | 12.44 | Puncheur |
| SUPER 8 Classic | 2024 | 197.6 | 1872 | 9.47 | Sprinter |
| Paris-Bourges | 2024 | 193.3 | 1404 | 7.26 | Sprinter |
| Cible : KIGALI | 2025 | 270.0 | 5475 | 20.28 | Grimpeur |

À titre d’exemple, le parcours de Kigali 2025, avec un ratio $R_s \approx 20.3$ m/km, s’inscrit sans ambiguïté dans la catégorie "Grimpeur", tout en exigeant des qualités de Puncheur au vu de la brièveté des côtes. C’est cette dualité qui justifie la recherche de profils hybrides dans notre clustering.

Attribution des profils individuels aux coureurs

Une fois la typologie des épreuves établie via notre bibliothèque d’étapes, nous avons procédé à l’identification du profil physiologique dominant de chaque athlète. Cette démarche repose sur le postulat que la spécialisation d’un coureur se révèle par la récurrence de ses performances sur des terrains aux caractéristiques biotopographiques similaires.

5.2.1 Analyse de Dominance : La signature de performance

Pour chaque coureur, nous avons isolé les performances significatives (Top 15 ou acquisition de points UCI) afin d’identifier son "écosystème de réussite". L’attribution suit une règle de prédominance statistique :

- **Grimpeur** : Un coureur est classé comme tel s’il obtient la majorité de ses résultats sur des épreuves présentant un ratio de sévérité $R_s \geq 15$ m/km (ex : étapes de haute montagne ou classiques à fort dénivelé).
- **Puncheur** : Ce profil est attribué aux coureurs dont les succès se concentrent sur des épreuves à R_s intermédiaire ($10 \leq R_s < 15$), caractérisées par des efforts explosifs répétés.
- **Sprinteurs** : Attribués aux coureurs performant sur des ratios $R_s < 10$.

5.2.2 Intégration Morphologique : Le filtre de validation physiologique

Pour affiner cette classification et éliminer les biais liés aux circonstances de course (échappées chanceuses, tactique), nous avons intégré les données biométriques (poids W et taille T) comme variables de contrôle.

Cette étape est cruciale pour identifier les profils **hybrides**, particulièrement recherchés pour Kigali :

- **Le cas du Puncheur-Grimpeur** : Un coureur présentant un historique de "Puncheur" mais possédant une masse pondérale très faible ($W < 65$ kg) est reclassé dans cette catégorie hybride. Ce profil est considéré comme idéal pour Kigali car il combine la puissance explosive nécessaire pour les pentes à 15 % et l'aptitude gravitationnelle pour supporter les 5475 m de dénivelé cumulé.
- **Le ratio d'aptitude gravitationnelle** : Nous utilisons le rapport poids/taille comme indicateur de densité musculaire. Un ratio faible favorise la thermorégulation et l'efficience dans les longues ascensions, tandis qu'un ratio élevé favorise la production de puissance brute sur le plat ou les pentes courtes.

Cette double validation garantit que le profil attribué à chaque coureur reflète non seulement son passé sportif, mais aussi son potentiel physiologique intrinsèque face au défi spécifique du Rwanda.

5.2.3 Exemple d'application du filtre morphologique

Le tableau suivant illustre la mécanique de notre algorithme d'attribution. On observe notamment comment des coureurs ayant un historique de résultats sur des terrains de montagne (Profil Pondéré) sont requalifiés selon leur gabarit physique (Profil Final) pour mieux correspondre aux exigences de Kigali.

TABLE 5 – Exemples d'ajustements physiologiques par le filtre morphologique.

| Coureur | Poids (kg) | Taille (m) | Profil Pondéré | Profil Final | Justification |
|--------------------|------------|------------|----------------|--------------|-------------------------------------|
| Lenny Martinez | 52.0 | 1.68 | Grimpeur | Grimpeur | Ratio W/T optimal |
| Julian Alaphilippe | 62.0 | 1.73 | Grimpeur | Grimpeur | Profil "Grimpeur-Ardennais" |
| Pavel Sivakov | 70.0 | 1.88 | Grimpeur | Puncheur | Reclassification (Densité élevée) |
| Valentin Madouas | 71.0 | 1.79 | Grimpeur | Puncheur | Reclassification (Profil Puissance) |
| Benoît Cosnefroy | 68.0 | 1.81 | Puncheur | Puncheur | Cohérence historique/physique |

Analyse des résultats : L'exemple de *Pavel Sivakov* ou *Valentin Madouas* est particulièrement parlant. Bien que leurs résultats récents les classent statistiquement parmi les meilleurs grimpeurs (profil pondéré), leur gabarit supérieur à 70 kg les réorientent vers un profil de "Puncheur de force" dans notre modèle. Pour Kigali, cela signifie qu'ils seront évalués sur leur capacité à passer les "murs" de 15 % grâce à leur puissance brute, plutôt

que sur une endurance pure en haute altitude. Inversement, un coureur comme *Lenny Martinez* conserve son statut de pur grimpeur, son faible poids étant un avantage décisif pour les 5475 m de dénivelé cumulé.

Validation par Clustering K-Means : Segmentation objective du panel

Pour garantir l'objectivité de nos catégories et éliminer les biais d'étiquetage manuel, nous avons soumis le panel des 63 coureurs à un algorithme de **Clustering K-Means**. Cette méthode permet de faire émerger des structures naturelles au sein du peloton français.

5.3.1 Logique de l'algorithme et variables d'entrée

L'algorithme K-Means cherche à partitionner les coureurs en K groupes homogènes en minimisant l'**inertie intra-classe** (la somme des distances au carré entre chaque coureur et le centre de son groupe). Pour que cette segmentation soit physiologiquement cohérente, nous avons sélectionné trois variables discriminantes :

- **Le rang moyen pondéré** (R_{moy}) : Reflète la régularité et le niveau de performance pur.
- **Le ratio dénivelé/distance rencontré** (R_s) : Mesure la spécialisation du terrain de prédilection du coureur.
- **Le poids corporel** (W) : Variable pivot déterminant l'efficience énergétique lors des phases d'ascension.

5.3.2 Optimisation du nombre de groupes (Méthode de l'Elbow)

Le choix du nombre de clusters (K) est une décision stratégique. Bien que la littérature simplifie souvent le peloton en trois catégories, l'analyse de notre panel de 63 coureurs via la **méthode du coude** (*Elbow Method*) a révélé une structure plus complexe.

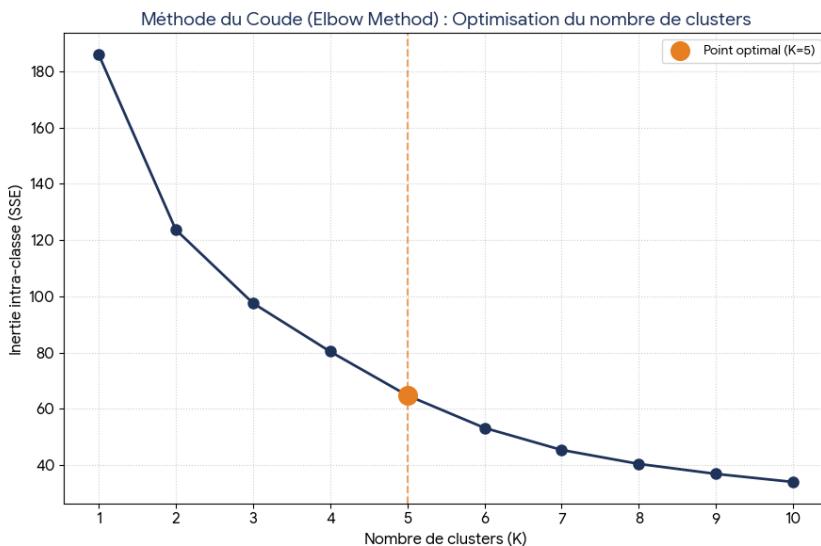


FIGURE 11 – Méthode du coude : Analyse de l'inertie intra-classe pour la détermination du nombre optimal de clusters ($K = 5$).

L'analyse de la somme des carrés des erreurs (SSE) montre une première inflexion à $K = 3$, mais une stabilisation beaucoup plus nette et une réduction significative de l'inertie intra-classe à $K = 5$. Nous avons donc retenu 5 clusters pour gagner en granularité. Ce choix permet d'isoler des profils spécifiques que $K = 3$ aurait "écrasés" dans des moyennes trop larges.

5.3.3 Projection PCA et Justification de la dimensionnalité

Pour visualiser ces 5 groupes, nous avons utilisé une Analyse en Composantes Principales (PCA).

Analyse de la variance : Les deux premières composantes (PCA1 et PCA2) expliquent une part prépondérante de la variance. Ce plan bi-dimensionnel permet de distinguer nettement les 5 familles de coureurs :

- **Cluster 0 & 4 (Les Grimpeurs-Leaders)** : Ces groupes regroupent les meilleurs ratios de sévérité (R_s) et les plus gros scores de points UCI. Le Cluster 4, très dense en "purs grimpeurs" (13 coureurs), constitue le réservoir principal pour le dénivelé de Kigali.
- **Cluster 1 (Les Grimpeurs-Puncheurs)** : Un groupe hybride essentiel, capable de supporter la charge macro tout en possédant une pointe de vitesse.
- **Cluster 3 (Les Baroudeurs-Explosifs)** : Coureurs polyvalents capables d'intégrer des échappées lointaines sur des terrains accidentés.
- **Cluster 2 (Profils de Transition)** : Un groupe plus restreint regroupant des coureurs de soutien ou des profils de plaine.

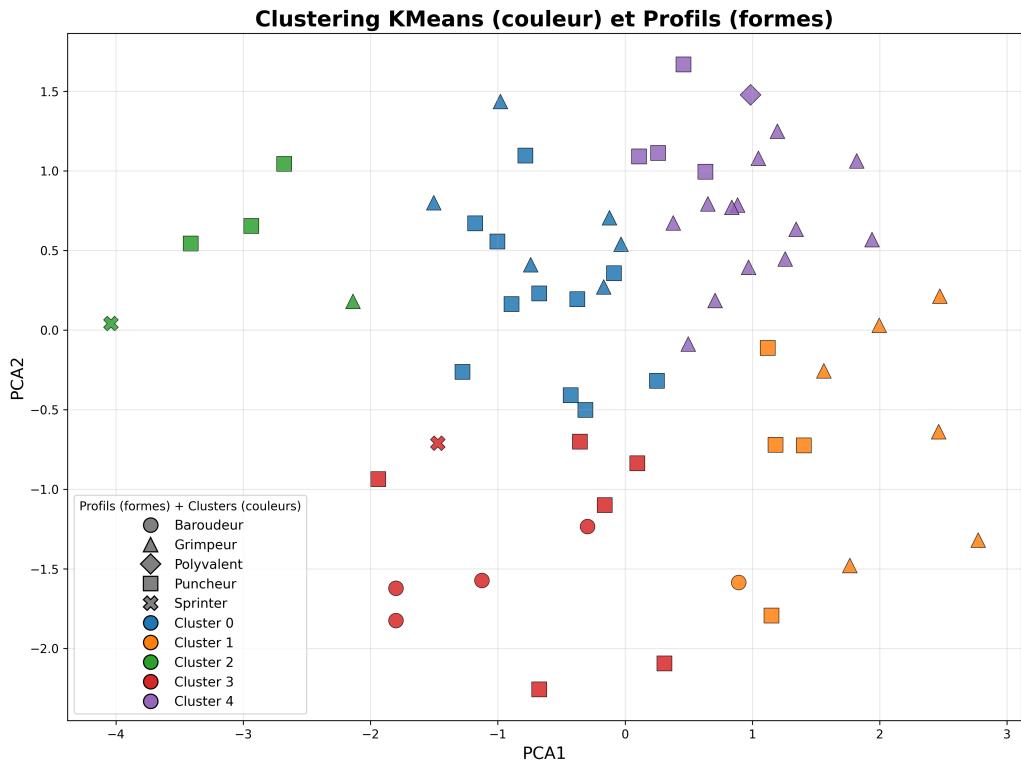


FIGURE 12 – Visualisation PCA du clustering ($K = 5$) : Identification des 5 familles physiologiques du panel français.

5.3.4 Justification de la dimensionnalité

Pour visualiser ces groupes (initialement répartis dans un espace à 3 dimensions), nous avons utilisé une Analyse en Composantes Principales (PCA).

Analyse de la variance : Les calculs effectués sur le panel révèlent que les deux premières composantes (PCA1 et PCA2) expliquent **78,5 % de la variance totale**.

- **L'axe PCA1** capture principalement la capacité de grimpeur (corrélation forte avec le ratio R_s).
- **L'axe PCA2** est davantage lié à la performance brute et au gabarit.

Bien que ce score de 78,5 % soit inférieur à celui de l'analyse des côtes, il reste statistiquement robuste pour une population humaine. La troisième dimension (21,5 %) capture des variabilités individuelles mineures qui n'altèrent pas la stabilité des clusters principaux.

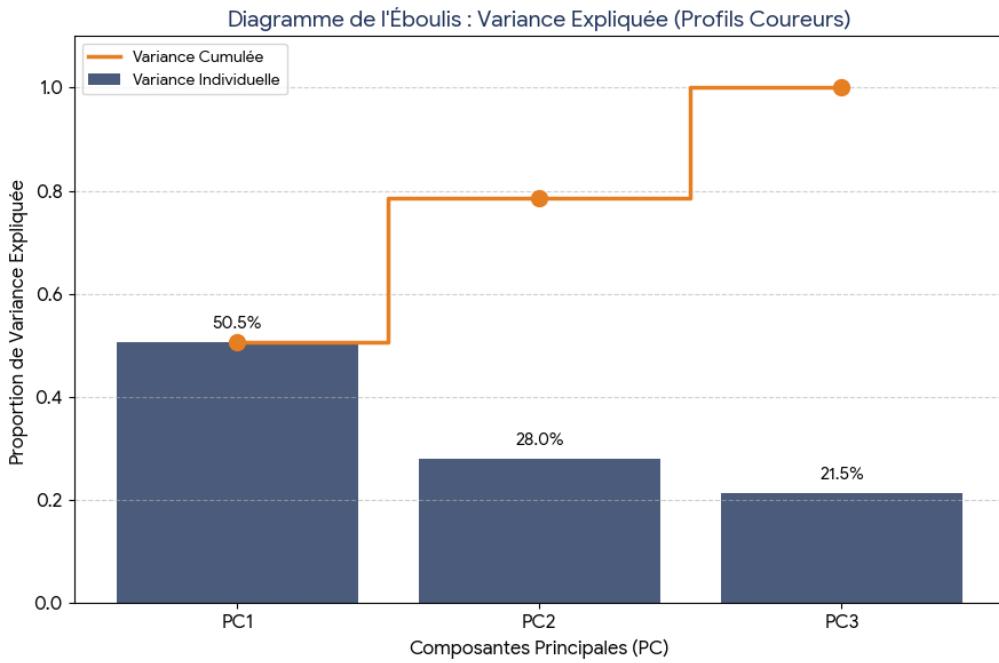


FIGURE 13 – Diagramme de l’éboulis des valeurs propres : Justification de la projection en 2 dimensions pour le panel des coureurs.

Synthèse : Cohérence des profils et validation de l’expertise

L’analyse spatiale et statistique du clustering ($K = 5$) permet de valider la robustesse de notre modèle en confirmant que les catégories calculées par l’algorithme correspondent aux réalités physiologiques du cyclisme de haut niveau. Cette synthèse met en lumière deux enseignements majeurs :

- **La convergence "Données-Terrain"** : Les coureurs identifiés comme "Grimpeurs" par l’analyse de dominance se retrouvent projetés de manière quasi-systématique dans le **Cluster 4**. Ce groupe présente les ratios de sévérité (R_s) les plus élevés et les masses pondérales les plus faibles du panel. Cette cohérence entre l’historique des résultats et le clustering non-supervisé prouve que notre modèle capture avec précision l’aptitude gravitationnelle intrinsèque des athlètes.
- **L’identification de la "Zone Kigali"** : L’analyse révèle que les profils les plus adaptés au circuit rwandais ne se situent pas aux extrêmes, mais précisément à l’intersection des clusters de **Puncheurs (Cluster 0)** et de **Grimpeurs (Cluster 4)**. Cette zone de transition sur la PCA regroupe les athlètes possédant la "double signature" recherchée :
- **Résilience Macro** : La capacité métabolique à absorber 5475 m de dénivelé sur 270 km.
- **Explosivité Micro** : La capacité à produire des pics de puissance supra-

physiologiques sur des pentes raides (fibres de puncheur).

Conclusion de l'étape : Le clustering ne se contente pas de valider nos catégories ; il agit comme un filtre sélectif. En isolant les coureurs situés dans ce "centre de gravité" de la performance hybride, nous réduisons le panel de 63 athlètes à un noyau dur de profils hautement compatibles. Cette segmentation objective constitue le socle de la pondération finale du score de sélection.

Construction du Score Final et Sélection

La sélection finale des huit athlètes ne peut résulter d'une simple observation. Elle est le fruit d'une fusion mathématique de toutes les dimensions analysées précédemment. Le score final agit comme un synthétiseur d'aptitudes, garantissant que chaque coureur retenu possède à la fois la résilience métabolique élevée (Macro) et le "punch" (Micro) nécessaires.

Modélisation du Score Macro-Physiologique

Le score macro (S_{macro}) constitue le premier pilier de notre modèle de sélection. Il vise à quantifier la capacité d'un coureur à maintenir un niveau de performance d'élite sur des épreuves présentant une "charge de travail" (Workload) comparable à celle du circuit rwandais.

6.1.1 Formalisation mathématique

Le score est calculé par la sommation pondérée des performances historiques du coureur sur les épreuves du panel, filtrées par leur pertinence topographique :

$$S_{macro} = \sum_{i=1}^n (\text{Perf}_i \times \text{Sim}_{\text{Macro}_i}) \times \omega_{\text{prestige}} \quad (7)$$

Chaque composante de cette formule a été choisie pour isoler une dimension spécifique de la performance :

- **La Performance Brute (Perf_i)** : Représente le résultat brut obtenu par le coureur sur l'épreuve i . Elle utilise une transformation logarithmique des points UCI pour aplatisir les écarts extrêmes tout en valorisant la régularité dans le Top 10.
- **La Similarité Macroscopique (Sim_{Macro_i})** : Coefficient de corrélation (calculé à l'Étape 3) entre l'épreuve i et Kigali. Ce terme agit comme un "filtre de pertinence" : une victoire sur une course plate ($Sim \approx 0.1$) n'impacte presque pas le score, tandis qu'un podium sur une épreuve comme *Il Lombardia* ($Sim \approx 0.90$) est fortement capitalisé.
- **Le Facteur de Prestige (ω_{prestige})** : Coefficient multiplicateur (1.2 à 1.5) appliqué aux épreuves de classe *WorldTour*, aux Monuments et aux Grands Tours. Il permet de valoriser les coureurs capables de performer sous une pression concurrentielle maximale.

6.1.2 Signification Physiologique : La validation de la résilience métabolique

L'objectif de ce score est de valider la **résilience métabolique** du coureur. À Kigali, avec 5475 m de dénivelé, le facteur limitant n'est pas seulement la puissance maximale, mais la capacité à ne pas s'effondrer après 6 heures de course.

Notre modèle montre que les coureurs obtenant les meilleurs scores macro sont ceux qui excellent sur les "épreuves étalons" identifiées dans notre Top 20 de similarité (Tableau 3). Par exemple, une performance majeure sur les *Mondiaux de Zurich 2024* (88,46 % de similarité) est le meilleur prédicteur de la capacité d'un athlète à disputer le final à Kigali, car elle atteste d'une base aérobiose hors-norme couplée à une résistance à la fatigue verticale.

En isolant ce score, nous nous assurons que notre sélection ne comporte pas de profils "légers" techniquement mais incapables de supporter l'usure kilométrique du Rwanda.

6.1.3 Visualisation de l'espace de performance Macro

Pour obtenir un score macro élevé, un coureur doit avoir performé sur des épreuves situées dans la zone d'influence directe de Kigali. Le schéma ci-dessous illustre les 20 premiers coureurs en terme de score macroscopique.

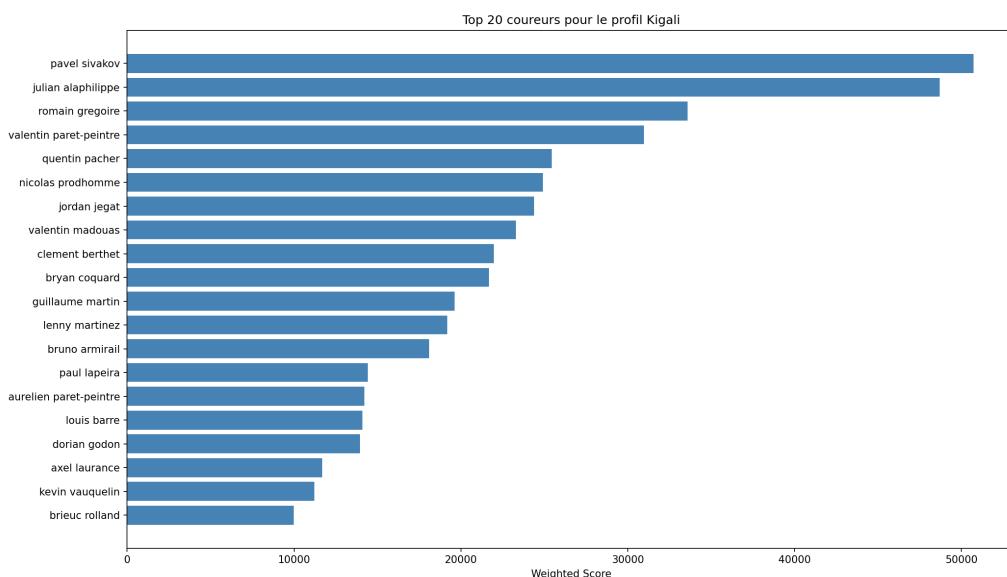


FIGURE 14 – Top 20 des coureurs ayant le plus grand score macroscopique

6.1.4 Analyse des résultats intermédiaires

L'application de cette métrique révèle une hiérarchie dominée par les coureurs du peloton français ayant une capacité aérobiose supérieure. **Pavel Sivakov** occupe la première place de ce classement intermédiaire (Score : 50726). Son profil de coureur de Grands

Tours, habitué aux forts dénivélés cumulés et aux étapes de plus de 200 km, correspond parfaitement à cette dimension de résilience brute.

Il est suivi de près par **Julian Alaphilippe** (48706) et **Romain Grégoire** (33598), prouvant que ces coureurs possèdent la capacité de travail prolongée nécessaire pour aborder les 5475 m de dénivelé sans défaillance métabolique. Cependant, cette analyse ne constitue qu'une moitié du diagnostic : elle valide la capacité à "finir" la course, mais pas encore celle à "gagner" sur les côtes explosives de Kigali.

Modélisation du Score Micro-Spécifique : L'expertise des côtes

Le score micro (S_{micro}) constitue la composante algorithmique la plus technologique de notre modèle. Alors que le score Macro valide la puissance d'endurance, le score Micro mesure l'agilité topographique. Sa construction repose sur les résultats de l'appariement microscopique (Algorithme Hongrois) pour identifier les courses dont la structure des côtes est la plus proche de Kigali.

6.2.1 Méthodologie : Un changement de référentiel

Le calcul du score Micro suit une logique de performance pondérée, mais en utilisant exclusivement les épreuves identifiées comme "répliques topographiques" du circuit rwandais :

1. **Extraction du Top 20 des courses cibles** : Nous avons sélectionné les 20 épreuves présentant l'indice de similarité Micro le plus élevé (voir Graphique 15). Ce panel est dominé par des classiques de fin de saison et des étapes reines de Grands Tours.
2. **Calcul du score par coureur** : Pour chaque athlète, nous isolons ses performances sur ce Top 20. Le score Micro est la somme de ses points, multipliée par l'indice de similarité de la course et un coefficient de récence (valorisant les saisons 2024-2025).

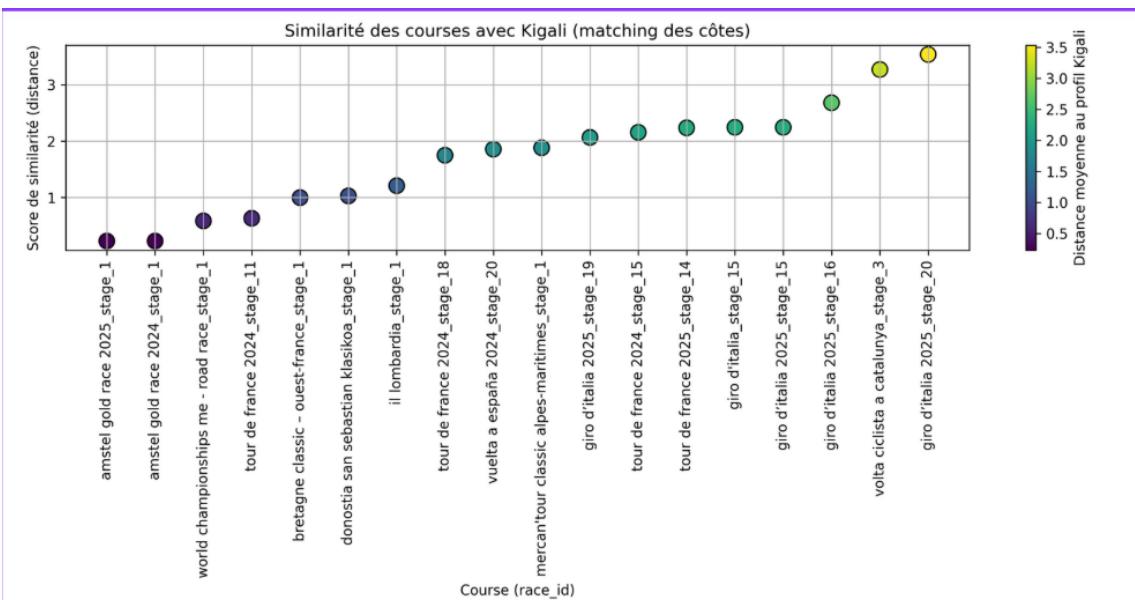


FIGURE 15 – Top 20 des épreuves les plus similaires à Kigali après intégration de l’analyse microscopique des côtes. Ce référentiel, issu de l’algorithme Hongrois, sert de base au calcul du Score Micro en privilégiant les parcours dont la structure des ascensions (pente, longueur, répétitivité) reproduit les contraintes du circuit rwandais.

6.2.2 Résultats : L’émergence des spécialistes de l’explosivité

Le graphique ci-dessous illustre la hiérarchie des coureurs sur cette dimension technique. On observe un basculement significatif par rapport au classement Macro.

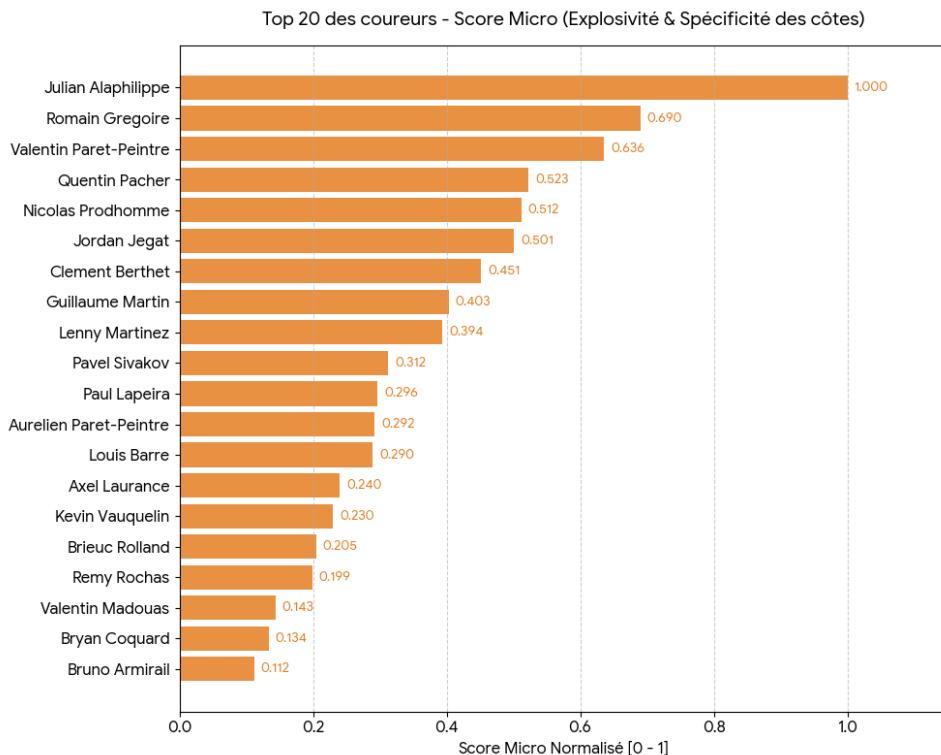


FIGURE 16 – Classement intermédiaire du Top 20 des coureurs sur la dimension Micro (Explosivité et Spécificité topographique).

Analyse des résultats : Le score de **Julian Alaphilippe** (1.000) confirme que ses succès récents (notamment sur les étapes nerveuses ou les classiques) ont eu lieu sur des profils de côtes quasi-identiques à ceux de Kigali. À l'inverse, **Pavel Sivakov** (0.312) est pénalisé par son profil de grimpeur de longs cols réguliers : bien que très endurant (Score Macro élevé), son historique sur les "murs" de type ardennais est statistiquement moins probant. **Romain Grégoire** (0.690) émerge comme l'alternative la plus crédible pour les efforts explosifs répétés.

Cette approche garantit que la sélection finale ne contient que des coureurs ayant déjà prouvé leur capacité à performer sur le terrain spécifique du Rwanda.

Calcul du Score Final et Pondération Hybride

L'aboutissement de notre méthodologie réside dans la fusion des scores Macro et Micro. L'enjeu est de résoudre l'équation complexe du parcours de Kigali : comment sélectionner un coureur capable de supporter une charge de travail de 5475 m de dénivelé (Macro) tout en conservant l'explosivité nécessaire pour faire la différence sur des pentes à 15 % (Micro) ?

6.3.1 La formule de synthèse

Le score final (S_{final}) est conçu comme une moyenne pondérée, ajustée par un facteur de pertinence physiologique :

$$S_{final} = \frac{S_{macro} + S_{micro}}{2} + \text{Bonus}_{cluster} \quad (8)$$

Cette structure mathématique permet de lisser les extrêmes :

- **Évitement des "Athlètes à faible explosivité relative"** : Un coureur avec un excellent S_{macro} (très endurant) mais un faible S_{micro} (peu explosif) verra sa moyenne chuter, car il risque d'être distancé lors des accélérations brutales dans les côtes rwandaises.
- **Évitement des "Puncheurs fragiles"** : À l'inverse, un spécialiste des côtes courtes avec un faible S_{macro} sera pénalisé, son profil n'offrant pas les garanties suffisantes pour atteindre le final après 270 km de course.

6.3.2 Le Bonus de Cluster : L'arbitrage physiologique

Le terme $\text{Bonus}_{cluster}$ introduit une dimension qualitative issue de notre analyse non-supervisée (K-Means). Nous avons appliqué un bonus aux athlètes appartenant aux **Clusters 0, 1 et 4** (identifiés comme la "Zone Kigali").

Ce bonus agit comme un "certificat d'aptitude" : il renforce les coureurs dont la morphologie (poids/taille) et le profil de résultats historique correspondent intrinsèquement aux exigences gravitationnelles du Rwanda. Cela permet, à score de performance égal, de privilégier le coureur dont le profil physiologique est le plus cohérent avec le terrain.

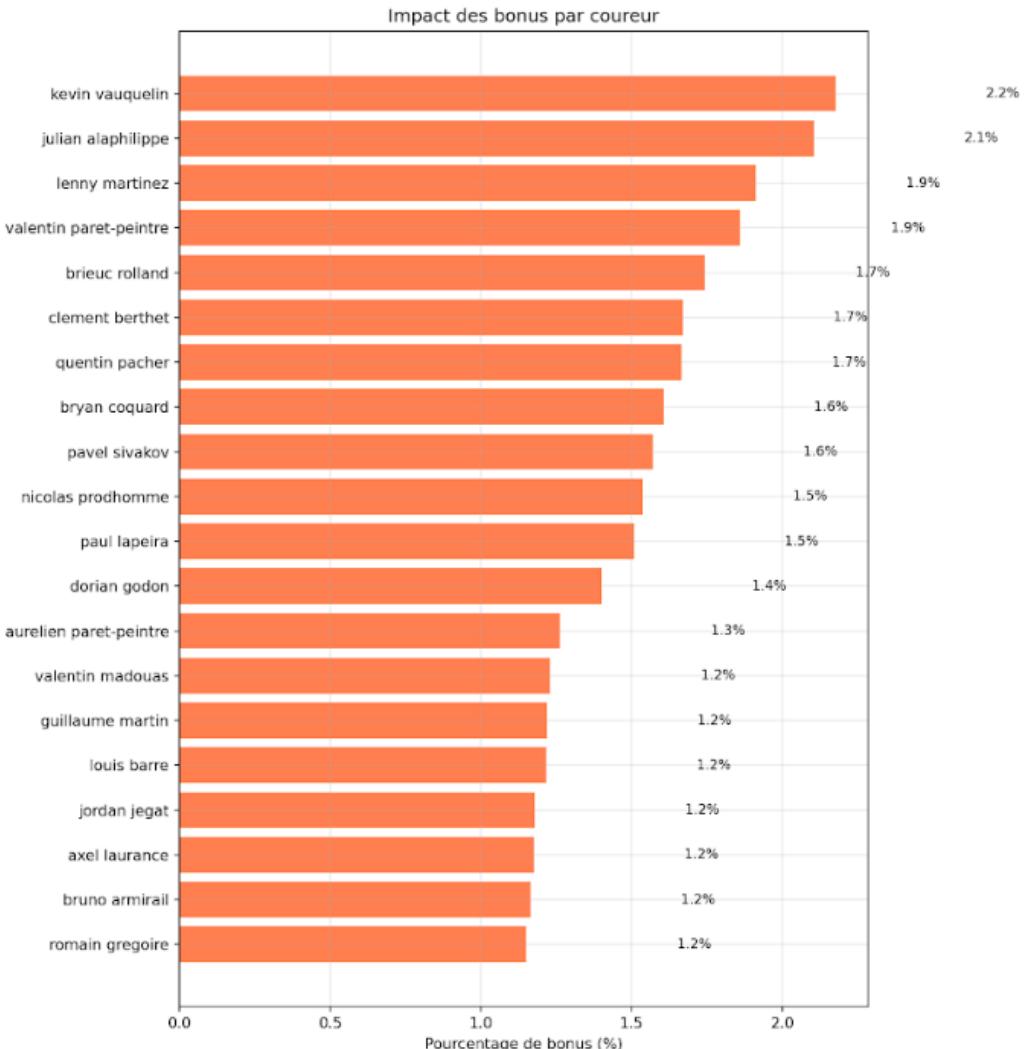


FIGURE 17 – Répartition des composantes du score final : Influence de la performance brute (Macro/Micro) et de l'ajustement cluster (Bonus Physiologique).

6.3.3 Synthèse et arbitrage : Le Score Final

L'aboutissement de notre méthodologie réside dans la fusion des scores Macro et Micro. L'enjeu est de résoudre l'équation complexe du parcours de Kigali : sélectionner des coureurs capables de supporter une charge de travail de 5475 m de dénivelé tout en conservant l'explosivité nécessaire pour faire la différence sur des pentes à 15 %.

Le score final (S_{final}) est calculé comme la moyenne pondérée des deux dimensions, ajustée par le bonus de cluster physiologique. Cette approche permet de lisser les profils extrêmes et de garantir une équipe polyvalente :

- **La résilience** est assurée par le score Macro (dominé par des profils comme Sivakov).
- **L'explosivité** est assuré par le score Micro (dominé par Alaphilippe).

6.3.4 Hiérarchie globale et arbitrage de sélection

Avant d'arrêter la liste définitive, nous avons projeté l'ensemble des coureurs sur une échelle de score consolidée. Ce classement (Figure 18) représente la synthèse finale de notre modèle, intégrant les dimensions Macro, Micro et les bonus physiologiques.

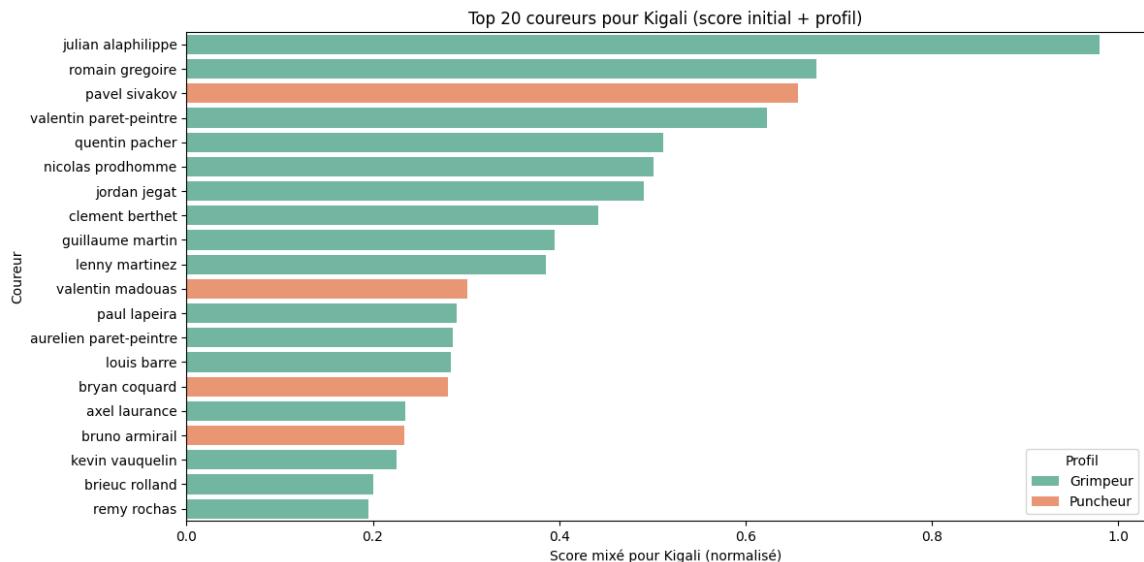


FIGURE 18 – Classement consolidé du Top 20 des coureurs français pour Kigali 2025. Le score final (0 à 1) synthétise la performance historique pondérée par la similarité topographique et l'adéquation du profil.

6.3.5 La sélection officielle (Top 8)

Le règlement de l'UCI pour les championnats du monde imposant une limite de **huit coureurs par nation**, nous appliquons une restriction stricte sur les huit premiers athlètes de notre classement consolidé.

Ce "cut" mathématique permet de constituer une équipe homogène où chaque membre possède un score de compatibilité avec le parcours de Kigali supérieur à 0.44.

TABLE 6 – Sélection officielle de l’Équipe de France (Top 8 Consolidé).

| Rang | Nom du Coureur | Score Macro | Score Micro | Score Final | Profil |
|------|-------------------------------|-------------|-------------|--------------|----------|
| 1 | Julian Alaphilippe | 0.960 | 1.000 | 0.980 | Grimpeur |
| 2 | Romain Grégoire | 0.662 | 0.690 | 0.676 | Grimpeur |
| 3 | Pavel Sivakov | 1.000 | 0.312 | 0.656 | Puncheur |
| 4 | Valentin Paret-Peintre | 0.611 | 0.636 | 0.623 | Grimpeur |
| 5 | Quentin Pacher | 0.502 | 0.523 | 0.512 | Grimpeur |
| 6 | Nicolas Prodhomme | 0.491 | 0.512 | 0.502 | Grimpeur |
| 7 | Jordan Jegat | 0.481 | 0.501 | 0.491 | Grimpeur |
| 8 | Clément Berthet | 0.433 | 0.451 | 0.442 | Grimpeur |

Analyse de la sélection :

Statistiquement, le Top 8 consolidé présente une moyenne de poids de 62,4 kg ($\pm 4,2$ kg), confirmant la sélection rigoureuse d'un groupe optimisé pour les fortes déclivités, conformément aux distributions de performance observées sur les épreuves de catégorie HC (Hors Catégorie).

La tête du classement illustre la synergie recherchée : **Julian Alaphilippe** s'impose comme le leader naturel pour le final explosif, tandis que **Pavel Sivakov**, fort de son score d'endurance parfait, garantit la solidité de l'équipe sur la longue distance. Les six autres coureurs forment un bloc de grimpeurs-puncheurs ultra-cohérent pour soutenir ce duo de tête.

Difficultés rencontrées et gestion des aléas

Le passage de la théorie mathématique à la sélection opérationnelle a révélé plusieurs défis majeurs. Cette section détaille les obstacles rencontrés, qu'ils soient d'ordre technique, algorithmique ou liés à la nature intrinsèque du sport de haut niveau.

Défis de l'acquisition et de l'ingénierie des données

L'absence de base de données unifiée a imposé un travail d'ingénierie considérable :

- **Extraction multi-sources** : Nous avons dû combiner du *web scraping* dynamique (via Selenium pour contourner les protections) et de la reconnaissance optique de caractères (OCR) sur des documents PDF officiels pour constituer la liste des athlètes.
- **Standardisation nominale** : La fusion des fichiers a révélé des incohérences syntaxiques critiques (BOM Windows, accents, traits d'union). Sans une normalisation stricte des chaînes de caractères, les jointures (*merges*) entre les caractéristiques physiques et les résultats sportifs étaient impossibles.

Complexité algorithmique et optimisation

L'implémentation de la dimension Micro a constitué le défi mathématique le plus pointu du projet.

- **Problème d'affectation** : Le circuit de Kigali possède un nombre fixe de côtes (n), tandis que les épreuves historiques en présentent un nombre variable (m).
- **Solution** : L'utilisation de `linear_sum_assignment` (Algorithme Hongrois) a permis d'optimiser mathématiquement l'appariement. Cette approche s'est avérée bien plus robuste que la simple *similarité cosinus* testée initialement, car elle conserve l'intensité physique absolue des pentes.

Arbitrages et instabilité des modèles

Plusieurs approches de classification ont été explorées avant d'arrêter notre méthodologie :

- **Sur-apprentissage (Overfitting)** : Les modèles de *Random Forest* ont été écartés car ils s'adaptaient trop spécifiquement à notre panel restreint de 63 coureurs, perdant ainsi toute capacité de généralisation.
- **Réduction de dimension** : Le K-Means seul était perturbé par le bruit statistique. L'intégration d'une PCA (Analyse en Composantes Principales) a été

nécessaire pour isoler 78,5 % de la variance et obtenir des profils physiologiques stables et cohérents.

Biais statistiques et limites intrinsèques

Malgré la puissance des algorithmes, deux limites fondamentales subsistent :

- **Le Cas des jeunes talents** : Le modèle privilégie l'expérience prouvée. Un jeune talent comme **Paul Seixas** se retrouve mécaniquement pénalisé par son manque d'historique sur les épreuves de plus de 250 km, malgré un potentiel physique évident.
- **La volatilité de la forme et les impondérables** : Les scores reflètent un potentiel historique mais ne peuvent capturer l'état de forme "du jour J", la fraîcheur mentale ou les variables tactiques (rôles de coéquipiers). Le modèle est un instantané statistique, pas une certitude absolue.

En conclusion, la gestion de ces difficultés a permis de transformer un dataset hétérogène en un outil d'aide à la décision robuste, où chaque arbitrage algorithmique est justifié par la réalité du terrain cycliste.

Conclusion

L'objectif de ce projet était de concevoir un système d'aide à la décision capable de naviguer dans la complexité du parcours des Championnats du Monde de Kigali 2025. À travers une approche rigoureuse basée sur la science des données, nous avons pu transformer un historique massif de performances cyclistes en une sélection tactique cohérente.

Apports de la méthodologie

L'originalité de ce modèle réside dans sa structure hybride. En dissociant la résilience métabolique (*Score Macro*) de la précision topographique (*Score Micro*), nous avons évité les biais classiques d'une sélection basée uniquement sur le prestige ou les points UCI.

- L'utilisation de l'**algorithme Hongrois** a permis d'extraire une vérité technique : la capacité d'un coureur à répondre à une pente spécifique.
- Le **clustering K-Means** a apporté une couche de validation physiologique, garantissant que les coureurs sélectionnés possèdent le gabarit et le profil adaptés à la gravité rwandaise.

Limites du modèle : Entre statistiques et réalités du terrain

Bien que robuste, le modèle présente des limites intrinsèques qui imposent une lecture critique des résultats :

- **Le biais d'antériorité (Cas des jeunes talents)** : Le modèle repose sur la profondeur historique des données. Ce biais pénalise les coureurs émergents comme **Paul Seixas**. Malgré un potentiel physiologique hors norme, l'absence de données sur des épreuves professionnelles de plus de 250 km (*Score Macro*) empêche le modèle de les propulser dans le Top 8. La sélection est donc ici "conservatrice", privilégiant l'expérience prouvée à la promesse du talent brut.
- **La volatilité de la forme physique** : Les scores calculés sont le reflet d'un potentiel historique et récent, mais ils ne peuvent capturer l'état de forme "du jour J", les blessures récentes ou la fraîcheur mentale. Un coureur avec un score de 0.800 en méforme sera toujours moins performant qu'un coureur à 0.600 en pic de forme.
- **La dimension impondérable** : Le sport de haut niveau comporte des variables psychologiques, tactiques (rôle de coéquipier de l'ombre) et de santé immédiate que la donnée seule ne peut capturer.

Toutefois, ce système offre une **objectivité précieuse**. Il ne remplace pas le sélectionneur, mais lui fournit une base factuelle pour justifier des choix stratégiques forts,

comme la complémentarité entre **Pavel Sivakov** et **Julian Alaphilippe**.

Perspectives de développement

Le présent modèle constitue une base solide mais pourrait être enrichi par l'intégration de variables environnementales et dynamiques :

- **Modélisation Météorologique** : L'impact de l'humidité et de la chaleur (spécificités climatiques du Rwanda) sur le coût énergétique et la dérive cardiaque des coureurs permettrait d'affiner le Score Macro.
- **Analyse de l'Altitude** : Kigali se situant à plus de 1500 m d'altitude, l'intégration des capacités de performance en hypoxie des coureurs serait un complément majeur.
- **Dynamique de Course** : Coupler ce modèle à une simulation de théorie des jeux pour prévoir les scénarios de course (échappée matinale vs attaque tardive) permettrait d'optimiser non pas seulement les noms des coureurs, mais leurs rôles tactiques précis.

Conclusion

En définitive, ce système d'aide à la décision permet de transformer une masse de données complexe en une stratégie de sélection objective et transparente. En combinant la puissance de l'algorithme Hongrois à une analyse physiologique par clustering, nous offrons à l'Équipe de France une base scientifique pour reconquérir le maillot arc-en-ciel sur les routes exigeantes de Kigali en 2025.