Projet de Fin d'Études

Prédiction de l'abandon des trailers

Etudiants :

Nolwenn PINCZON DU SEL
Mahalia STEPANOFF
Marine HAMONIAUX
Aurélien CRESPIN

Tuteur pédagogique :
Sébastien BAILLY

Responsable pédagogique :
Salima EL KOLEI

2020-2021

Sommaire

Introduction **4**

1 L'Ultra-Trail : un défi pour les coureurs et les analystes **6**
1.1 Présentation de l'Ultra-Trail du Mont-Blanc . . . 6
1.1.1 L'UTMB, une course mythique . . . 6
1.1.2 Quels indicateurs pour analyser la course ? . . . 7
1.2 Décrire l'abandon sur l'UTMB 8
1.2.1 ... par une classification dichotomique 9
1.2.2 ... et vers une typologie plus fine . . . 11

2 Une approche statique par les modèles linéaires généralisés **13**
2.1 Comment modéliser l'abandon a priori ? . . . 13
2.2 Par un modèle logistique sans effet aléatoire... . . . 14
2.3 ...sans pouvoir intégrer la variabilité d'une édition à l'autre . . . 16

3 Une première approche dynamique par les modèles de Markov cachés **17**
3.1 Vers un choix de trois modèles de Markov Cachés par sélection de variables et choix du meilleur d'entre eux . . . 17
3.2 Analyse du modèle retenu : $\mathcal{M}_{\text{ANOVA},5}$ . . . 19

4 Une seconde approche dynamique par les modèles joints **21**

5 Discussion **23**

Conclusion **24**

Bibliographie **26**

Annexe **26**

Liste des tableaux

1.1 Proportion d'abandon par année et par sexe . . . 9
1.2 Comparaison des moyennes des variables explicatives potentielles pour les deux sous-populations (abandon et pas d'abandon) . . . 9
1.3 Comparaison des délais d'avance moyens par rapport aux barrières horaire selon le type d'abandon . 11

2.1 Odds-Ratios et intervalle de confiance à 95% associés (sur l'échantillon d'apprentissage / $n = 7552$) . 14
2.2 Matrice de confusion sur l'échantillon de test pour le modèle logistique complet (69 valeurs manquantes). 15

3.1 Sommes des log-vraisemblances et taux de bonnes prédictions calculés sur l'ensemble des états de toutes les séquences de l'échantillon test et pour les trois modèles de Markov cachés . . . 19
3.2 Taux de bonnes prédictions pour le modèle de Markov caché $\mathcal{M}_{\text{ANOVA},5}$ . . . 19

A-1 Table de contingence des effectifs des coureurs par année et par sexe . . . 27
A-2 Description des variables de la base de données . . . 27
A-2 Description des variables de la base de données . . . 28
A-2 Description des variables de la base de données . . . 29
A-2 Tableau récapitulatif des variables de la base totale . . . 29
A-3 Proportion d'abandon selon leur type, par année et par sexe . . . 29
A-4 Comparaison des moyennes des variables explicatives potentielles pour les trois populations (pas d'abandon, abandon pour dépassement de limite horaire et abandon pour autre cause) . . . 30
A-5 Tableau récapitulatif des p-valeurs de l'analyse de la variance entre chaque variable et la variable du type d'abandon de la course, puis des moyennes correspondantes pour chaque modalité du type d'abandon (celles-ci sont ici calculées directement sur l'ensemble des états de toutes les courses) . . . 30
A-6 p-valeurs des ANOVA pour l'effet de la variable du type d'abandon par balise sur chaque variable quantitative variant en fonction des balises et moyennes correspondantes . . . 31
A-7 Récapitulatif des paires de facteurs (de la variable du type d'abandon par balise) dont les moyennes pour la variable considérée sont significativement différentes au niveau 0.05 pour le test de Student . 32
A-8 p-valeurs des tests de comparaisons de moyennes et de proportions entre les 2 clusters pour chaque variable en fonction des différentes bases de coureurs . . . 34
A-9 Vérification de l'équilibre des échantillons de travail sur l'abandon . . . 35
A-10 Tableau des coefficients estimés et des p-valeurs des cinq variables pré-sélectionnées utilisées comme unique régresseur non-constant dans les cinq régressions logistiques (échantillon d'apprentissage / $n = 7552$) . . . 35
A-11 Tableau des coefficients estimés et des p-valeurs pour le modèle complet (échantillon d'apprentissage / $n = 7552$) . . . 35
A-12 Quartiles des probabilités d'abandon estimées sur l'échantillon de test . . . 36

Table des figures

1.1 Logos des sept parcours de l'Ultra-Trail du Mont-Blanc (Image extraite du site de l'UTMB [1]) . . . 6
1.2 Carte du parcours de l'UTMB (Image extraite du site de l'UTMB [1]) 7
1.3 Dénivelé et nombre d'abandons par balise (moyenné sur les 5 années de course) 10

2.1 Forest plot des odds-ratios pour le modèle complet (sur l'échantillon d'apprentissage / $n = 7552$) . . 15

A-1 Inertie relative du dendrogramme selon le nombre de clusters sur la base 2014 33
A-2 Partition du dendrogramme en 2 classes sur la base 2014 33
A-3 Courbe ROC pour la régression logistique (échantillon de test / $n = 2554$) 36

Glossaire

Assistance tolérée Équipement de balise : zone (partie du poste de ravitaillement réservée à cet usage) pendant laquelle une assistance personnelle est tolérée. Cette assistance est régie par des règles strictes. Elle ne peut être assurée que par une seule personne à la fois, sans matériel spécifique autre qu'un sac d'un volume maximum de 30 litres. L'assistance fournie par un professionnel médical ou paramédical – même si c'est un parent – est strictement interdite. L'assistant est désigné en amont de la course et dispose d'un "Ticket assistance" qu'il doit présenter au chef de poste pour pouvoir exercer sa fonction. 7

Balise Point étape de la course. Dès qu'un coureur atteint une balise, son temps de passage est renseigné. Les balises sont réparties de façon régulière sur le parcours. 5, 6

Barrière horaire Balise qui dispose d'une limite horaire. Les coureurs n'ayant pas atteint la balise concernée avant le délai sont contraints à l'abandon. 5, 6, 8, 11

Cote ITRA Indicateur de niveau et d'expérience d'un athlète dans le domaine du trail. Chaque course terminée (sans abandon) par un trailer lui octroie un certains nombre de points ; et à chaque nouvelle course, sa cote (son total de points) est mise à jour. Le nombre de points accordé varie de 1 à 6 points selon la course, et le barème est fixé par l'International Trail Running Association (ITRA). 6, 8, 10

Indice finisher Indice de performance minimum nécessaire pour pouvoir terminer la course dans le délai imparti. Plus cet indice est élevé, plus la course est signalée comme difficile. 6

Point de contrôle Équipement de balise : zone d'une longueur de quelques centaines de mètres avant et après chaque ravitaillement pendant laquelle le coureur peut être accompagné. 7

Poste de secours fixe (PSF) Équipement de balise : présence d'un poste de secours permanent (et donc de professionnels de santé disponibles) au niveau de la balise. Cela concerne uniquement quelques balises, reconnues comme particulièrement difficiles et nécessitant donc un dispositif préventif pour assister et soigner les éventuels blessés. 7

Temps d'arrêt Temps total passé aux balises disposant de zones de repos. Ce temps d'arrêt est calculé par la différence temporelle entre l'arrivée à la balise (au point de chronométrage noté "in") et la sortie (au point de chronométrage noté "out"). 8, 10

Temps de course Temps total passé en mouvement. On considère ici qu'un coureur est en mouvement à tout instant où il n'est pas sur une balise de repos. 8, 10

Vitesse ascensionnelle Valeur de la projection du vecteur vitesse sur l'axe vertical. Il s'agit en fait d'une vitesse ascensionnelle moyenne calculée entre deux balises consécutives. La formule de calcul est la suivante : $V = \frac{\text{dénivelé entre deux balises}}{\text{temps réalisé pour parcourir ce dénivelé}}$. À titre informatif, en trail, les vitesses ascensionnelles moyennes peuvent varier de 300 à 1 600 m/h (en fonction de la distance, de la pente et du niveau du coureur) et atteindre les 2 000 m/h pour les meilleurs. 8

Introduction

Et s'il était possible d'évaluer votre capacité à compléter un trail avant même votre inscription à la course, souhaiteriez-vous connaître votre probabilité de réussite ? Donner à chacun cette possibilité est tout l'enjeu de ce projet de fin d'études qui se concentre sur la prédiction de l'abandon des participants à l'Ultra-Trail du Mont-Blanc.

Cet évènement sportif créé en 2003 rassemble chaque année 10 000 coureurs de 100 nationalités différentes, 20 000 accompagnateurs et des dizaines de milliers de spectateurs répartis tout le long du parcours. L'UTMB (l'Ultra-Trail du Mont-Blanc) traverse trois pays (France/Italie/Suisse) et trois grandes régions alpines (Auvergne-Rhône-Alpes, la Vallée d'Aoste, le Valais). Il emprunte principalement le sentier de grande randonnée du tour du Mont Blanc à-travers dix-huit communes françaises, italiennes et suisses.

Composé de sept trails de différentes longueurs et aux dénivelés variables, dont quatre en ultra-endurance en pleine nature, cet évènement met à rude épreuve les capacités physiques et psychologiques des trailers et bon nombre d'entre eux abandonnent la course pour différentes raisons. En effet, les conditions de course comme le dénivelé et la présence de **barrières horaires**, les caractéristiques physiques comme le sexe et l'âge, ou encore les conditions météorologiques comme la vitesse du vent et les précipitations, peuvent avoir un impact sur l'abandon lors de la course.

L'objectif de ce projet est d'identifier les facteurs permettant de prédire l'abandon des trailers avant le départ et durant la course. Deux cadres d'étude ont donc pu être considérés : en statique, pour évaluer la probabilité d'abandon avant la course, et en dynamique, afin de prédire en temps réel l'abandon d'un coureur d'une **balise** à l'autre lors d'une course. Pour les participants, cela vise à remplir deux objectifs : d'une part leur permettre de mieux préparer leur course en fonction de leur profil, et d'autre part en les aidant à identifier les points critiques du parcours. Pour les organisateurs, l'étude sur le plan dynamique vise à mieux comprendre le phénomène de l'abandon et à améliorer la prise en charge des coureurs grâce à une répartition optimisée des postes de secours. Ce dernier point est primordial car certaines sections du parcours sont inaccessibles par voies aériennes, et les situations d'urgence, extrêmement délicates à gérer, doivent à tout prix être évitées.

Des études précédentes ont permis d'amorcer la réflexion sur le plan statique mais l'aspect dynamique, central pour la problématique d'organisation des secours, a été peu abordé dans la littérature. C'est d'ailleurs l'ouverture de l'article *A Predictive Framework for Trail Running Assessment of Performance* [2] qui présente les travaux réalisés par Riccardo Fogliato, Natalia L. Oliveira et Ronald Yurko, pour prédire l'abandon des coureurs de l'UTMB avec des modèles statiques de discrimination.

C'est pourquoi ce projet s'inscrit dans la continuité du projet méthodologique, qui portait sur la compréhension théorique et la mise en place pratique des modèles de Markov cachés. Ceux-ci, ainsi que des extentions des modèles de survie, permettent d'intégrer l'aspect dynamique et de donner pour chaque balise, les coureurs qui risquent de ne pas pouvoir atteindre la balise suivante.

Nous présenterons dans un premier temps le contexte de course de l'UTMB afin de comprendre l'articulation des données recueillies, et d'orienter au mieux nos analyses. Nous détaillerons ensuite l'implémentation de nos modèles prédictifs ; d'abord sur le plan statique, puis sur le plan dynamique. Enfin, nous discuterons des limites de ces approches et d'éventuelles pistes d'améliorations pour des travaux futurs.

1 L'Ultra-Trail : un défi pour les coureurs et les analystes

À la différence du trail, qui est une discipline accessible à tous allant de plusieurs kilomètres à quelques dizaines de kilomètres, l'Ultra-Trail est une épreuve de course particulièrement ardue pour tous les athlètes qui s'y confrontent, et dont la distance à parcourir est supérieure à 80 kilomètres. Analyser et modéliser ces données de course - au format très spécifique - constitue également un défi majeur. Nous nous attacherons d'abord à présenter dans cette partie le contexte atypique et les enjeux de notre étude.

1.1 Présentation de l'Ultra-Trail du Mont-Blanc

L'objectif de notre étude est d'analyser les cinétiques de course des trailers afin d'identifier des facteurs potentiels permettant de prédire l'abandon ; ou du moins de discriminer les athlètes les plus à risque. Aussi, pour orienter au mieux notre plan d'analyse, il est primordial d'appréhender correctement le contexte dans lequel s'inscrit notre sujet, et donc de comprendre les spécificités et le fonctionnement de la course.

1.1.1 L'UTMB, une course mythique

L'Ultra-Trail du Mont-Blanc est un évènement sportif réputé dans le monde entier. Chaque année, il rassemble plus de 10 000 coureurs qui viennent tester leurs limites sur l'un des sept parcours proposés dont les noms et les logos sont présentés dans la FIGURE 1.1 ci-dessous.

Note de lecture : *PTL = Petite Trotte à Léon, CCC = Courmayeur-Champex-Chamonix, UTMB = Ultra-Trail du Mont-Blanc, TDS = Sur les Traces des Ducs de Savoie, OCC = Orsières-Champex-Chamonix, MCC = Martigny-Combe-Chamonix, YCC = Youth Chamonix Courmayeur*

FIGURE 1.1 – Logos des sept parcours de l'Ultra-Trail du Mont-Blanc (Image extraite du site de l'UTMB [1])

Notre étude se focalise sur la course principale de l'évènement : l'UTMB. Chaque année, plus de 2 000 coureurs participent à cette course se déroulant en semi-autonomie. Ils réalisent une boucle de 171km autour de Chamonix, en traversant les frontières suisse et italienne, pour un total de plus de 10 000 mètres de dénivelé positif. L'**indice finisher** annoncé est de 420, ce qui signifie que les athlètes qui s'inscrivent à cette course sont invités à avoir une **cote ITRA** supérieure à ce chiffre. La durée maximale autorisée pour franchir la ligne d'arrivée est de 46 heures et 30 minutes. À titre informatif, le temps moyen des athlètes terminant la course est de 39 heures, 49 minutes et 59 secondes.

Afin de suivre le parcours des ultra-trailers dans un contexte rendu difficile par la nature même de l'itinéraire (sentiers escarpés, altitude) et des conditions de course (conditions climatiques et course en continu, y compris la nuit), la course fonctionne sur le principe de **balises**. Réparties de manière régulière tout au long des 171km, ce sont des points de chronométrage fixes qui permettent de mesurer l'avancée des coureurs, d'observer des tendances de cinétique de course (en calculant des indicateurs comme la vitesse moyenne ou la vitesse moyenne ascensionnelle par exemple) et d'enregistrer un classement évolutif. La course de l'UTMB compte 33 balises qui sont présentées dans la FIGURE 1.2.

Certaines de ces balises - appelées **barrières horaires** dans le vocabulaire du trail - présentent une limite temporelle : les coureurs n'ayant pas atteint ces balises dans le délai imparti doivent abandonner la course. Pour l'UTMB,

15 balises sont concernées, parmi lesquelles on retrouve Saint-Gervais, les Contamines et la Balme notamment. Le détail est présenté en FIGURE 1.2.

On note également que plusieurs balises disposent de divers équipements, qui permettent aux coureurs de se ravitailler, se reposer ou se soigner. Dans une course d'ultra-endurance, la possibilité de pouvoir se ressourcer est un élément essentiel. On distingue ainsi trois types de ravitaillement : uniquement boisson, plat chaud ou repas complet. Pour les balises présentant des zone de repos (les Chapieux, Courmayeur, la Fouly, Champex, Trient et Vallorcine), elles disposent d'un point chronométrage à l'arrivée et à la sortie (précisément, on dispose donc de 39 mesures temporelles pour 33 balises différentes). Cela permet de mesurer le temps d'arrêt des coureurs, et ainsi d'obtenir des indicateurs sur le repos qu'ils se sont accordés (ou dont ils ont eu besoin) avant de repartir pour la suite de la course. Enfin, d'autres balises permettent un accompagnement lié aux soins. Cela peut se matérialiser par la présence d'un **poste de secours fixe (PSF)** ou d'un **point contrôle** (avec ou sans **assistance tolérée**). Le détail de la répartition de ces équipements est encore présenté en FIGURE 1.2 ci-dessous.

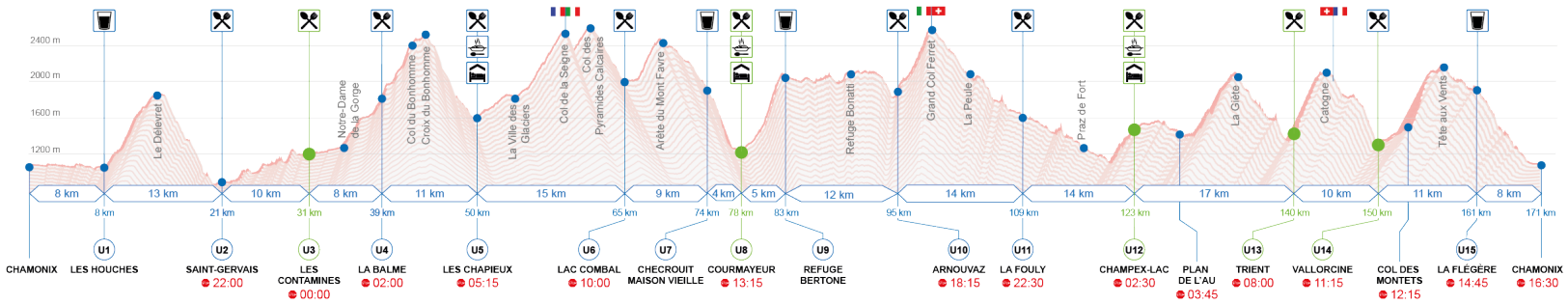

FIGURE 1.2 – Carte du parcours de l'UTMB (Image extraite du site de l'UTMB [1])

Afin d'étudier cette course mythique qu'est l'UTMB, et particulièrement son déroulement, son parcours, ses points de difficulté et ses coureurs, nous disposons des données de cinq années d'édition allant de 2014 à 2018. Nous nous appliquerons dans la partie suivante à détailler leur architecture et à dégager quelques pistes de réflexion pour orienter nos analyses.

1.1.2 Quels indicateurs pour analyser la course ?

En amont de ce projet de fin d'études, un travail de data management réalisé par Enora Alaoui (lors de son stage de première année à l'ENSAI) a permis de constituer la base de données sur laquelle nous avons mené notre étude. Afin de regrouper des données provenant d'années différentes - et donc de parcours quelque peu variables -, mais aussi dans le but de maximiser leur qualité informative, son choix a été de ne conserver que les balises apparaissant sur chacune des cinq années de course. Sur les 39 points de chronométrage évoqués précédemment, seuls 30 ont finalement été conservés.

Notons d'abord que l'architecture de nos données est très particulière. Comme expliqué précédemment, nous disposons uniquement d'informations sur la course aux points fixes que constituent les balises. Pour chaque coureur, on possède donc autant d'observations que de balises atteintes. Dès lors, si un coureur abandonne, on arrête de recenser ses informations après la balise pour laquelle son abandon est indiqué. Il y a donc une inégale répartition des données des coureurs dans le temps.

De plus il est nécessaire de considérer le fait que certains coureurs ont participé à plusieurs courses sur l'intervalle 2014 - 2018. Précisément, 1 682 trailers ont réalisé plus d'une course, pour une moyenne de 2.2 participations. La TABLE A-1 en annexe page 27 présente les effectifs des coureurs par année et par sexe.

Pour analyser les trajectoires individuelles de ces coureurs, et extraire des tendances collectives, nous disposons d'une cinquantaine de variables que nous avons décidé de regrouper en cinq types différents, selon l'information qu'elles offrent :

— sur l'état du coureur (abandon ou non)

— sur la cinétique de course (vitesse, **vitesse ascensionnelle**, **temps de course** ou **temps d'arrêt** notamment)
— sur les caractéristiques personnelles du coureur (son âge, son sexe et sa **cote ITRA** par exemple)
— sur le type de balise (variables indicatrices sur les équipements proposés et dénivelé ou altitude)
— sur les conditions de course (conditions météorologiques, objectives et ressenties)

La TABLE A-2 présentée en annexe page 29 détaille cette répartition.

Parmi ces variables explicatives potentielles, certaines proviennent de notre construction et d'un échange métier avec notre tuteur Sébastien Bailly, dans l'objectif de décrire au mieux les cinétiques de courses des athlètes. Nous avons ainsi choisi de distinguer les **temps de course** des **temps d'arrêt**, afin d'étudier un effet éventuel du temps de repos que s'octroient les trailers sur les balises prévues à cet effet. Ces deux variables sont construites de la même façon : les temps respectifs sont cumulés à chaque point de chronométrage. Dans cette logique, nous avons regroupé les points de chronométrage des balises multiples (c'est-à-dire les balises disposant de zones de repos et donc d'une mesure temporelle à l'arrivée et à la sortie, et les points chronomètre à 1km et 100m de l'arrivée finale). Sur les 30 mesures temporelles de la base de données initiale, nous en avons finalement conservé 22. Nous avons également considéré les mesures de **vitesses ascensionnelles moyennes** entre les balises consécutives, potentiellement plus informatives que les vitesses moyennes. Il a aussi été décidé de quantifier les délais consécutifs des coureurs vis-à-vis des **barrières horaires** réparties tout au long du parcours. Plus précisément, nous avons discrétisé cette information pour obtenir une variable à trois modalités : "0" si la balise ne présente pas de limite temporelle, "1" si le coureur atteint la balise moins de 30 minutes avant le délai éliminatoire et "2" si son avance est plus conséquente. Enfin, pour les trailers concernés par plusieurs participations à la course, nous avons calculé le nombre (évolutif) de participations précédentes. On anticipe que cette variable peut traduire un effet expérience, résolument positif pour terminer la course.

Finalement, nous avons à notre disposition une quantité d'informations importante pour décrire et suivre les trajectoires de course des trailers. Il s'agit à présent de revenir à notre objectif initial, à savoir, expliquer puis prédire l'abandon. Pour cela, il apparaît évident que l'année de course ne doit pas être prise en compte comme variable explicative. En effet, notre but est de prédire les abandons potentiels d'une nouvelle édition de l'UTMB. L'année de course constitue donc un effet aléatoire qu'il faudra considérer dans nos modèles pour estimer des coefficients fixes plus justes. Également, les mesures obtenues a posteriori (comme le temps total de course ou le classement final) ne peuvent être exploitées dans nos analyses. Enfin, on utilisera uniquement l'historique des données jusqu'à la balise $B-1$ pour prédire l'état du coureur (abandon ou non) à la balise B (on reviendrait sinon au problème d'exploitation d'informations obtenues a posteriori).

Maintenant que nous avons établi le cadre de fonctionnement de la course étudiée et posé les bases de notre analyse, nous pouvons nous concentrer plus précisément sur le cœur du sujet : l'abandon des trailers.

1.2 Décrire l'abandon sur l'UTMB ...

Les courses d'ultra-endurance comme l'UTMB poussent les athlètes dans leurs retranchements - à la fois physiques et mentaux - ; à tel point que certains sont contraints à l'abandon, incapables de poursuivre plus longtemps un effort particulièrement difficile. En moyenne, 35% des coureurs par édition sont concernés. Les abandons sont donc des événements habituels en trail, mais la complexité des parcours - qui peut aggraver les situations de détresse - invite à s'intéresser particulièrement à leurs occurrences, afin de protéger les athlètes et de fournir aux secouristes un outil d'aide à la décision. Notre objectif dans un premier temps est alors de bien comprendre les mécanismes qui conduisent à l'abandon et de mettre à jour les trajectoires de course les plus à risque.

Notre intérêt se porte donc ici sur les variables informant sur le statut du coureur, à savoir "en course" ou "abandon". On dispose en fait de deux moyens de renseignement. Une première variable indique si le trailer a abandonné au cours de sa course ou non (c'est donc une mesure répétée à chaque point de temps pour un coureur). La seconde renseigne sur l'étape d'abandon du coureur : l'indicatrice prend la valeur 1 pour l'observation correspondant au référencement de l'abandon (donc la dernière observation de l'individu), et conserve la valeur 0 pour toutes les autres. Ces deux variables sont à considérer dans deux cadres différents : statique pour la première, et dynamique pour la seconde.

1.2.1 ... par une classification dichotomique ...

Pour débuter notre analyse, nous souhaitons visualiser plus précisément les proportions de coureurs concernés par un abandon, selon l'année d'édition et le sexe. Les résultats sont donnés dans la TABLE 1.1 ci-dessous.

Variable	Année					Sexe	
Modalité	2014 (N=2 429)	2015 (N=2 560)	2016 (N=2 554)	2017 (N=2 526)	2018 (N=2 554)	Femme (N=1 186)	Homme (N=11 437)
Proportion d'abandon	0.35	0.36	0.42	0.33	0.30	0.42	0.35

TABLE 1.1 – Proportion d'abandon par année et par sexe

On conclut que les proportions d'abandon sont équilibrées chaque année, ce qui nous permet de travailler à l'échelle macro en regroupant les années. On note néanmoins une légère hausse de la proportion d'abandon en 2016 par rapport aux autres éditions. Cela pourrait s'expliquer par des conditions météorologiques plus médiocres, mais également peut-être par un niveau global plus faible des coureurs. Cette dernière hypothèse n'est pas vérifiée au seuil de 5%, la p-valeur du test du χ^2 entre l'année de course et les cotes ITRA des coureurs valant 0.074. Á l'inverse, on constate une différence significative des conditions météorologiques sur les années d'édition. Á titre informatif, les p-valeurs des tests du χ^2 pour les intempéries et les températures moyennes avec l'année de course sont inférieures à 2.2×10^{-16}. Les différences entre éditions au niveau du nombre d'abandons sont donc dues à des causes indépendantes de l'organisation.

Au niveau du sexe, on remarque que les femmes ont une proportion d'abandon de 42% plus élevée de sept points que celle des hommes (35%). On note que cette différence est significative : la p-valeur du test du χ^2 vaut 2.5×10^{-6}. C'est un constat que l'on approfondira dans la suite de nos analyses.

L'étude de l'abandon des ultra-trailers conduit à chercher des variables permettant de discriminer les coureurs ayant abandonné de ceux ayant terminé la course. Les premières analyses sont effectuées en divisant la population totale selon le statut d'abandon.

Pour une première visualisation des différents profils de coureurs par rapport à l'abandon, nous avons choisi d'effectuer une classification ascendante hiérarchique (CAH) dans l'objectif de faire ressortir des variables discriminantes pour l'abandon. Cette analyse complète est présentée en annexe page 32.

Afin d'étudier les effets potentiels des caractéristiques personnelles des trailers et de leur cinétique de course, nous avons calculé puis comparé les moyennes obtenues pour chacune des populations par des test de Student. La TABLE 1.2 ci-dessous reprend toutes ces informations.

Type de variable	**Variable** Moyenne (écart-type)	**Pas d'abandon (N= 8 146)**	**Abandon (N= 4 477)**	**Total (N= 12 623)**	**P-valeur du test d'égalité de moyenne (pas d'abandon) VS abandon)**
Statique	Âge	42 (8)	45 (9)	43 (8)	$< 2.2 \times 10^{-16}$
	Cote ITRA	559.3 (87.9)	515.0 (107.7)	556.0 (96.1)	$< 2.2 \times 10^{-16}$
Dynamique	Temps de course	39h49m59s (5h18m22s)	18h20m40s (8h45m24s)	32h12m48s (12h17m38s)	$< 2.2 \times 10^{-16}$
	Temps d'arrêt	2h46m35s (1h09m38s)	1h03m27s (0h52m25s)	2h04m19s (1h25m52s)	$< 2.2 \times 10^{-16}$
	Vitesse	4.92km/h (0.74km/h)	4.88km/h (1.07km/h)	4.90km/h (0.87km/h)	3.9×10^{-2}
	Vitesse ascensionnelle positive	299m/s (40m/s)	293m/s (70m/s)	297m/s (53m/s)	1.7×10^{-7}
	Vitesse ascensionnelle négative	528m/s (90m/s)	432m/s (198m/s)	492m/s (148m/s)	$< 2.2 \times 10^{-16}$

TABLE 1.2 – Comparaison des moyennes des variables explicatives potentielles pour les deux sous-populations (abandon et pas d'abandon)

On conclut d'abord que plus un coureur est âgé, plus il a de risque de ne pas terminer la course. Ce constat est logique : les effets de l'âge sur les capacités physiques sont bien connus. Pour autant, il est essentiel de considérer un effet entraînement ou expérience de course, porté par la **cote ITRA**, qui réduit le risque d'abandon lorsque cette cote est plus élevée.

Pour ce qui est des variables à caractère dynamique traduisant la trajectoire de course des individus, les effets apparaissent tous significatifs. Notons que les valeurs présentées pour les vitesses résultent d'une double moyenne : d'abord par individu puis sur la sous-population. L'étude du **temps de course** présentée ici nous permet de porter l'attention sur un potentiel facteur de confusion : deux temps de course égaux peuvent appartenir à deux coureurs aux profils très différents, l'un ayant abandonné la course et l'autre l'ayant terminé relativement rapidement. En observant le **temps d'arrêt cumulé**, on remarque que la durée totale de repos des trailers ayant terminé le parcours est bien plus élevée en moyenne que celle des trailers ayant abandonné. Cela peut traduire une bonne gestion du couple effort/repos, potentiellement de la part de coureurs plus expérimentés. Il faut néanmoins nuancer cette interprétation, puisque les coureurs ayant abandonné n'ont pas atteint toutes les balises de repos et n'ont donc pas eu autant d'occasions de se reposer que les coureurs ayant franchi la ligne d'arrivée. Logiquement enfin, les mesures de vitesse indiquent que les coureurs terminant le parcours courent plus vite en moyenne que ceux poussés à l'abandon. Cela caractérise peut-être un effet fatigue (conduisant à l'abandon) mais est aussi dû aux abandons par dépassement de limite horaire ; les coureurs les plus lents (et même trop lents par rapport aux attendus) étant classés parmi les abandons.

Au-delà des mesures spécifiques à l'individu, on peut s'interroger sur un "effet balise" sur l'abandon. On peut en effet concevoir que certaines balises sont plus difficiles à atteindre que d'autres, notamment d'un point de vue du dénivelé. D'autres effets peuvent être anticipés par rapport aux équipements proposés par les balises. Par exemple, savoir que la prochaine balise est une balise de repos peut pousser un coureur proche de la rupture à poursuivre son effort. Dans cette optique, nous avons décidé d'observer le nombre d'abandons par balise, moyenné sur les années d'édition, et de le superposer au parcours de course. Le résultat est présenté en FIGURE 1.3 ci-dessous. Notons que le nombre moyen d'abandons par balise est un baromètre plus difficilement exploitable qui aurait pu être un indicateur de proportion, mais dans notre situation celui-ci aurait été fortement biaisé puisque les coureurs ayant abandonné à une balise ne sont plus représentés aux balises suivantes.

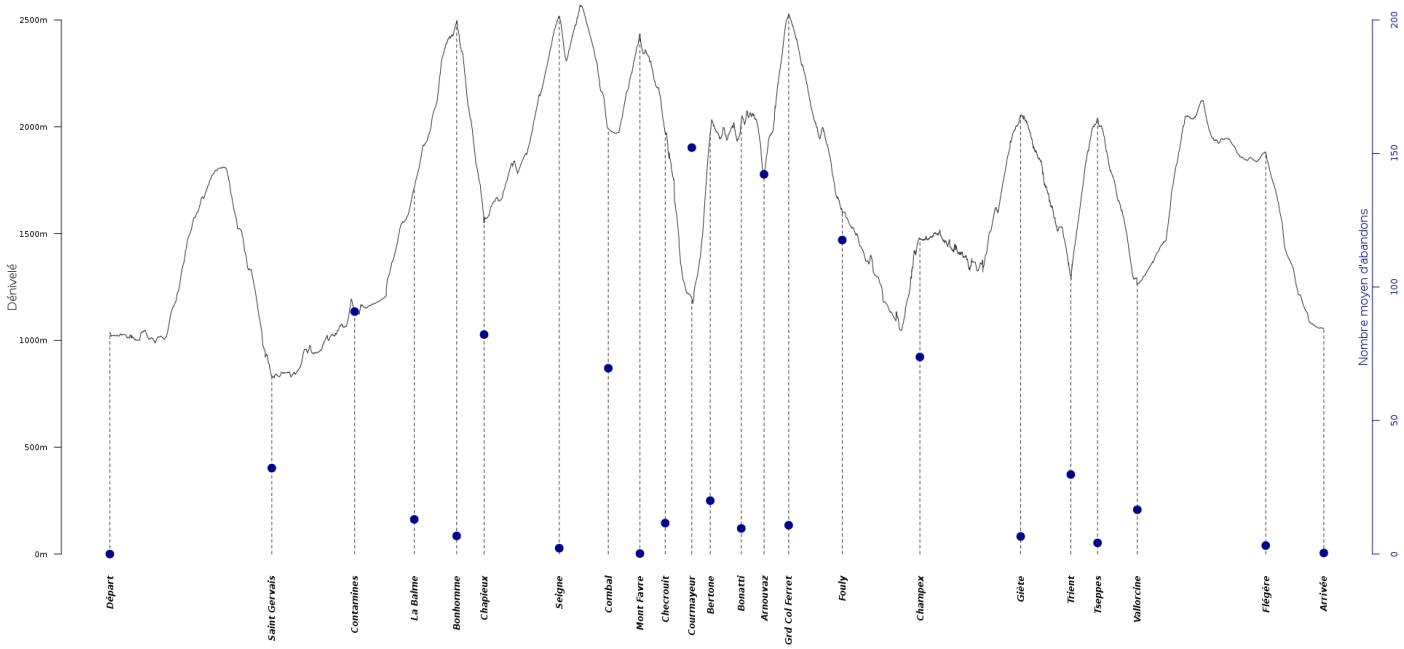

FIGURE 1.3 – Dénivelé et nombre d'abandons par balise (moyenné sur les 5 années de course)

On constate une coupure entre les balises, du point de vue du nombre moyen d'abandons : sept balises, que l'on peut considérer comme particulièrement liées à l'abandon des coureurs, présentent bien plus de 50 abandons par édition quand les vingt-deux autres atteignent à peine une trentaine. En séparant ces balises selon ce critère de 50 abandons, les nombres moyens d'abandon valent 104 et 10 respectivement. Cela renforce bien l'idée d'un effet

balise. On remarque logiquement que la fin de course est peu propice aux abandons : sachant qu'ils sont proches de l'arrivée, les coureurs consentent plus à poursuivre leur effort que s'ils étaient au début ou au milieu de la course, à niveau de fatigue égal. Si l'on se concentre à présent sur les sept balises référencées comme plus difficiles par leur nombre moyen d'abandon, on remarque qu'elles se situent souvent après un dénivelé fortement négatif. Cela peut apparaître contre-intuitif à premier abord, puisque l'on anticipe que les étapes les plus compliquées sont celles ayant un dénivelé positif important. C'est en fait ce que l'on retrouve : en majorité, les sept balises se situent juste avant une étape de fort dénivelé positif. De plus, il n'est pas possible d'abandonner en cours de dénivelé, car certaines zones sont impossible d'accès sauf en hélicoptère. Sachant cela et au vu de leur fatigue, les coureurs peuvent se décourager et décider d'abandonner. Finalement, ce graphique nous permet de faire ressortir le rôle prépondérant du mental sur l'abandon. On note également que l'effet balise remarqué ici devra être approfondi dans nos analyses futures.

Ces premières considérations ont été menées en séparant les individus selon qu'ils aient ou non abandonné lors de leur participation. Pourtant, on peut distinguer deux grands types d'abandon : les abandons par dépassement de **barrière horaire** et les autres abandons (pour raisons médicales ou par décision du coureur). Cette distinction pourrait nous amener à décrire plus précisément les mécanismes d'abandon.

1.2.2 ... et vers une typologie plus fine

Deux variables d'abandon ont été définies : une statique et répétée pour toutes les observations du coureur, et une autre évolutive, indiquant l'étape à laquelle l'abandon s'est produit.

En moyenne, sur les cinq années de course, 2.7% des abandons sont dus à un dépassement de limite horaire. Le détail de cette proportion, par année et par sexe, est présenté dans la TABLE A-3 en annexe page 29. À titre indicatif, la moyenne du dépassement des barrières horaires est de 26 minutes et 24 secondes, pour un écart-type de 21 minutes et 57 secondes. Certains dépassements sont d'ailleurs très faibles, le minimum étant de deux secondes.

Afin de mieux comprendre les trajectoires de course conduisant à un dépassement de limite horaire, nous avons calculé les délais d'avance moyens de chaque coureur par rapport aux barrières horaires et comparé les moyennes et écarts-types de chacune des trois populations (pas d'abandon, abandon pour dépassement et autre abandon). Notons que pour les individus ayant abandonné pour cause de dépassement, la dernière observation n'est pas prise en compte dans le calcul, puisqu'elle fausserait les résultats. Les résultats de cette analyse sont donnés dans la TABLE 1.3 ci-dessous.

Délai d'avance moyen par rapport aux barrières horaire	**Pas d'abandon (N = 8 146)**	**Abandon pour dépassement de limite horaire (N = 118)**	**Abandon pour autre cause (N = 4 359)**
Moyenne	4h14m57s	00h36m18s	2h02m36s
Écart-type	2h41m00s	00h22m28s	1h44m34s

TABLE 1.3 – Comparaison des délais d'avance moyens par rapport aux barrières horaire selon le type d'abandon

On constate que les ultra-trailers ayant dû abandonner pour cause de dépassement de limite horaire ont en moyenne beaucoup moins de marge par rapport aux barrières horaire que les deux autres populations de coureurs. Ce résultat est logique : à force d'être à la limite des barrières horaire, il y a plus de risque de finir par en dépasser une, ne serait-ce que de deux secondes.

Après avoir précisé les modalités d'abandon des coureurs, nous observons à nouveau les différences de moyenne des variables explicatives quantitatives pour chacune des trois sous-populations (sans abandon, abandon pour cause de dépassement de limite horaire et abandon pour autres causes). Les résultats sont présentés dans la TABLE A-4 en annexe page 30. On conclut, comme précédemment, que les caractéristiques des trois sous-populations discriminées par le type d'abandon sont significativement différentes.

Une première analyse des liens

Afin d'avoir un premier aperçu avant la modélisation du pouvoir discriminant de chacune de nos variables explicatives, nous étudions leur variation en fonction du type d'abandon à l'aide d'une analyse de variance (ANOVA). Le but est de déterminer si les moyennes des variables quantitatives varient en fonction de la modalité de la variable qualitative d'intérêt. La limite de cette méthode est de ne pas pouvoir observer le comportement des variables

qualitatives par rapport à la variable d'intérêt, elle aussi qualitative.

Nous réalisons en premier lieu des ANOVA entre chaque variable quantitative statique (qui dépend uniquement du coureur) et la variable qualitative du type d'abandon de la course. Les p-valeurs des tests d'égalité des moyennes et les moyennes correspondantes sont regroupées en annexe dans la TABLE A-5.

Afin de comparer les moyennes des modalités du type d'abandon deux à deux, nous utilisons le test de Student. Pour chaque paire de modalités, la statistique de test est significative à 5%, sauf pour la paire *abandon pour cause de barrière horaire - autre type d'abandon* pour les variables correspondant au nombre de participations antérieures et au classement par sexe.

Nous faisons maintenant de même avec les variables quantitatives qui varient en fonction des balises et la variable du type d'abandon de la balise. Les p-valeurs du test d'égalité des moyennes et les moyennes correspondantes calculées sur l'ensemble des états de toutes les courses, regroupées dans la TABLE A-6, nous serviront plus tard pour ajuster nos modèles de Markov cachés. Pour chacune de ces variables, les paires de facteurs (de la variable se référant au type d'abandon par balise) dont les moyennes sont significativement différentes au niveau 0.05 d'après le test de Student, sont regroupées dans la TABLE A-7 en annexe. On remarque que les variables ayant les p-valeurs les plus petites pour l'ANOVA ont toutes des moyennes significativement différentes pour chaque paire de facteurs correspondant au type d'abandon par balise.

En observant les différences de moyennes des variables en fonction du type d'abandon par balise ou par course dans les TABLEAUX A-5 et A-6, on constate des différences significatives pour certaines variables :

1. pour le type d'abandon par course : classement, classement par catégorie, classement par sexe
2. pour le type d'abandon par balise : différence entre l'heure d'arrivée et l'heure de la barrière horaire, vitesse, dénivelé, pourcentage de dénivelé, vitesse ascensionnelle

Nos premières analyses nous ont permis d'observer et quantifier les incidences des différentes variables à notre disposition sur l'abandon des trailers. Nous souhaitons à présent réaliser une première modélisation, dans un cadre statique, afin de prédire en amont de la course la capacité d'un coureur à compléter celle-ci, selon son profil.

2 Une approche statique par les modèles linéaires généralisés

Deux angles de modélisations statistiques ont pu être considérés au cours de ce projet afin de prédire l'abandon des coureurs : statique et dynamique. Les modèles construits ont pour objectif d'expliquer et de prédire au mieux l'abandon d'un coureur respectivement pour une année de course ou pour une balise donnée lors d'une course.

Construction des échantillons d'apprentissage, de validation et de test

Dans les deux cadres de prédiction, la première étape, tous modèles confondus, a été de définir des échantillons d'apprentissage, de validation et de test.

Étant donné que ce projet vise à prédire l'abandon des participants pour une année de course donnée, les modèles ont été entraînés et validés sur les quatre premières années de course puis testés sur la cinquième année, à savoir 2018. Ainsi, nous avons tiré aléatoirement 75% des observations pour les années 2014 à 2017 afin de construire un échantillon d'entraînement de 7552 coureurs, les 25% restants (2517 coureurs) composant l'échantillon de validation. L'échantillon de test est composé des observations des 2554 coureurs disponibles pour l'année 2018. Pour s'assurer de ne pas biaiser nos conclusions, nous avons effectué en amont de nos analyses une vérification de l'équilibre des trois échantillons sur l'abandon. Cette vérification de la bonne répartition des modalités d'abandon se trouve en annexe (TABLE A-9 page 35).

2.1 Comment modéliser l'abandon a priori ?

Les modèles linéaires généralisés considérés dans un premier temps, sont une approche statique simple qui vise à prédire, lors de l'inscription d'un coureur à la course, sa probabilité d'abandon en fonction de son profil : sexe, âge, taille et cote ITRA. Cette information donnée a priori pourrait orienter l'entraînement et donner un indicateur de réussite. Nous considérons donc un modèle logistique pour expliquer l'abandon d'un coureur pour une course donnée.

Nous avons à notre disposition :

— **Y**, l'abandon au cours d'une course, tel que :
  - $Y = 1$, si le coureur a abandonné lors de la course,
  - $Y = 0$, sinon.

— **X**, le vecteur ligne des régresseurs, contenant :
  - le régresseur constant $X_0 \equiv 1$,
  - $X_j, 1 \leq j \leq p$, les p variables explicatives choisies, à déterminer.

— $n =$ **7 552** réalisations indépendantes du vecteur $(Y, X_1, \ldots, X_p)$.

Le modèle s'écrit donc :

$$logit(p(X)) = ln(\frac{p(X)}{1-p(X)}) = X.\beta$$

où $p(X) = P(Y = 1/X)$ est la probabilité que le coureur ait abandonné lors de la course considérée, sachant les caractéristiques relevées, et $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ est le vecteur des coefficients à estimer.

On peut donc réécrire : $X.\beta = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p$.

Dans le cadre statique, seules les variables propres au coureur comptent. Nous considérons ainsi d'inclure les variables : sexe, cote ITRA, âge et taille.

Plusieurs angles ont été envisagés, en travaillant simultanément sur les quatre années de courses pour lesquelles les échantillons ont été construits :

— considérer les 7 552 participations de la base d'apprentissage comme des coureurs indépendants, ou,

— prendre en compte les 7 552 participations des 6 786 coureurs et en incluant donc un effet aléatoire année

Nous nous concentrons sur le cas de la prédiction en considérant d'abord les coureurs de la base d'apprentissage comme des participants indépendants.

2.2 Par un modèle logistique sans effet aléatoire...

Dans un premier temps, nous considérons un modèle sans effet aléatoire où les 7 552 participations aux différentes courses sont considérées comme indépendantes. Pour sélectionner les variables à inclure dans le modèle, nous avons d'abord testé leur significativité une à une en construisant des modèles où elles étaient le seul régresseur non constant. Après évaluation de ces modèles, nous conservons les quatre variables initiales : sexe, cote ITRA, âge et taille (en annexe FIGURE A-10 page 35).

Construction et résultats du modèle complet

Le modèle logistique complet a donc été envisagé. Le TABLEAU A-11 en annexe page 35 donne les coefficients estimés ainsi que la p-valeur associée à chaque variable. Tous les coefficients sont significatifs au niveau 5%.

Le TABLEAU 2.1 donne les odds-ratios, leurs intervalles de confiance à 95% ainsi que les p-valeurs associées. Notons que :

$$OR(x_1, x_2) = \frac{odds(x_1)}{odds(x_2)} = \frac{\frac{p(x_1)}{1-p(x_1)}}{\frac{p(x_2)}{1-p(x_2)}} = e^{\beta.(x_1 - x_2)}$$

Variable	**Odds-ratios (OR)**	**Intervalle de confiance OR à 95%**	**P-valeur**
Sexe H/F	0.68	$[0.56; 0.0.82]$	3.2×10^{-5}
Âge (écart de 10 ans)	1.45	$[1.35; 1.55]$	$< 2.2 \times 10^{-16}$
Taille S/XS	2.30	$[1.19; 4.76]$	1.7×10^{-2}
Taille M/XS	2.20	$[1.13; 4.56]$	2.5×10^{-2}
Taille L/XS	2.42	$[1.23; 5.04]$	1.3×10^{-2}
Taille XL/XS	2.83	$[1.39; 6.09]$	5.4×10^{-3}
Côte ITRA (écart de 120 pts)	0.85	$[0.79; 0.91]$	1.1×10^{-6}

TABLE 2.1 – Odds-Ratios et intervalle de confiance à 95% associés (sur l'échantillon d'apprentissage / $n = 7552$)

Toutes les variables considérées sont significatives et aucun des intervalles de confiance des odds-ratios ne contient la valeur 1, ce qui assure leur interprétabilité. La note de lecture du *forest plot* de la FIGURE 2.1 précise l'interprétation des odds-ratios.

Toutes choses égales par ailleurs, il est possible de noter que :

— être un homme fait diminuer la probabilité d'abandon sur la course ;

— plus le coureur est âgé, plus sa probabilité d'abandonner la course augmente ;

— plus le coureur a une cote ITRA élevée, plus sa probabilité d'abandonner la course diminue ; et,

— les coureurs de taille S, M, L et XL ont une probabilité d'abandon plus élevée que les coureurs de taille XS. Hormis la taille M, on remarque que plus la taille est élevée, plus la probabilité d'abandon augmente.

Notons également que les intervalles de confiance pour les différentes modalités de taille sont étendus puisqu'on dispose de peu de données associées à la modalité de référence XS.

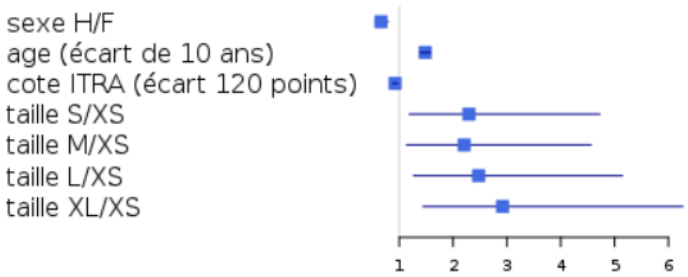

FIGURE 2.1 – Forest plot des odds-ratios pour le modèle complet (sur l'échantillon d'apprentissage / $n = 7552$)

Note de lecture : La droite d'équation $x = 1$ permet de séparer les facteurs associés à l'abandon lors d'une course, à droite, et ceux associés à une course complète (pas d'abandon), à gauche. Le fait d'être un homme ou d'avoir une cote ITRA plus élevée contribue donc à la diminution de la probabilité d'abandonner, alors qu'un âge plus élevé contribue à l'augmentation de celle-ci (plus le coureur est âgé plus sa probabilité d'abandonner augmente).

Cette partie met donc en évidence des éléments expliquant l'abandon. Précisons que ce modèle complet a été retenu suite à une étape de validation, avec une erreur quadratique moyenne de 0.35.

Afin de vérifier la qualité de l'ajustement de notre modèle aux données, nous avons effectué un test d'Hosmer-Lemeshow. La p-valeur obtenue est très faible, inférieure à 2.2×10^{-16} ; ce qui met en évidence une mauvaise spécification du modèle. Ce résultat nous semble logique puisque les seules variables de sexe, d'âge, de taille et de cote ITRA ne peuvent porter toutes les subtilités du mécanisme d'abandon. Cela nous pousse à émettre des réserves sur l'analyse menée dans cette première partie.

Pour autant, cette étude descriptive pourrait permettre de coacher au mieux des coureurs en amont de la course, pour pouvoir donner des indications sur les chances et l'entraînement nécessaire lors de l'inscription au trail, en fonction du profil de chacun. Nous allons maintenant étudier les performances du modèle pour l'aspect prédictif.

Analyses prédictives

Dans le cadre de l'évaluation de la performance prédictive du modèle complet, l'échantillon de test de 2554 coureurs est considéré. L'erreur quadratique moyenne vaut alors 0.30 environ, c'est-à-dire que si l'on considère 100 coureurs, le modèle prédit en moyenne la bonne issue pour 70 d'entre eux. La TABLE 2.2 donne la matrice de confusion, permettant de préciser la nature des erreurs.

		Référence		
		Non abandon	Abandon	Total
Prédiction	Non abandon	1641	660	2301
	Abandon	91	93	184
	Total	1732	753	2485

TABLE 2.2 – Matrice de confusion sur l'échantillon de test pour le modèle logistique complet (69 valeurs manquantes).

Notons que la sensibilité vaut 12% alors que la spécificité s'élève à 95%. La courbe ROC se trouve en annexe page 36, FIGURE A-3. Sur les 30% d'erreurs, on note donc qu'une majeure partie correspond à des coureurs qui ont abandonné et que le modèle n'identifie pas comme tels. Le modèle sur-estime donc la capacité des coureurs de l'échantillon test à compléter la course. Ce constat n'est pas étonnant ; il est difficile pour un modèle de ne pas évaluer à la baisse le nombre d'abandons avec si peu d'informations et des profils proches de coureurs, tous entraînés pour une telle course. Les probabilités d'abandons estimées sont en effet en grande majorité au dessous du seuil discriminant des 0.5, le troisième quartile valant 0.42. À titre informatif, la TABLE A-12 en annexe, page 36, donne les quartiles des probabilités d'abandon estimées. Le maximum de 0.66 montre bien que celles-ci restent plafonnées. Dans le cadre d'une nouvelle édition de l'UTMB, il serait donc pertinent de ré-évaluer ce seuil à la baisse pour améliorer les capacités prédictives du modèle.

2.3 ...sans pouvoir intégrer la variabilité d'une édition à l'autre

Le modèle présenté dans la partie précédente ne donne pas une probabilité d'abandon assez fiable. Celle-ci peut néanmoins servir d'un indicateur *a priori* de réussite, à savoir nuancer, lors de l'inscription à la course. Cette étude a également permis de mettre en exergue les facteurs favorisant l'abandon ou non.

L'étude d'un effet aléatoire année a également été envisagée afin d'intégrer au modèle statique une considération des conditions propres à chaque année de course. Cela permettrait en effet d'apprécier des difficultés liées, par exemple, à la météo globale d'une année à l'autre ; sans pour autant avoir à intégrer les variables dynamiques. Cependant, cette analyse n'a pas permis de déceler un effet année ou individu sur la base d'un modèle plus simple (comprenant uniquement les variables sexe et âge) car le modèle complet n'a pas convergé. Ce constat est d'ailleurs appuyé par les analyses de Riccardo Fogliato, Natalia L. Oliveira et Ronald Yurko [2] qui ont également rencontré un défaut de convergence sur ces mêmes données. Ce résultat ne nous semble pas étonnant au vu des statistiques descriptives sur l'abandon par année (TABLE 1.1) ; il n'a donc pas été jugé nécessaire de continuer les analyses avec une pénalisation pour palier le défaut de convergence.

Dans ce cadre statique, le modèle intégre uniquement les informations *a priori*, récoltées lors de l'inscription à la course. Ce travail sur les caractéristiques des coureurs a permis de répondre à une première attente : l'estimation de la probabilité de réussite, et des analyses permettant d'aiguiller la préparation au trail en fonction du profil de chacun. La performance prédictive de ce modèle logistique est cependant relativement faible. Les coureurs étant tous entraînés pour réaliser une telle course, leurs informations ne sont pas particulièrement discriminantes. Il est nécessaire d'intégrer les variables relatives à la course (cinétique de course, informations des balises et météo) et de tenir compte de l'aspect dynamique pour rendre nos prédictions plus pertinentes et de tenter de répondre à la problématique de l'organisation des secours. C'est tout l'intérêt des modèles de Markov cachés, qui visent à prédire l'abandon d'un coureur à chaque balise. L'introduction du cadre dynamique pourrait effectivement palier notre défaut de performance prédictive.

3 Une première approche dynamique par les modèles de Markov cachés

Dans cette étude, nous considérons trois modalités pour décrire l'abandon ou non de chaque trailer : la première correspond au cas où il a terminé la course sans abandonner, la seconde au cas où il a abandonné à cause d'une barrière horaire non respectée, et la dernière s'il a abandonné pour une autre raison, comme la fatigue physique notamment.

Les types d'abandons observés sur les différentes balises peuvent être vus comme des données temporelles observées pour toutes les unités sur un intervalle de temps donné. Cet intervalle de longueur 22 correspond au nombre de balises après regroupement de certaines d'entre elles et il est tronqué à droite pour les coureurs qui abandonnent. Le type d'abandon à la balise t dépend du type d'abandon à la balise $t-1$. En effet, le coureur ne peut pas abandonner à une balise s'il l'a déjà fait à la balise précédente et il faut prendre en compte cet effet. Le type d'abandon ou le non abandon du coureur à la balise est aussi influencé par d'autres variables variant à chaque balise comme par exemple les variables de cinétique de course (dénivelé, altitude, vitesse du coureur, ...) ou encore les conditions météorologiques (vitesse du vent, précipitations, ...). Nous cherchons ainsi à découvrir une séquence de type d'abandon à chaque balise t pour chacun des coureurs. Pour ces raisons, nous choisissons donc de considérer une modélisation dynamique par modèles de Markov cachés.

Un processus de Markov caché est un processus bivarié $\{(Y_t, Z_t)\}$ composé d'un processus observable $Y = \{Y_t\}$ dont la distribution est déterminée par les états cachés $Z = \{Z_t\}$ modélisés par un processus de Markov. Ainsi, $\mathcal{L}(Z_t|Y_1, Z_1, \ \dots\ , Y_{t-1}, Z_{t-1}) = \mathcal{L}(Z_t|Z_{t-1})$ et $\mathcal{L}(Y_t|Y_1, Z_1, \ \dots\ , Y_{t-1}, Z_{t-1}, Z_t) = \mathcal{L}(Y_t|Z_t)$.

Dans le cadre de notre étude, les états $\{Z_t\}$ dits *cachés* sont les types d'abandons observés à chaque balise t où $t \in [\![1 : T]\!]$ et $T \in [\![1 : 22]\!]$. T, qui est soit la dernière balise du parcours, soit la balise d'abandon du coureur, varie dans notre cas en fonction de chaque course. On note P la matrice de transition du processus markovien Z.

Les états observables Y sont les autres variables qui varient en fonction des balises, comme par exemple l'altitude, ou la vitesse du coureur. Ainsi, Y_t est un vecteur d'observations de variables à l'instant t. Dans la pratique, on sait pertinemment que ce sont les variables Y_t qui influencent le type d'abandon Z_t. Mais afin de pouvoir utiliser un modèle de Markov caché, on considère que dans le cadre théorique, l'abandon Z_t influence le vecteur de variables Y_t. Cette influence est définie par les probabilités d'émission qui sont les probabilités d'obtenir une certaine valeur pour Y_t sachant la valeur de Z_t.

Ainsi, un modèle de Markov caché est défini par le vecteur π des probabilités initiales de chaque état possible de Z, la matrice de transition P de Z et les probabilités d'émission. Une première étape sera d'initialiser le modèle en définissant les variables qui constitueront le processus Y et de définir les paramètres du modèle. Pour cela, nous utilisons le package *pomegranate* de Python. À partir des séquences observées pour chaque coureur et chaque course de l'échantillon d'apprentissage, l'algorithme initialise aléatoirement les distributions des états cachés et la matrice de transition. Les probabilités d'émission sont initialisées grâce à une distribution gaussienne multivariée puis elles seront affinées grâce à un partitionnement en k-moyennes.

Afin de construire au mieux notre modèle final, nous construisons d'abord trois modèles sur l'échantillon d'apprentissage par sélection de variables, puis nous les comparons sur l'échantillon de validation et choisissons le meilleur d'entre eux. Enfin, nous évaluons la performance du modèle retenu sur l'échantillon test.

3.1 Vers un choix de trois modèles de Markov Cachés par sélection de variables et choix du meilleur d'entre eux

Le package *pomegranate* ne permet pas de construire un modèle à la fois à partir de données d'apprentissage et en initialisant les probabilités d'émission avec une distribution multivariée discrète. Nous construisons donc nos

modèles de Markov cachés uniquement avec des variables à valeurs continues et avec une distribution gaussienne multivariée pour l'initialisation des probabilités d'émission. Quant à la matrice de transition, le package ne nous permet malheureusement pas de préciser ses coefficients, en imposant par exemple un état absorbant. Il serait intéressant de trouver un moyen de donner la valeur 0 aux probabilités de passage des états d'abandon à d'autres états d'abandon et la valeur 1 aux probabilités de passage des états d'abandon à l'état de non abandon. En effet, le modèle peut prédire un abandon dans une séquence alors qu'en réalité il n'y en a pas, mais l'idéal dans ce cas serait de rendre les passages d'un état à un autre le plus réaliste possible.

Nous n'intégrons pas la variable de la différence entre l'heure d'arrivée à la balise et l'heure de la barrière horaire car celle-ci détermine complètement la modalité 1 ("abandon pour cause d'arrivée après la barrière horaire") de la variable du type d'abandon. En effet, si cette différence est négative, cela signifie que le coureur est arrivé après la barrière horaire et le type d'abandon prend alors la valeur 1.

Afin de choisir les variables que nous incluons dans nos modèles, nous utilisons les analyses de variance effectuées dans la partie **1.2.2** puis nous construisons les modèles retenus sur l'échantillon d'apprentissage.

Notre premier modèle, noté $\mathcal{M}_{\text{total}}$, prend en compte toutes les variables quantitatives à valeurs continues (sauf le délai avec la barrière horaire) :

barrière horaire	distance par rapport à la balise précédente	vitesse du vent
humidité	temps	température moyenne
température ressentie	altitude	vitesse ascensionnelle
temps arrêté cumulé	vitesse	dénivelé
horaire	pourcentage de dénivelé	

Nous gardons les variables de vitesse et vitesse ascensionnelle car leur V de Cramer vaut 0.2, ce qui fait preuve de peu de corrélation. Malgré le fait que le V de Cramer entre le dénivelé et le pourcentage de dénivelé soit de 0.7, nous conservons aussi ces deux variables car les prédictions sont de meilleure qualité qu'en excluant une des deux.

Pour notre second modèle, noté $\mathcal{M}_{\text{ANOVA},10}$, nous conservons les dix variables avec les plus petites p-valeurs pour le test d'égalité des moyennes. Ces p-valeurs sont disponibles en annexe dans la TABLE **A-6**. Les dix variables sont les suivantes, de gauche à droite et du haut vers le bas, ordonnées de manière croissante en fonction de leur p-valeur :

pourcentage de dénivelé	horaire	dénivelé
vitesse	temps arrêté cumulé	vitesse ascensionnelle
altitude	température ressentie	température moyenne
	temps	

Pour le troisième, noté $\mathcal{M}_{\text{ANOVA},5}$, nous gardons les cinq variables avec les plus petites p-valeurs pour le test d'égalité des moyennes :

pourcentage de dénivelé, horaire, dénivelé, vitesse, temps arrêté cumulé

Les trois modèles peuvent être comparés sur l'échantillon de validation. Pour cela, nous utilisons la somme des log-vraisemblances calculées pour chaque séquence de la forme $Y_1, ..., Y_T$ de l'échantillon de validation. La log-vraisemblance prend en compte l'intégralité de la séquence, ce qui nous permet de cibler les modèles qui prédisent correctement l'abandon ou non d'un coureur à une course. Cependant, notre objectif est de prédire le type d'abandon balise par balise et, pour évaluer ce critère, nous utilisons le taux de bonnes prédictions calculé sur l'ensemble des états de toutes les séquences. En choisissant un modèle qui sera un compromis entre une somme des log-vraisemblances et un taux de bonnes prédictions des états assez élevés, nous nous assurons de prédire au mieux le type d'abandon état par état et pour chaque séquence. Ces résultats sont regroupés dans le tableau ci-dessous :

Modèle	$\mathcal{M}_{\text{total}}$	$\mathcal{M}_{\text{ANOVA,10}}$	$\mathcal{M}_{\text{ANOVA,5}}$
Somme des log-vraisemblances	-3402397	-2843391	-1511887
Pourcentage de bonnes prédictions calculé état par état	98%	73%	76%

TABLE 3.1 – Sommes des log-vraisemblances et taux de bonnes prédictions calculés sur l'ensemble des états de toutes les séquences de l'échantillon test et pour les trois modèles de Markov cachés

Le modèle qui correspond au meilleur compromis entre la maximisation de la somme des log-vraisemblances et la maximisation du taux de bonnes prédictions des états est le modèle $\mathcal{M}_{\text{ANOVA,5}}$ construit à partir des 5 variables avec les plus petites p-valeurs pour le test d'égalité des moyennes. En effet, ce modèle maximise la somme des log-vraisemblances et a le deuxième meilleur pourcentage de bonnes prédictions parmi tous les états. $\mathcal{M}_{\text{ANOVA,10}}$ ne maximise aucun de ces deux critères et $\mathcal{M}_{\text{total}}$ maximise le taux de bonnes prédictions mais minimise la somme des log-vraisemblances. C'est donc le modèle $\mathcal{M}_{\text{ANOVA,5}}$ que nous choisissons.

3.2 Analyse du modèle retenu : $\mathcal{M}_{\text{ANOVA,5}}$

Nous obtenons pour le modèle $\mathcal{M}_{\text{ANOVA,5}}$ construit la matrice de transition suivante :

$$P = \begin{pmatrix} 0.979 & 0.002 & 0.019 \\ 0.333 & 0.333 & 0.333 \\ 0.333 & 0.333 & 0.333 \end{pmatrix}$$

où les premier, deuxième et troisième états correspondent respectivement au non abandon, à l'abandon pour cause d'arrivée à la balise après la barrière horaire et à l'abandon pour une autre cause. Afin que notre modèle soit mieux adapté aux données, il aurait été préférable que les transitions des états d'abandon vers l'état de non abandon valent 1. Ceci est une piste d'amélioration. Il faudrait que l'on obtienne la matrice suivante :

$$\begin{pmatrix} 0.979 & 0.002 & 0.019 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Afin d'évaluer les performances de notre modèle, nous déterminons son taux de bonnes prédictions sur l'échantillon test. Nous utilisons ici différentes façons de le calculer. La première consiste à le calculer état par état et pour toutes les séquences comme nous l'avons fait dans la partie **3.1** dans la TABLE **3.1** pour l'échantillon de validation. Nous faisons de même mais cette fois-ci parmi les coureurs qui ont effectivement abandonné. Pour les troisième et quatrième taux de bonnes prédictions, nous comparons les séquences prédites aux vraies séquences de la base de test plutôt que de comparer tous les états, et ce parmi tous les individus de la base de test puis parmi les coureurs qui ont effectivement abandonné. Ces quatre types de pourcentage de bonnes prédictions sont regroupés dans le tableau ci-dessous :

Données considérées pour le calcul	Pourcentage de bonnes prédictions
Tous les états	77%
Tous les états des coureurs qui ont effectivement abandonné	48%
Les séquences	36%
Les séquences des coureurs qui ont effectivement abandonné	2%

TABLE 3.2 – Taux de bonnes prédictions pour le modèle de Markov caché $\mathcal{M}_{\text{ANOVA,5}}$

Les taux de bonnes prédictions s'affaiblissent fortement lorsque l'on prend en compte l'ensemble de la course. Cependant, la prédiction état par état, qui est celle qui nous intéresse le plus, est relativement bonne. Elle diminue lorsque l'on considère uniquement les coureurs de la base de test qui ont effectivement abandonné, ce qui est prévisible puisque les probabilités de transitions de la matrice P ne sont pas bonnes pour les deuxième et troisième lignes.

Si l'on compare le taux de bonnes prédictions pour tous les états de l'échantillon test, qui est de 77%, à celui de l'échantillon de validation obtenu dans la TABLE **3.1**, qui est de 76%, on constate alors une légère diminution de la performance du modèle entre l'échantillon de validation et l'échantillon test. On peut expliquer cela par le fait que l'échantillon d'apprentissage sur lequel le modèle a été construit contient à peu près autant de balises avec abandons (toutes causes confondues) que l'échantillon de validation, soit 2.1%. L'échantillon test quant à lui est composé à 1.6% de balises avec abandons. Le modèle s'adapte donc mieux aux données de l'échantillon de validation qu'à celles de l'échantillon test car sa structure est celle qui est la plus similaire à celle de l'échantillon d'apprentissage. On précise ici que le pourcentage de balises avec abandons n'est pas le même que le pourcentage de coureurs ayant abandonné, puisque le nombre de balises n'est pas le même pour les coureurs ayant abandonné et en fonction de la balise à laquelle ils ont abandonné.

Il est difficile de prédire l'abandon par un modèle de Markov caché dans lequel nous ne pouvons pas choisir la matrice de transition car cela engendre des transitions entre les états improbables en pratique. En effet, il est impossible d'abandonner à deux balises consécutives et c'est pourtant ce que suppose parfois notre modèle. De plus, il serait intéressant de pouvoir aussi prendre en compte les variables dynamiques qualitatives ainsi que les variables statiques. Une autre limite de ce modèle est l'aspect sans mémoire des chaînes de Markov, qui ne permet pas de prendre en compte l'historique de la course pour prédire l'abandon. C'est pourquoi une analyse de survie par modèles joints pourrait être plus adaptée à nos données.

4 Une seconde approche dynamique par les modèles joints

Les données dont on dispose s'inscrivent dans le cadre d'une analyse de survie : l'abandon d'un coureur peut être apparenté à un évènement marquant la fin de la course, que l'on peut comparer à un décès dans une étude clinique. Notre objectif est donc d'étudier la survie des trailers au cours du temps selon les différentes informations recueillies le jour de la course.

La modélisation statistique de ces données de survie est sujette à de nombreuses contraintes rappelées ici.

D'abord, le suivi longitudinal du parcours des trailers (marqué par la mesure de leurs informations aux temps fixes constitués par les balises) invite à prendre en compte un effet aléatoire individuel.

Ensuite, la position et l'état du coureur à un temps t de la course dépend nécessairement de l'historique de celle-ci. Plus précisément, les mesures obtenues pour chacune des balises sont corrélées à celles des balises précédentes. On peut aussi anticiper un effet potentiel de la balise suivante sur l'état moral du coureur : en effet, s'il sait que la prochaine balise dispose d'une aire de repos ou propose un ravitaillement, le trailer sera peut-être plus motivé à poursuivre sa course - malgré son épuisement physique -, en sachant qu'il pourra bientôt récupérer des forces. On peut donc imaginer la création d'une variable de distance de la prochaine aire de repos.

Enfin, à l'instant t, les données mesurées pour chacun des coureurs sont corrélées entre elles car elles dépendent de la course du coureur lui-même. Les variables de météo et de cinétiques de course sont ainsi particulièrement corrélées : d'une part, les caractéristiques météorologiques influencent la cinétique de course des coureurs ; et d'autre part, la cinétique des coureurs détermine la combinaison "position sur le parcours / conditions météorologiques" qui peut avoir un impact non négligeable sur le risque d'abandon. Prenons ici l'exemple d'une averse particulièrement puissante. Cette météo très défavorable n'aura pas les mêmes répercussions sur les coureurs (et donc sur leur risque d'abandon) selon qu'ils soient dans une côte abrupte rendue glissante ou arrivés à une balise de repos leur permettant de s'abriter.

Pour toutes ces raisons, il apparaît nécessaire de considérer l'ajustement de modèles joints, qui permettraient de modéliser nos contraintes de données longitudinales et dynamiques, tout en tenant compte de la notion de survie inhérente à notre problématique. Nous nous attardons donc dans la partie suivante à présenter le cadre théorique de notre modélisation.

Cadre théorique

Posons dans un premier temps le cadre temporel : le temps de survie, nécessairement continu, est celui du temps de course des coureurs. Pour autant, ce n'est pas lui que l'on cherche à prédire. De toute évidence, il serait absurde d'envisager la prédiction d'une variable si volatile que le temps total de course ; et d'autant plus que l'on sait que l'abandon y crée un effet ambigu (un coureur ayant abandonné durant la course peut avoir le même temps total qu'un coureur très rapide). On s'intéresse plutôt à estimer, à partir de ce temps de course, l'intervalle inter-balise auquel se trouve le coureur. On pourra ainsi calculer une probabilité d'abandon à la balise horizon (la prochaine à atteindre par le coureur). C'est en effet plutôt cette succession dynamique d'horizons (discrets) qui constitue un paramètre d'intérêt pour notre étude.

Soit $D_n = \{T_i, y_i; i = 1, ..., n\}$ un échantillon de la population, avec :

— T_i le temps de course final du coureur i

— δ_i l'indicateur d'évènement ($\delta_i = 0$ si le coureur i n'a pas abandonné et $\delta_i = 1$ sinon)

— y_i le vecteur de réponse longitudinal (de taille $n_i \times 1$) pour le i^e coureur, avec n_i le nombre de balises atteintes par le coureur

Pour tenir compte du caractère longitudinal de la mesure de l'abandon, nous postulons un modèle linéaire généralisé à effets mixtes. On suppose donc que la distribution conditionnelle de Y_i sachant un vecteur d'effets aléatoires a_i est linéaire et appartient à la famille exponentielle. On note :

$$g\left(\mathbb{E}[y_i(t)|a_i]\right) = \eta_i(t) = \beta \times u_i(t) + a_i \times v_i(t)$$

où $g(.)$ désigne une fonction de lien logistique, $y_i(t)$ désigne l'état (abandon ou non) du i^e coureur au point temporel t, et $u_i(t)$ et $v_i(t)$ désignent respectivement les vecteurs des régresseurs dépendants du temps pour les effets fixes β et pour les effets aléatoires a_i. On suppose que les effets aléatoires suivent une distribution normale multivariée : $a_i \sim \mathcal{N}(0, D)$.

Pour l'analyse de survie, on suppose que le risque d'abandon dépend d'une fonction du prédicteur linéaire η_i particulier au coureur i. On note :

$$h_i\left(t|\mathcal{H}_i(t), \omega_i(t)\right) = h_0(t) \times exp\left[\gamma\omega_i(t) + f\{\mathcal{H}_i(t), a_i, \alpha\}\right]$$

où $\mathcal{H}_i(t) = \{\eta_i(s), 0 \leq s < t\}$ indique l'historique de l'état d'abandon jusqu'au temps t et $h_0(.)$ indique la fonction de risque constante. $\omega_i(t)$ est le vecteur de covariables exogènes - éventuellement variables dans le temps - (soit les caractéristiques individuelles des coureurs et la balise à laquelle ils se situent), et on note γ le vecteur des coefficients de régression correspondants. Le vecteur paramètre α quantifie l'association entre les caractéristiques du processus d'abandon jusqu'au temps t et le risque d'un abandon au même moment (dans le cas où celui-ci n'a pas encore eu lieu).

Pour compléter la spécification de notre modèle de survie, nous devons formuler des hypothèses appropriées sur la fonction de risque constante $h_0(.)$. On utilise une approche de type B-splines afin de modéliser $h_0(.)$ en conservant une certaine flexibilité. Le logarithme de la fonction de risque de base s'exprime de la façon suivante :

$$log\left(h_0(t)\right) = \gamma_{h_0,0} \times \sum_{q=1}^{Q} \gamma_{h_0,q} B_q(t, k)$$

où $B_q(t, k)$ désigne la q^e fonction de base d'une spline B avec noeuds $k_1, ..., k_Q$ et γ_{h_0} le vecteur des coefficients de spline.

On estime ensuite les paramètres de notre modèle sous le paradigme bayésien par l'intermédiaire d'un algorithme MCMC (Markov Chain Monte Carlo).

Difficultés d'implémentation

Nous n'avons malheureusement pas réussi à implémenter cette modélisation sur nos données. Après avoir imputé nos valeurs manquantes sur les temps et vitesses de course, nous obtenons une erreur de valeurs manquantes que nous n'expliquons pas. Nous estimons cependant que poursuivre les investigations - au niveau des fonctions implémentées dans les packages - pourrait permettre de résoudre ce problème d'applicabilité. En effet, la solution que nous anticipons serait de re-travailler les packages pour les adapter au mieux à nos données.

5 Discussion

Au vu du caractère très particulier de nos données, une grande partie de notre travail s'est davantage inscrite dans une réflexion méthodologique sur l'adaptabilité des différents modèles à mettre en place que dans l'application concrète de ceux-ci. Nous présentons dans cette partie l'état de nos observations et questionnements, dans l'objectif d'orienter au mieux les travaux futurs.

Reprenons d'abord l'architecture des données : doublement longitudinales et définissant une bulle temporelle, elles ne respectent aucune des hypothèses standards des modèles habituellement appliqués. Les résultats que nous avons présentés dans les parties précédentes sont donc à nuancer fortement. En effet, nos données peuvent s'apparenter aussi bien à :

- de la survie, puisque l'abandon peut être considéré comme l'évènement de "mort" du coureur dans la bulle spatio-temporelle de l'UTMB
- des séries temporelles avec un effet individuel, puisque le passé de la course a - pour chaque individu - une incidence non négligeable sur la suite du parcours
- des états de Markov cachés, en considérant que l'état mental et physique du coureur - non observé - est modélisable à partir des données mesurées et prend des valeurs dans un ensemble discret (par exemple, au début de la course, le trailer est dans un état "Très bon", puis il entre dans différents degrés de fatigue, jusqu'à un potentiel point de non-retour qui déclenchera l'abandon)

La difficulté principale que nous avons relevée est en fait la dualité des temps d'intérêt de nos données : on a d'un côté, le temps de course des coureurs, continu, et de l'autre, les temps fixes des balises, discrets. En effet, pour anticiper les abandons des athlètes en zones critiques, il est nécessaire de prendre en compte leur situation géographique par rapport aux points clés du parcours (*i.e.* les balises) ; mais également leur cinétique de course, qui définit le couple "position sur le parcours / conditions météorologiques". Le suivi des courses est ainsi d'autant plus difficile que pour un temps donné, la position des trailers sur le parcours suit une loi non définie et dépendante du temps considéré. Cet effet a été mis en évidence dans la publication *A Predictive Framework for Trail Running Assessment of Performance* de Riccardo Fogliato, Natalia L. Oliveira et Ronald Yurko [2], grâce à une estimation des densités de distribution des temps de passage des coureurs aux différents points de contrôle (*i.e.* les balises). On peut également souligner l'incidence inhomogène du passé sur la course, puisque dépendante du tracé du parcours ainsi que des équipements proposés aux différents points fixes.

Pour toutes ces raisons, il semble très délicat d'expliquer et d'anticiper les abandons ; d'autant plus qu'ils proviennent directement des ressources en force mentale des individus, non quantifiables et non observables. C'est en fait la complexité physique et psychique du corps humain que l'on cherche à objectiver ici.

Revenons à présent sur les modèles que nous avons testés pour analyser ces données, en mettant en avant leurs facteurs limitants.

D'abord, les modèles de Markov cachés ont permis de bien prédire l'absence d'abandon à chaque balise pour les coureurs ayant complété la course et ceux ayant abandonné par la suite. Cependant, ils ne prennent pas en compte l'impossibilité de certaines transitions entre les différents états d'abandon, comme souligné en PARTIE 3. D'autre part, par définition des chaînes de Markov, le processus est jugé sans mémoire : on considère que l'information utile pour la prédiction du futur est entièrement contenue dans l'état présent du processus et est indépendante des états antérieurs (propriété de Markov). Enfin, dans la pratique, l'état physique et mental du coureur n'a pas été modélisé comme processus caché ; ce qui constituait pourtant la justification de l'implémentation de cette méthode sur nos données.

Nous avons également étudié l'application des modèles de Landmark avec l'intention de pouvoir visualiser, aux temps fixes des balises et par rapport à leurs caractéristiques (équipements, dénivelé pour les atteindre...), l'évolution des probabilités de survie estimées pour chaque individu. Mais la nature discrète de ces temps a empêché la réalisation de cet objectif, c'est pourquoi nous n'avons pas mentionné cette méthode dans ce rapport. Les exemples présentés par Hans C. van Houwelinger et Hein Putter dans le livre *Dynamic Prediction in Clinical Survival Analysis* [3] couplé au package *dynsurv* de R le démontrent.

Suite à ce constat, nous avons essayé d'ajuster des modèles joints en prenant en compte les temps de course continus comme temps d'analyse et les numéros de balises comme variable explicative pour ensuite assembler ces

deux informations temporelles. Cette approche a été rendue possible par l'implémentation des modèles joints proposée dans le package *JMbayes* conçu par Dimitris Rizopoulos [4]. Pour autant, cette approche nous limite sur le nombre de variables explicatives dynamiques, ce qui est très coûteux pour notre analyse. On revient ici sur l'idée que toute tentative d'objectivation d'un mécanisme aussi complexe que l'abandon en trail doit nécessairement intégrer un maximum d'informations.

Dans la continuité de cette réflexion, nous avons pensé à une solution consistant à créer un indicateur ou un score de l'état mental et physique des coureurs. Cet indicateur permettrait d'intégrer la multiplicité des interactions de notre problème, tout en allégeant les modèles. Une étude spécifiquement dédiée à la construction de ce baromètre serait pertinente. On pourrait par exemple discuter d'autres variables d'intérêt potentielles comme un différentiel de vitesse rapporté au dénivelé comparé à la tendance générale par exemple, ou encore un indice lié à la régularité de la vitesse. Sur ce dernier point, on peut en effet anticiper qu'un athlète a plus de chance de terminer la course s'il fournit un effort constant plutôt que s'il part trop rapidement, provoquant un effondrement de ses capacités physiques ensuite.

Cette prise de recul sur nos modélisations statistiques nous a conduit à considérer l'application de méthodes de machine learning - comme les réseaux de neurones - sur nos données. Cela permettrait de gérer le caractère très atypique des données étudiées, sans avoir à poser d'hypothèses limitant la phase de modélisation. On anticipe que les réseaux de neurones LSTM (Long-Short Term Memory) pourraient être plutôt adaptés à la situation. Le modèle pourrait ajuster des poids distincts pour les différents temps du passé, intégrant ainsi la dimension temporelle des données ; mais aussi éventuellement pour différents types de cinétiques de course, afin d'inclure l'hétérogénéité des effets du passé selon le temps de course (notamment à cause des conditions météorologiques changeantes et de la dualité jour/nuit). Par ailleurs, des modèles de prédiction de séries temporelles, ou encore des approches par classifieurs de trajectoires multivariées pourraient être considérées, grâce aux packages *forecast* et *prophet* respectivement.

Nous souhaitons à présent, pour achever cette discussion, émettre l'idée de la création d'un package répondant parfaitement à la problématique posée. La perspective proposée serait de permettre dans un premier temps l'ajustement de modèles de survie sur chacun des intervalles inter-balises, en prenant en compte la difficulté du parcours sur cet intervalle (notamment par rapport au dénivelé et à la proximité ou non - passée ou future - d'équipements permettant de se ressourcer) ; mais aussi les conditions météorologiques et les caractéristiques des coureurs. On pourrait ensuite agréger les estimations obtenues à l'aide de splines pour obtenir un suivi dynamique de la course et tenir compte de la dépendance temporelle des intervalles inter-balises. Un tel dispositif pourrait constituer un outil d'aide à la décision conséquent pour les secouristes dont les interventions d'urgence sont rendues difficiles, voire même impossibles parfois, sur les sentiers escarpés de l'UTMB.

Notons pour conclure que les idées présentées ici ne sont pas exhaustives et que la problématique sur laquelle nous avons travaillé pourrait ouvrir des opportunités de recherche relativement larges pour modéliser ces données particulièrement composites.

Conclusion

L'Ultra-Trail Mont-Blanc est l'une des courses les plus difficiles au monde, en raison de son dénivelé de 10 000 mètres mais aussi à cause des conditions météorologiques de haute montagne, comme le froid, le vent ou la neige. Des milliers de trailers aux profils très différents s'élancent chaque année pour tenter de venir à bout des 171 kilomètres à parcourir. Pourtant, près d'un tiers d'entre eux abandonnent avant de passer la ligne d'arrivée, que ce soit pour dépassement des délais impartis ou par épuisement physique et/ou psychologique.

Face à ce contexte, l'organisation des secours est une préoccupation de premier ordre, qui est d'autant plus mise en difficulté par les conditions extrêmes. Pour protéger les athlètes dont l'intégrité physique et psychologique peut être mise en jeu, il est donc primordial de parvenir à anticiper les situations d'urgence afin de les prévenir.

Malheureusement, sur l'aspect prédictif, les modèles implémentés ont été très peu performants. Du côté statique, les mauvaises capacités prédictives de la régression logistique, légèrement meilleures que le hasard, étaient attendues. Le seul profil des coureurs ne permet pas de regrouper la multiplicité des facteurs influant sur la faculté à compléter le trail. Du côté dynamique, les hypothèses a priori des modèles envisagés sont un frein à l'applicabilité des estimations comme discuté précédemment.

L'architecture et la qualité des données ont rendu la tâche particulièrement complexe. Face aux données manquantes et à la présentation des données, il a été nécessaire de prendre certaines décisions qui restent encore discutables. Ces choix sont en partie à l'origine de limites dans l'application des modèles. Une réflexion supplémentaire sur la création d'indicateurs de performance de course pourrait condenser plus d'information que prise en compte jusqu'à maintenant. Considérer un indicateur de rythme relatif, comme proposé par les auteurs de l'article *A Predictive Framework for Trail Running Assessment of Performance* [2], a en effet plus de sens qu'étudier une vitesse absolue.

Nous nous sommes ici concentrés sur deux approches non exhaustives, d'évolution d'un état discret d'une part, et de survie d'autre part. Or la complexité des données pourrait être prise en charge par des méthodes de machine learning, beaucoup plus flexibles vis-à-vis des hypothèses de départ. Il serait ainsi intéressant de mettre en place, entre autres, des réseaux de neurones de type *Long Short-Term Memory*.

Finalement, les difficultés soulevées par nos analyses nous conduisent à affiner la réflexion menée. Un travail de recherche plus approfondi nous semble nécessaire pour construire un modèle mieux adapté à la problématique. On s'abstiendrait ainsi des limites de l'implémentation de méthodes statistiques en partie inadéquates.

Bibliographie

[1] Site répertoriant toutes les informations de l'UTMB. [Online]. Available : `https://utmbmontblanc.com/`.

[2] Riccardo Fogliato, Natalia L. Oliveira, and Ronald Yurko. A Predictive Framework for Trail Running Assessment of Performance. July 2020. [Online]. Available : `https://arxiv.org/pdf/2002.01328.pdf`.

[3] Hans C. van Houwelinger and Hein Putter. *Dynamic* P*rediction in* C*linical* S*urvival* A*nalysis*. 2012.

[4] Dimitris Rizopoulos. The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC, year=2016, note= [Online]. Available : `https://arxiv.org/pdf/1404.7625.pdf`.

Annexe

[1] Statistiques descriptives

1.1 Description des données

À titre informatif, nous avons réalisé la table de contingence des effectifs des coureurs par année et par sexe. On constate une stabilité des effectifs chaque année, et on peut noter une tendance à la hausse de la participation féminine.

	2014	**2015**	**2016**	**2017**	**2018**	**Total**
Femme	199	222	257	241	267	1186
Homme	2230	2338	2297	2285	2287	11437
Total	2429	2560	2554	2526	2554	12623

TABLE A-1 – Table de contingence des effectifs des coureurs par année et par sexe

Notre base de données est constituée d'une cinquantaine de variables, que nous avons regroupées en six grandes catégories détaillées ci-dessous.

TABLE A-2 – Description des variables de la base de données

Catégorie des variables	**Nom des variables**	**Description des variables**
Informations sur l'abandon	Abandon	Variable indicatrice indiquant si le coureur a abandonné au cours de la course.
	Abandon à l'étape	Variable indicatrice indiquant l'étape à laquelle le coureur a abandonné la course.
	Type d'abandon	Variable catégorielle à 3 modalités indiquant si le coureur a abandonné et pourquoi. La modalité "0" correspond à l'absence d'abandon, la modalité "1" à un abandon pour cause de dépassement de limite horaire et la modalité "2" à un abandon pour une autre cause.
	Type d'abandon à l'étape	Variable catégorielle à 3 modalités indiquant l'étape à laquelle le coureur a abandonné et pourquoi. La modalité "0" correspond à l'absence d'abandon, la modalité "1" à un abandon pour cause de dépassement de limite horaire et la modalité "2" à un abandon pour une autre cause.
Cinétique de course	Date	Date d'arrivée à la balise (au format POSIXct ; soit aaaa-mm-jj h :m :s).
	Horaire	Horaire d'arrivée du coureur à la balise (au format numérique).
	Temps de course	Temps de course cumulé (au format hms).
	Temps d'arrêt	Temps d'arrêt cumulé (au format hms) aux balises permettant aux coureurs de se reposer.
	Vitesse	Vitesse moyenne du coureur entre la balise concernée et la balise précédente.
	Vitesse ascensionnelle	Vitesse ascensionnelle moyenne du coureur entre la balise concernée et la balise précédente (calculée par rapport au dénivelé de la balise).

Table A-2 – Description des variables de la base de données

Catégorie des variables	Nom des variables	Description des variables
	Délai avec la barrière horaire	Délai de passage (au format hms) avant une limite horaire (s'il y en a une, sinon on note une valeur manquante). Si le coureur passe en avance à la barrière horaire, son délai est négatif et s'il passe en retard son délai est positif.
	Trente minutes avant la barrière horaire	Variable catégorielle à 3 modalités indiquant la tranquillité du coureur par rapport aux barrières horaires. La modalité "0" correspond à une absence de barrière horaire, la modalité "1" à un passage moins de 30 minutes avant la limite ou à un passage après la limite et la modalité "2" à un passage plus large, supérieur à 30 minutes.
Informations personnelles sur le coureur	Identifiant	Identifiant du coureur, unique pour chaque individu.
	Identifiant année	Identifiant du coureur, unique pour chacune des courses d'un individu.
	Âge	Âge du coureur à la date de la course.
	Code postal	Code postal du coureur.
	Ville	Ville de résidence du coureur.
	Pays	Nationalité du coureur.
	Année de naissance	Année de naissance du coureur.
	Sexe	Sexe du coureur.
	Taille	Taille du coureur. Modalités : "XS", "S", "M", "L", "XL".
	Cote ITRA	Cote ITRA du coureur (indicateur de son expérience/ses performances en trail).
	Nombre de participations précédentes	Nombre de participations précédentes à l'UTMB.
Informations sur les balises	Balise horaire	Numéro et nom de la balise.
	Numéro de balise	Numéro de balise (variable utilisée pour les affichages, pour une meilleure visibilité).
	Altitude	Altitude de la balise.
	Barrière horaire	Si la balise concernée possède un temps limite, alors la variable présente l'horaire limite (au format hms) ; sinon est convenue une valeur manquante (NA).
	Ravitaillement boisson	Variable indicatrice indiquant si la balise présente un ravitaillement boisson.
	Ravitaillement plat chaud	Variable indicatrice indiquant si la balise présente un ravitaillement plat chaud.
	Ravitaillement complet	Variable indicatrice indiquant si la balise présente un ravitaillement complet.
	Repos	Variable indicatrice indiquant si la balise présente un lieu de repos.
	Point de contrôle	Variable indicatrice indiquant si la balise dispose d'un point de contrôle (i.e. le coureur peut être accompagné).
	Assistance tolérée	Variable indicatrice indiquant si la balise dispose d'un point de contrôle avec assistance tolérée (i.e. le coureur peut être assisté par un proche).
	Poste de secours fixe	Variable indicatrice indiquant si la balise dipose d'un poste de secours fixe (i.e. des professionnels de santé sont présents).

Table A-2 – Description des variables de la base de données

Catégorie des variables	Nom des variables	Description des variables
	Dénivelé	Dénivelé (en m) par rapport à la balise précédente.
	Distance	Distance (en km) par rapport à la balise précédente.
	Pourcentage de dénivelé	Pourcentage de dénivelé sur l'étape (dénivelé/distance).
Informations sur les conditions de course	Humidité	Humidité enregistrée au niveau de la balise.
	Température moyenne	Température moyenne à la balise.
	Température ressentie	Température ressentie à la balise.
	Intempéries	Type d'intempéries à la balise (15 modalités).
	Vitesse du vent	Vitesse du vent à la balise.
	Heure de repas	Variable indicatrice indiquant si le passage à la balise s'est effectué sur une heure de repas. Les heures de repas sont définies arbitrairement comme : 8h-9h, 12h-13h, 19h-20h.
	Jour-nuit	Variable indiquant si le passage à la balise s'est effectué le jour ou la nuit.
Informations supplémentaires sur la course	Année de course	Année de la course.
	Temps total	Temps total de course (s'il y a eu abandon, le temps est arrêté au moment de l'abandon).
	Classement	Classement final individuel sur la course.
	Classement par catégorie	Classement final par catégories. Les catégories construites pour cette variable ne sont pas connues.
	Classement par sexe	Classement final individuel sur la course selon le sexe.

Table A-2 – Tableau récapitulatif des variables de la base totale

1.2 Étude des liens entre variables

Nous présentons ici le détail de proportion des modalités d'abandon selon leur type, par année et par sexe. On remarque que les abandons par dépassement de limite horaire sont très peu représentés.

Variable	Année					Sexe	
Modalité	2014 (N=2 429)	2015 (N=2 560)	2016 (N=2 554)	2017 (N=2 526)	2018 (N=2 554)	Femme (N=1 186)	Homme (N=11 437)
Proportion de non abandon	0.652	0.637	0.575	0.668	0.696	0.582	0.652
Proportion d'abandon par dépassement de limite horaire	0	0.003	0.003	0.039	0.002	0.021	0.008
Proportion d'abandon pour autre cause	0.348	0.360	0.422	0.293	0.302	0.396	0.340

Table A-3 – Proportion d'abandon selon leur type, par année et par sexe

Il s'agit maintenant de détailler les moyennes et écarts-types pour les variables quantitatives en fonction du type d'abandon.

Type de variable	**Variable** moyenne (écart-type)	**Pas d'abandon**	**Abandon pour dépassement de limite horaire**	**Abandon pour autre cause**	**Total**
Statique	Âge	42 (8)	50 (10)	45 (9)	43 (8)
	Cote ITRA	559.3 (87.9)	454.3 (51.7)	538.3 (108.0)	551.0 (96.1)
Dynamique	Temps de course	39h49m59s (5h18m22s)	14h11m20s (9h28m29s)	18h27m26s (8h42m36s)	32h12m48s (12h17m38s)
	Temps d'arrêt	2h46m35s (1h09m38s)	0h31m24s (0h25m11s)	1h04m02s (0h52m36s)	2h04m19s (1h25m52s)
	Vitesse	4.92km/h (0.74km/h)	4.54km/h (0.56km/h)	4.89km/h (1.07km/h)	4.90km/h (0.87km/h)
	Vitesse ascensionnelle positive	299m/s (40m/s)	239m/s (50m/s)	294m/s (69m/s)	297m/s (53m/s)
	Vitesse ascensionnelle négative	528m/s (90m/s)	274m/s (211m/s)	436m/s (196m/s)	492m/s (148m/s)
	Vitesse du vent	17.4km/h (6.9km/h)	16.8km/h (3.4km/h)	17.2km/h (5.8km/h)	17.3km/h (6.5km/h)
	Humidité	0.70 (0.17)	0.83 (0.05)	0.71 (0.15)	0.71 (0.16)
	Température moyenne	11.3° (4.7°)	6.6° (3.7°)	13.1° (5.4°)	11.9° (5.0°)
	Température ressentie	6.7° (4.1°)	3.5° (3.6°)	8.5° (5.0°)	7.2° (4.6°)

TABLE A-4 – Comparaison des moyennes des variables explicatives potentielles pour les trois populations (pas d'abandon, abandon pour dépassement de limite horaire et abandon pour autre cause)

La TABLE A-4, construite pour étudier la significativité des différences entre les sous-populations discriminées par le type d'abandon, a été complétée par des tests d'analyse de variance, tous significatifs au niveau 5 %. Cependant, des tests similaires ont été réalisés pour la préparation du modèle de Markov caché, en posant des hypothèses propres au modèle. Les résultats sont contenus dans les tables suivantes.

Variable	**p-valeur de l'ANOVA**	**Moyenne des coureurs n'ayant pas abandonné**	**Moyenne des coureurs ayant abandonné à cause de la barrière horaire**	**Moyenne des coureurs ayant abandonné pour une autre raison**
Nombre participations lors des années précédentes	<0.05	0.17	0.33	0.22
Cote ITRA	<0.05	559	454	538
Âge	<0.05	42	50	45
Classement	<0.05	819	2316	2066
Classement par catégorie	<0.05	252	426	565
Classement par sexe	<0.05	691	1693	1697

TABLE A-5 – Tableau récapitulatif des p-valeurs de l'analyse de la variance entre chaque variable et la variable du type d'abandon de la course, puis des moyennes correspondantes pour chaque modalité du type d'abandon (celles-ci sont ici calculées directement sur l'ensemble des états de toutes les courses)

Variable	p-valeur de l'ANOVA	Moyenne des coureurs n'ayant pas abandonné	Moyenne des coureurs ayant abandonné à cause de la barrière horaire	Moyenne des coureurs ayant abandonné pour une autre raison
Barrière horaire (en secondes)	3.8e-3	39545.37	42742.7	38525.66
Distance (en km) par rapport à la balise précédente	7.6e-4	7.88	7.38	7.66
Vitesse du vent	3.5e-4	17.39	17.22	16.89
Humidité	9.1e-19	0.71	0.77	0.71
Temps (en secondes)	1.0e-23	73139.19	66031.43	66700.11
Température moyenne	3.7e-46	11.47	12.35	12.78
Température ressentie	4.4e-59	6.88	7.67	8.31
Altitude	5.0e-63	1646.76	1582.45	1521.03
Vitesse ascensionnelle	5.6e-64	-77.55	-233.41	-228.9
Temps arrêté cumulé	3.4e-70	3926.09	1723.71	3011.38
Vitesse	1.4e-73	5.10	3.62	4.28
Dénivelé	1.1e-113	13.26	-337.63	-215.31
Horaire (en secondes par rapport à 0H00)	5.9e-142	39225.82	44326.48	49382.80
Pourcentage de dénivelé (par rapport à la distance pour arriver à la balise)	4.6e-145	0.56	-5.93	-4.00
Différence (en secondes) entre l'heure d'arrivée et l'heure de la barrière horaire	3.6e-246	-13872.33	1583.78	-8394.96

TABLE A-6 – p-valeurs des ANOVA pour l'effet de la variable du type d'abandon par balise sur chaque variable quantitative variant en fonction des balises et moyennes correspondantes

Variable	Paires de facteurs dont les moyennes pour la variable considérée sont significativement différentes au niveau 5% (0 pour "pas d'abandon", 1 pour "abandon à cause de la barrière horaire" et 2 pour "abandon pour une autre raison"
Barrière horaire	0-1 1-2
Distance (en km) par rapport à la balise précédente	0-1 0-2
Vitesse du vent	0-2
Humidité	0-1 1-2
Temps (en secondes)	0-1 0-2
Température moyenne	0-1 0-2
Température ressentie	toutes
Altitude	toutes
Vitesse ascensionnelle	0-1 0-2
Temps arrêté cumulé	toutes
Vitesse	toutes
Dénivelé	toutes
Horaire (en secondes par rapport à 0H00)	toutes
Pourcentage de dénivelé (par rapport à la distance pour arriver à la balise)	toutes

Différence (en secondes) entre l'heure d'arrivée et l'heure de la barrière horaire	toutes

TABLE A-7 – Récapitulatif des paires de facteurs (de la variable du type d'abandon par balise) dont les moyennes pour la variable considérée sont significativement différentes au niveau 0.05 pour le test de Student

1.3 Classification ascendante hiérarchique

La classification ascendante hiérarchique (CAH) est une méthode de classification dont l'objectif est de répartir un ensemble Ω de n individus dans un certain nombre de classes. Elle part d'une situation où tous les individus sont seuls dans une classe, puis sont rassemblés en classes de plus en plus grandes. Elle produit une hiérarchie notée H, l'ensemble des classes à toutes les étapes de l'algorithme, qui vérifie les propriétés suivantes :

- $\Omega \in H$: au sommet de la hiérarchie, lorsqu'on groupe de manière à obtenir une seule classe, tous les individus sont regroupés
- $\forall \omega \in \Omega, \{\omega\} \in H$: en bas de la hiérarchie, tous les individus sont seuls
- $\forall (h, h') \in H^2, h \cap h' = h$ ou $h \subset h'$ ou $h' \subset h$: si l'on considère deux classes du regroupement, alors soit elles n'ont pas d'individus en commun, soit l'une est incluse dans l'autre.

La méthode suppose qu'on dispose d'une mesure de dissimilarité entre les individus x et y, et est notée $dissim(x, y)$. Ici, on utilise la distance de Gower qui est un indice compris entre 0 et 1. Si l'indice vaut 1, les deux individus sont identiques. À l'opposé, s'il vaut 0, les deux individus considérés n'ont pas de point commun. Si l'on note S_g l'indice de similarité de Gower, la distance de Gower D_g s'obtient simplement de la manière suivante : $D_g = 1 - S_g$.

Ainsi, lindice de similarité de Gower entre deux individus x_1 et x_2 se calcule de la manière suivante : $S_g(x_1, x_2) = \frac{1}{p} \sum_{j=1}^{p} s_{12j}$. p représente le nombre total de variables utilisées pour comparer les deux individus et s_{12j} représente la similarité partielle entre les individus 1 et 2 concernant le descripteur j. Cette similarité partielle se calcule différemment s'il s'agit d'une variable qualitative ou quantitative :

— **variable qualitative :** s_{12j} vaut 1 si la variable j prend la même valeur pour les individus 1 et 2, et vaut 0 sinon

— **variable quantitative :** la différence absolue entre les valeurs des deux variables est tout d'abord calculée, soit $|y_{1j} - y_{2j}|$. L'écart maximum observé est ensuite déterminé sur l'ensemble du fichier et est noté R_j. Dès lors, la similarité partielle vaut $s_{12j} = 1 - \frac{|y_{1j} - y_{2j}|}{R_j}$.

Pour réaliser une classification ascendante hiérarchique, il est nécessaire de choisir une méthode d'agrégation des classes C_1 et C_2 à partir d'une mesure de dissimilarité entre classes. On utilise ici le critère de Ward qui vise à maximiser l'inertie inter-classe :

- $dissim(C_1, C_2) = \frac{n_1 \times n_2}{n_1 + n_2} \times dissim(G_1, G_2)$ avec n_1 et n_2 les effectifs des deux classes, G_1 et G_2 leurs centres de gravité respectifs.

Ces regroupements successifs produisent un dendrogramme, dont la racine correspond à la classe regroupant l'ensemble des individus. Cet arbre de classification binaire représente une hiérarchie de partitions que l'on peut tronquer à un niveau donné.

On souhaite ici construire des partitions évolutives entre 2014 et 2018 afin d'observer l'effet des caractéristiques des coureurs sur leur abandon, notamment la significativité des différences entre chaque cluster de coureurs. On construit donc 5 jeux de données prenant en compte les mêmes variables et incluant l'ensemble cumulé des coureurs d'une année à l'autre. Les variables utilisées pour chaque jeu de données sont les suivantes : *sexe*, *année_naissance*, *taille*, *cote_ITRA_année*, *nb_abandon_année* et *prop_abandon_année*. Chacun des jeux de données possède respectivement 2429, 4758, 6834, 8726 et 10605 coureurs, auxquels on associe les 7 variables citées précédemment.

La FIGURE A-1 ci-dessous présente l'inertie relative du dendrogramme selon le nombre de clusters retenus sur la base des coureurs 2014 :

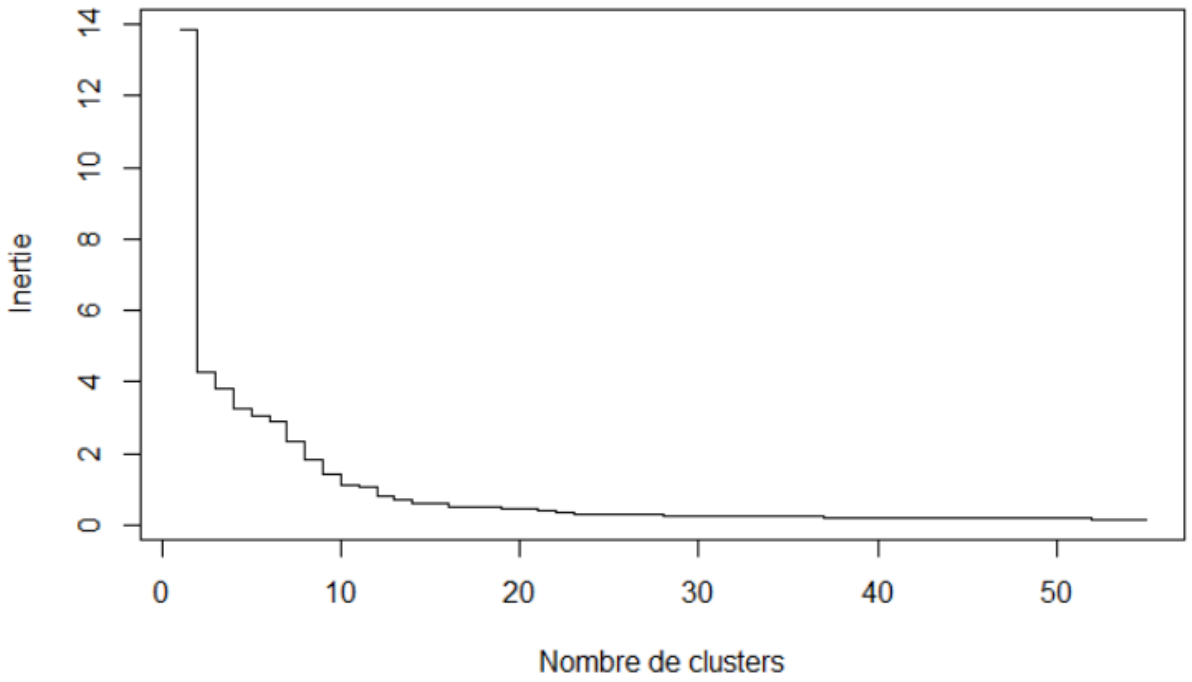

FIGURE A-1 – Inertie relative du dendrogramme selon le nombre de clusters sur la base 2014

La règle du coude nous permet ainsi de retenir 2 clusters. On peut alors afficher la partition obtenue du dengrogramme en 2 classes sur la base des coureurs 2014 comme le montre la FIGURE A-2 :

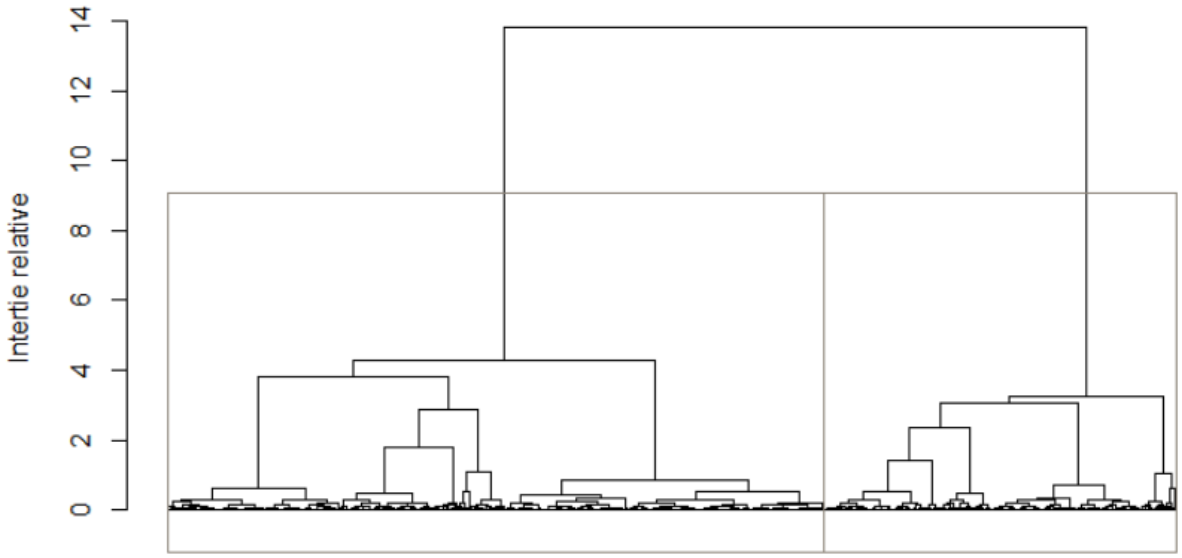

FIGURE A-2 – Partition du dendrogramme en 2 classes sur la base 2014

On obtient ainsi une partition en 2 clusters comprenant respectivement 1582 et 847 coureurs. L'indice silhouette, compris entre -1 et 1, permet de caractériser la qualité d'appartenance des coureurs à leur cluster. Ici, cet indice vaut respectivement 0.75 et 0.72 pour les 2 clusters, ce qui est proche de 1 et permet de conclure que la partition obtenue est relativement bonne pour la base des coureurs 2014.

On peut à présent interpréter les caractéristiques des clusters de coureurs et déterminer si les moyennes et les proportions pour chaque variable quantitative et qualitative sont significativement différentes entre les 2 classes. Pour cela, on réalise un test de Student de comparaison de moyennes pour les variables quantitatives. On vérifie

au préalable deux hypothèses : les ensembles de valeur comparés doivent suivre approximativement une loi normale et être de même variance, sans quoi le test de Student n'est pas valide. Pour vérifier que l'hypothèse de normalité est acceptable, on peut utiliser le test de normalité de Shapiro-Wilk et pour vérifier celle d'égalité des variances, on réalise un test de Fisher. Si l'une ou les deux hypothèses ne sont pas validées, on utilise alors le test non paramétrique de Wilcoxon/Mann-Whitney, qui ne repose sur aucune hypothèse sur la distribution des échantillons comparés. Enfin, pour comparer deux proportions, on effectue un test du chi-2 de comparaison de proportions, qui ne repose sur aucune hypothèse non plus.

Cependant, les tests effectués pour chacune des variables pour comparer les moyennes et proportions entre chaque cluster souffrent de certaines limites. D'une part, certaines proportions sont pratiquement nulles pour certaines modalités de variables, à cause d'effectifs très faibles. D'autre part, le test non paramétrique de Wilcoxon/Mann-Whitney suppose néanmoins que les distributions des échantillons comparés ont des formes à peu près comparables en termes de variance, ce qui n'est pas toujours le cas.

Les mêmes analyses ont été réalisées pour les CAH obtenus à partir des bases 2014-2015, 2014-2016, 2014-2017 et 2014-2018. D'après la règle du coude, 2 clusters ont été retenus pour l'interprétation des partitions de coureurs. Les indices silhouette sont globalement assez élevés et supérieurs à 0.5, indiquant que les partitions obtenues sont plutôt bonnes. Les résultats obtenus sont finalement présentés ci-dessous dans la TABLE A-8 :

Variable	**2014**	**2014-2015**	**2014-2016**	**2014-2017**	**2014-2018**
Sexe	0.02	0.03	3×10^{-5}	2×10^{-6}	2×10^{-7}
Année de naissance	3×10^{-4}	2×10^{-11}	$<2 \times 10^{-12}$	$<2 \times 10^{-12}$	$<2 \times 10^{-12}$
taille	0.2	0.001	0.01	0.002	7×10^{-5}
Cote ITRA	1×10^{-4}	2×10^{-11}	$<2 \times 10^{-12}$	$<2 \times 10^{-12}$	$<2 \times 10^{-12}$
Nombre de participations	X	$<2 \times 10^{-12}$	$<2 \times 10^{-12}$	$<2 \times 10^{-12}$	0.04
Nombre d'abandons	$<2 \times 10^{-12}$	$<2 \times 10^{-12}$	$<2 \times 10^{-12}$	$<2 \times 10^{-12}$	$<2 \times 10^{-12}$
Proportion d'abandon	$<2 \times 10^{-12}$	$<2 \times 10^{-12}$	$<2 \times 10^{-12}$	$<2 \times 10^{-12}$	$<2 \times 10^{-12}$

TABLE A-8 – p-valeurs des tests de comparaisons de moyennes et de proportions entre les 2 clusters pour chaque variable en fonction des différentes bases de coureurs

Globalement, sur l'ensemble des différentes bases de coureurs, les moyennes et les proportions des variables utilisées pour construire les CAH, sont significativement différentes entre les 2 clusters. En effet, toutes les p-valeurs sont inférieures voire très largement inférieures à 0.05, exceptée une à cause d'une valeur unique pour la variable associée. Ainsi, les moyennes et proportions les plus significativement différentes entre les 2 clusters, sont celles des variables du nombre d'abandons, de la proportion d'abandon ainsi que de l'année de naissance et de la cote ITRA.

[2] Première modélisation statistique

2.1 Construction des échantillons

Après construction des échantillons d'entraînement, de validation et de test, nous avons vérifié la cohérence de la répartition des modalités de la variable abandon.

Échantillon	**Apprentissage**	**Validation**	**Test**
Proportion d'abandon	0.37	0.37	0.30

TABLE A-9 – Vérification de l'équilibre des échantillons de travail sur l'abandon

L'équilibre des échantillons en termes d'abandon est bien vérifiée.

2.2 Modélisation logistique

Afin de sélectionner les régresseurs à inclure dans ce modèle, nous avons construit cinq modèles où chacune des variables (sexe, cote ITRA, âge et taille) étaient le seul régresseur non constant.

Variable	**Coefficient estimé**	**P-valeur**
Sexe H/F	-0.30	2.0×10^{-4}
Âge	0.04	$\leq 2 \times 10^{-16}$
Taille S/XS	0.48	0.15
Taille M/XS	0.42	0.21
Taille L/XS	0.60	0.08
Taille XL/XS	0.81	0.02
Cote ITRA	-2.5×10^{-3}	$\leq 2 \times 10^{-16}$

TABLE A-10 – Tableau des coefficients estimés et des p-valeurs des cinq variables pré-sélectionnées utilisées comme unique régresseur non-constant dans les cinq régressions logistiques (échantillon d'apprentissage / $n = 7552$)

Avec le TABLEAU A-10, nous constatons que les variables sexe, age et cote ITRA sont significatives au niveau 20%, raisonnable pour une pré-sélection de variables. La modalité M de la variable taille, considérée par rapport à la modalité de référence XS, n'est pas significative au niveau souhaité.

Malgré cette réserve, la procédure stepwise a permis de sélectionner toutes ces variables au sein du modèle complet. Le TABLEAU A-11 permet de préciser les coefficients et les p-valeurs des variables.

Variable	**Coefficient estimé**	**P-valeur**
Sexe H/F	-0.38	3.2×10^{-5}
Âge	0.04	$\leq 2.2 \times 10^{-16}$
Taille S/XS	0.83	1.7×10^{-2}
Taille M/XS	0.79	2.5×10^{-2}
Taille L/XS	0.88	1.3×10^{-2}
Taille XL/XS	1.04	5.4×10^{-3}
Cote ITRA	-1.36×10^{-3}	1.1×10^{-6}

TABLE A-11 – Tableau des coefficients estimés et des p-valeurs pour le modèle complet (échantillon d'apprentissage / $n = 7552$)

Tous les coefficients sont significatifs au niveau 5%. Nous avons détaillé leur interprétation via les odds-ratios présentés dans le corps de ce rapport.

La phase de test décrite dans le corps de ce rapport montre un défaut de performance de notre modèle. Sa courbe ROC représentée par la FIGURE A-3 confirme ce constat ; le modèle paraît effectivement légèrement meilleur que le hasard seulement.

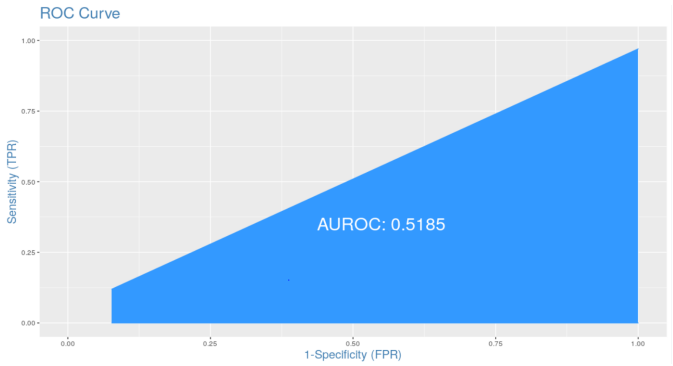

FIGURE A-3 – Courbe ROC pour la régression logistique (échantillon de test / $n = 2554$)

Le modèle est en effet peu discriminant. Il a été nécessaire de vérifier la dispersion des probabilités d'abandon estimées sur l'échantillon de test avec la TABLE A-12, afin d'évaluer la pertinence du seuil de 0.5 et de comprendre les erreurs de prédiction induites.

Minimum	Q1	Q2	Q3	Maximum
0.15	0.30	0.36	0.42	0.66

TABLE A-12 – Quartiles des probabilités d'abandon estimées sur l'échantillon de test