

RAPPORT DE STAGE

Etude des déterminants de la performance en natation de haut-niveau

GOUSSET Léonard

Du 12 Avril au 28 Mai 2021

**DUT Statistiques et Informatique
Décisionnelle**

2020 / 2021

REMERCIEMENTS

Tout d'abord, je tiens à remercier infiniment mes maîtres de stages :

Arthur Leroy, Doctorant en mathématiques appliquées et enseignant à l'IUT de Paris – Rives de Seine pour m'avoir enseigné les qualités d'un excellent statisticien et pour la confiance qu'il m'a accordée durant ce stage.

Robin Pla, conseiller technique national à la Fédération Française de Natation, pour l'apport de ses connaissances du terrain et tous ses conseils.

Je les remercie tous les deux pour leur disponibilité et leur écoute, malgré les conditions actuelles.

Je tiens également à remercier tous les membres du laboratoire MAP5 que j'ai pu croiser, ils m'ont intégré au groupe dès mon arrivée et m'ont mis dans des conditions de travail optimales pour ce stage.

TABLE DES MATIERES

PARTIE 1 : INTRODUCTION.....	6
1.1 – Choix du stage	7
1.2 – Contexte.....	8
1.2.1 – Contexte général	8
1.2.2 – Objectifs et problématique.....	8
1.3 – Le MAP5	9
1.3.1 – Présentation générale	9
1.3.2 – L'équipe Statistique	10
1.4 – Les ressources utilisées pour l'analyse	11
1.4.1 – Langage R et quelques packages	11
1.4.2 – Les statistiques bayésiennes.....	12
1.4.2.1 – Méthode Fréquentiste / Méthode Bayésienne	12
1.4.2.2 – Modèle Bayésien	13
1.5 –Données à disposition	14
1.5.1 – Description de la base	14
1.5.2 – Récupération des données.....	15
1.6 – Introduction aux analyses	16
PARTIE 2 : ANALYSE	17
2.1 – Quantifier les écarts de course	18
2.1.1 – Introduction.....	18
2.1.2 – Résultats.....	18
2.2 – LA VITESSE	25
2.2.1 – Un outil purement déterministe	25
2.2.1.1 – Introduction	25
2.2.2 – Comment répartir sa vitesse ?.....	28
2.2.2.1 – Introduction	28
2.2.2.1.1 – Contexte et objectifs.....	28
2.2.2.1.2 – Compétence acquise	28
2.2.2.2 – Analyse.....	29
2.2.2.2.1 – Classement des portions	29
2.2.2.2.2 – Contribution des vitesses.....	32

2.3 – BONUS : LES HEATMAPS	35
2.3.1 – Introduction.....	35
2.3.2 – Analyses	36
2.3.2.1 – Création des Heatmaps	36
2.3.2.2 – Résultats	39
PARTIE 3 : CONCLUSION	44
3.1 – CONCLUSION DE L’ETUDE	45
3.2 – BILAN PERSONNEL ET PROFESSIONNEL	46
3.3 – ANNEXE	46
3.4 – REFERENCES	47

TABLE DES FIGURES/TABLEAUX

Figure 1 : 18èmes championnats du monde de natation à Gwangju, en Corée du Sud	14
Figure 2 : Matériel de mesure pour récupérer les données	15
Figure 3 : Ecart de courses des hommes (en pourcentage).....	19
Figure 4 : Exemple Nage libre	20
Figure 5 : Ecart de course chez les femmes (en pourcentage).....	21
Figure 6 : Exemple brasse	22
Figure 7 : Comparaison des écarts entre les hommes et les femmes	23
Figure 8 : Exemple de code de la partie UI.....	25
Figure 9 : Exemple de code de la partie Serveur.....	25
Figure 10 : Mon application Shiny.....	26
Figure 11 : Code pour créer les menus déroulant pour la nage et le sexe	27
Figure 12 : Menu déroulant pour les nages	27
Figure 13 : Menu déroulant pour les sexes	27
Figure 14 : Rendu de l'analyse des vitesses dans mon application shiny.....	29
Figure 15 : Exemple de contribution des vitesses pour la brasse.....	32
Figure 16 : Exemple du temps d'entraînement idéal pour la brasse chez les hommes.....	33
Figure 17 : Création Heatmap	36
Figure 18 : Rajout des données	36
Figure 19 : filtre des données pour rendre le graphique plus lisible	37
Figure 20 : Ajout des courbes de vitesse.....	37
Figure 21 : Ajout de la fonction ggplotly	38
Figure 22 : Rendu des heatmaps dans mon application shiny.....	38
Figure 23 : Exemple sur le 2ème quart de course chez les femmes en brasse	39
Figure 24 : Exemple brasse femmes 0-25m	39
Figure 25 : Nage libre Homme 50-75m	40
Figure 26 : Nage libre Homme 25-50m	40
Figure 27 : Nage libre Homme 75-100m	40
Figure 28 : Papillon deuxième 50-75m et 75-100m homme.....	41
Figure 29 : Papillon deuxième 50-75m et 75-100m femme.....	41
Tableau 1 : Classement des portions par nage et par sexe de la contribution de la vitesse sur le temps final.....	30
Tableau 2 : Classement des portions selon le sexe.....	30
Tableau 3 : Classement des portions en dos.....	31
Tableau 4 : Classement des portions en nage libre	31

PARTIE 1 : INTRODUCTION

1.1 – Choix du stage

Pour valider notre DUT Statistiques et Informatique Décisionnelle, nous devons effectuer un stage d'une durée de 10 semaines minimum (6 cette année à cause de la COVID-19) à la fin de la deuxième année. Ce stage est une opportunité de mettre en pratique tout ce que l'on a appris durant les deux ans de DUT, il nous permet également de découvrir le monde de l'entreprise, auquel nous serons confrontés dans quelques années.

L'offre à laquelle j'ai postulé, en binôme avec James HAZIZA portait sur une étude de la natation de haut niveau, avec une introduction aux méthodes statistiques Bayésiennes et aux approches Monte Carlo Markov Chains (MCMC). Dans l'offre étaient incluses 2 références dont un article sur l'impact de la morphologie au niveau de la performance des nageurs de haut-niveau : <https://hal-insep.archives-ouvertes.fr/hal-02925019/document>. Il a été réalisé en partie par mon encadrant principal : Arthur LEROY et mon co-encadrant : Robin PLA. Durant la deuxième année de mon DUT, j'avais eu l'occasion de travailler sur le même sujet en utilisant également les méthodes Bayésiennes et approches MCMC, et je trouvais cela intéressant de voir jusqu'où des statisticiens sont allés avec un sujet plutôt simple.

Travailler dans un laboratoire de recherche pour la Fédération Française de Natation était pour moi une offre très intéressante, d'autant plus que nous avons à disposition une base de données encore inexplorée. Dans ce cadre-là, j'ai eu l'opportunité de découvrir le métier de chercheur en effectuant mon stage du 12 Avril au 28 Mai 2021 au laboratoire de recherche MAP5 (Mathématiques Appliquées Paris 5), avec pour sujet de stage :

« L'étude des déterminants de la performance en natation de haut-niveau »



FÉDÉRATION FRANÇAISE
NATATION

1.2 – Contexte

1.2.1 – Contexte général

Ayant déjà eu l'occasion de travailler ensemble, Robin PLA a demandé une collaboration à Arthur LEROY afin de réaliser une étude sur les déterminants de la performance des athlètes de haut-niveau en natation afin d'aider les entraîneurs, les cadres techniques voir même les nageurs. Ceux-ci bénéficieront d'un étude purement statistique réalisée sur la base des 18èmes championnats du monde de natation se déroulant du 12 au 28 Juillet 2019 à Gwangju, en Corée du Sud. 12^{ème} au classement des médailles, l'équipe de France n'avait pas connu une place si basse depuis les championnats du monde de 2009 à Rome, en Italie (21^{ème} à l'époque).

1.2.2 – Objectifs et problématique

Le but de ce stage est d'aider les membres de la FFN à améliorer la performance des nageurs, ou du moins proposer des solutions pour améliorer cette performance, en conséquence, nous avons défini plusieurs objectifs :

- Apporter des informations encore inconnues avec un regard nouveau.
- Construire des outils faciles d'utilisation.
- Avancer des solutions pour palier à certains problèmes.

Pour compléter chacun de ses objectifs, j'ai décidé de répondre à la problématique suivante :

Quels sont les déterminants de la performance en natation de haut-niveau, et comment transformer les résultats en outils simple d'utilisation ?

1.3 – Le MAP5

1.3.1 – Présentation générale

Le MAP5 est un laboratoire en mathématiques appliquées situé dans Paris 5, il a pris la suite d'équipes principalement tournées vers les applications de la statistique au domaine biomédical et vers l'image.

Fondé en 2004, le MAP5 ne cesse de se développer en ayant pour ambition d'établir un excellent niveau de recherche en mathématiques appliquées. Bien que rattaché à l'Université de Paris et à l'Institut National des Sciences Mathématiques et de leurs Interactions (INSMI) du CNRS, il fait également partie de la Fondation Sciences Mathématiques de Paris (FSMP).



Dans un laboratoire de ce type, il n'y a pas de hiérarchie comme dans une entreprise plus « classique ». Il y a certes une directrice (Anne Estrade), mais elle n'a pas de pouvoir sur les recherches effectuées et leurs thèmes.

Aujourd'hui, le laboratoire est composé de 55 permanents (25 de plus qu'en 2005), mais également de doctorants, de postdocs et d'ATER (Attaché Temporaire d'Enseignement et de Recherche) qui sont répartis en 4 équipes, travaillant chacune dans un domaine particulier des mathématiques appliquées :

- L'équipe Statistique
- L'équipe Probabilité
- L'équipe Traitement d'images
- L'équipe Modélisation, analyse et simulation

1.3.2 – L'équipe Statistique

Pour ma part, j'ai intégré le pôle statistique, où se trouve mon encadrant principal : Arthur LEROY. Cette équipe est composée de 17 enseignants-chercheurs ainsi que de 11 doctorants et post-doctorants. Tous les 15 jours en moyenne, un séminaire où sont exposés les travaux en cours des chercheurs de l'équipe statistique, se réunit. Ces recherches menées au sein de cette équipe s'organisent autour de 4 thèmes principaux :

- Thème 1 : Apprentissage et estimation
 - Méthodes de classification et régression.
 - Estimation non-paramétrique et semi paramétrique.
 - Analyse de causalité et médiation
 - Apprentissage ciblé
 - Application en médecine personnalisée, étude des réseaux, performance sportive.

- Thème 2 : Statistique des modèles discrets
 - Erreur de mesure.
 - Modèles de Markov cachés.
 - Modèles mixtes-Données longitudinales.
 - Outils statistiques : Mesure de dépendance, processus empirique et coût de transport.

- Thème 3 : Statistique des processus à temps continu
 - Processus ponctuel et survie, processus de comptage.
 - Equations différentielles stochastiques.
 - Inférence statistique pour les processus de Lévy et Lévy mixtes.
 - Modèles de shot noise.
 - Théorie des valeurs extrêmes.

- Thème 4 : Epidémiologie et génétique
 - Epidémiologie du VIH/SIDA, dengue et paludisme.
 - Fertilité, cancer et recherche clinique.
 - Génétique épidémiologique et du vieillissement.
 - Biopuces et données NGS.

Lors des différentes recherches, les chercheurs peuvent être amenés à collaborer avec des personnes extérieures, des professionnels d'un sujet en particulier. Par exemple mon co-encadrant Robin PLA (conseiller technique national de la FFN) a déjà travaillé avec des membres du MAP5, pour réaliser un article sur l'impact de la morphologie sur la performance en natation de haut-niveau, comme je l'ai dit dans l'introduction.

1.4 – Les ressources utilisées pour l’analyse

1.4.1 – Langage R et quelques packages

L’ensemble du traitement et de l’analyse de données a été effectué avec le langage de programmation R, plus précisément sur Rstudio qui est un IDE (Integrated Development Environment) de R, c’est un environnement qui facilite la prise en main, l’exécution du code et la data visualisation.



Au sein même de R, plusieurs packages m’ont été nécessaires, que ce soit pour le traitement des données ou pour la data visualisation :

« tidyverse »

Facilite la conception et la compréhension du code



« ggplot2 »

Pour la réalisation de tous les graphiques



« Shiny »

Réaliser un interface utilisateur interactive



1.4.2 – Les statistiques bayésiennes

1.4.2.1 – Méthode Fréquentiste / Méthode Bayésienne

Comme il était annoncé dans l'offre de stage, il était question d'introduire, à l'analyse de données, les méthodes de statistiques Bayésiennes et les approches Monte Carlo Markov Chains (MCMC), mais qu'est-ce que les statistiques Bayésiennes ?

Aujourd'hui, il existe deux méthodes statistiques, la méthode fréquentiste qui correspond aux statistiques que l'on voit tous les jours à la télévision ou dans des articles, et la méthode Bayésienne. En un mot, les statistiques Fréquentistes correspondent à la probabilité des événements selon une certaine théorie, ils peuvent être mis sous forme de fréquence, alors que les statistiques Bayésiennes correspondent à la probabilité des théories au vu de certains événements.

Le Bayésianisme s'inspire tout droit de la formule de Bayes, elle-même inventée par Thomas Bayes qui dit que :

$$P(A \cap B) = P(A) * P(B|A)$$

$$P(A \cap B) = P(B) * P(A|B)$$

Donc : $P(A) * P(B|A) = P(B) * P(A|B)$

Et :
$$P(B|A) = \frac{P(B)*P(A|B)}{P(A)}$$

Avant cela on ne pouvait calculer la probabilité des événements que si on en connaissait la cause (soit $P(A|B)$), mais grâce à la formule précédente, le célèbre Pierre Simon Laplace écrit un mémoire sur la probabilité des causes par les événements ((soit $P(B|A)$), il appelle ça « les probabilités inverses »).

1.4.2.2 – Modèle Bayésien

Dans le cadre de mon stage, j'ai eu l'opportunité de traiter une partie des données avec la méthode Bayésienne, lors de la conception d'un modèle.

Pour réaliser un modèle linéaire Bayésien sur R, j'ai dû installer le package « bayestR », il permet d'écrire le modèle, d'estimer les paramètres des variables d'entrée par rapport à une variable de sortie et d'analyser les lois obtenues à posteriori. Cela se fait grâce à la fonction « stan_glm » (issue du logiciel stan lui-même) :

```
```{r}
modele <- stan_glm(var_sortie ~ var_entree1 + var_entree2 + var_entree3, data = db)
```
```



J'expliquerai plus en détail ce que fait cette fonction et ce que représente un modèle bayésien dans la partie qui lui est réservé pour les analyses.

1.5 – Données à disposition

1.5.1 – Description de la base

La base de données sur laquelle nous avons fait notre étude comprend les performances des 32 finalistes Hommes et 32 finalistes Femmes des finales de 100m (Nage libre / Dos / Brasse / Papillon) lors des championnats du monde de natation de 2019 à Gwangju, en Corée du sud.



Figure 1 : 18èmes championnats du monde de natation à Gwangju, en Corée du Sud

Aux 64 nageurs sont attribués 48 valeurs correspondant aux 48 variables de la base de données, nous permettant de déterminer les indicateurs de performance, afin d'aider les entraîneurs et conseiller techniques de la FFN. Au sein des 48 variables, on peut en discerner plusieurs types :

- Les variables Temporelles : Le temps de chaque nageur sur différentes parties de course, le temps final, le temps de réaction, le temps de vol, etc...
- Les variables Physiologiques : La fréquence de coups de bras par minute, le nombre de coups de bras sur le 50m ou encore la distance par cycle de bras¹

¹ 1 cycle de bras = 2 coups de bras en Nage libre et dos

1 cycle de bras = 1 mouvement de bras en Papillon et Brasse

1.5.2 – Récupération des données

Depuis 1987, le service recherche et développement, de la FFN, dirigé aujourd’hui par Philippe Hellard recueille et analyse différents paramètres scientifiques pour permettre aux nageurs d’améliorer leurs performances.

Depuis 2008 ils utilisaient le logiciel « Swimwatch » pour ce qui est des analyses de course, et le programme « Ciren » pour les entraînements. Malheureusement, Simwatch était incomplet et difficile d’accès, ils ont donc développé un nouveau programme pour les compétitions : « Espadon ». Plus pratique et plus accessible, ce programme permet par exemple calculer la vitesse instantanée des nageurs, ou de comparer des courses qui ont plusieurs années d’écart. Testé pour la première fois pendant les championnats de France de Montpellier en novembre 2014, il dispose d’un système de reconnaissance automatique des nageurs, il suffit aux utilisateurs de choisir l’athlète en question pour en récolter les données.

Pour les championnats du monde 2019, c’est mon co-encadrant Robin PLA qui s’est chargé de filmer les 8 courses de 100m, puis de recueillir les données pour les mettre dans la base.

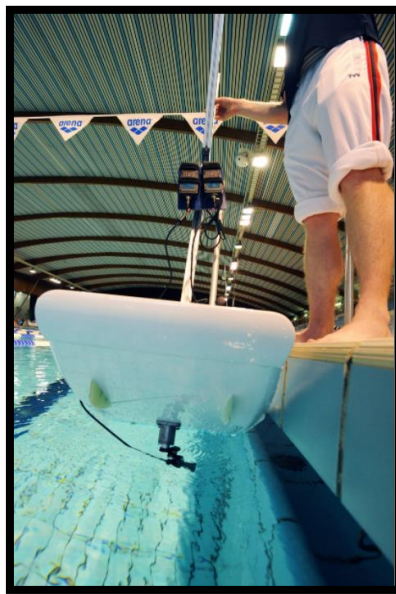


Figure 2 : Matériel de mesure pour récupérer les données

1.6 – Introduction aux analyses

Pendant les deux premières semaines, j'ai appliqué aux données de l'étude, toutes sortes de méthodes vues durant les deux ans de DUT (analyses univariées, bivariés, classification des données, relation linéaire, etc...) sans vraiment savoir où j'allais. C'est seulement après ces deux semaines que je suis passé dans l'optique « recherche », en reprenant les objectifs fixés, à savoir « Apporter des informations encore inconnues avec un regard nouveau ».

Avec mes encadrants et mon binôme, nous avons repris tous les résultats que nous avons jusqu'à présent pour ne garder que ceux qui avaient le plus de potentiel. Nous avons décidé de garder les résultats traitant :

- Des écarts de courses :
Pour cette partie nous avons pensé différemment, l'important dans une course n'est pas forcément d'aller vite, mais plus vite que les autres. Nous nous sommes donc intéressés aux écarts créés par différentes parties de course, si cela évoluait d'une course à l'autre, entre les hommes et les femmes, etc...
- De la vitesse :
Cette partie est consacrée à la modélisation bayésienne. Elle traitera des contributions de la vitesse sur certaines parties de course afin d'optimiser au maximum son énergie et donc diminuer au maximum son temps final.
- De la relation fréquence de coups de bras ~ Distance par cycle de bras :
Cette dernière partie a pour but de déterminer s'il y a une relation entre ces deux variables, un certain ratio d'équilibre. Mieux vaut être régulier tout au long de la course ou bien faut-il privilégier une fréquence plus élevée sur certaines parties de course ?

PARTIE 2 : ANALYSE

2.1 – Quantifier les écarts de course

2.1.1 – Introduction

Parmi tous les résultats obtenus lors de la phase d'exploration des données (2 semaines), la quantification des écarts de course a été le premier résultat assez intéressant pour l'approfondir et en faire une étude. Par quantification, j'entends donner une mesure (ici en pourcentage ou en seconde) sur les écarts créés entre les nageurs sur une partie de course. Vont-ils tous à la même vitesse sur une certaine portion où ont-ils une manière de gérer leur course qui leur est propre ?

Formule quantification des écarts² :

$$\frac{1}{8} \left[\sum_{i=1}^8 (|vitesse\ nageur_j\ portion_1 - mean(vitesse\ nageur_j\ portion_1)|) \right] \quad (1)$$

Cette formule correspond à une seule portion d'une seule nage d'un seul sexe, elle a donc été répétée 8 (nombre de portions) x 4 (nombre de nage) x 2 (homme/femmes)

Les enjeux de cette étude sont d'observer s'il y a des différences ou des similitudes entre les nages, entre les parties de course et/ou entre les hommes et les femmes. Plus précisément, cette étude cherche à déterminer quelles sont les portions où tous les nageurs vont à la même vitesse (donc peu d'écarts), ou au contraire de déterminer les portions où les nageurs ne vont pas à la même vitesse (c'est ici que écarts se créent). Ces deux objectifs sont tout aussi importants, en effet, si les nageurs ne travaillent que les parties qui créent les écarts et délaissent celles où ils vont tous à la même vitesse, l'effet inverse se produira et ce sera dans ces dernières qu'ils vont perdre du temps par rapport à leurs concurrents.

Toujours dans l'optique de rendre notre analyse la plus claire possible pour les membres de la FFN, on a opté pour la réalisation d'un simple tableau (ou heatmap) sur R avec la fonction `geom_tile` de `ggplot`. A chaque case est attribuée une couleur d'un certain degré de la valeur correspondante.

2.1.2 – Résultats

Sur les figures ci-dessous, chaque case correspond à une partie de course d'une nage précise. La couleur attribuée à chaque case correspond à l'écart moyen créé par cette portion, avec sa valeur exacte. Comme c'est en pourcentage, il fallait choisir une valeur de référence qui est à chaque fois la portion 0-15m du 100m Nage libre.

Exemple : Chez les hommes en nage libre, la partie 50-65m (52.3%) crée presque deux fois moins d'écarts que la partie 0-15m. (100%)

² pour une seule nage, un seul sexe et une seule portion ; ex : portion 0-15m du 100m Nage libre Homme

Pour les hommes :



Figure 3: Ecarts de courses des hommes
(en pourcentage)

La première chose que nous apprend ce graphique, c'est que la partie qui crée le plus d'écart chez les hommes est le départ en papillon (0-15m) et que celle qui en crée le moins est le 25-45m en nage libre.

En additionnant les pourcentages de chaque course, on remarque que le papillon est la nage qui crée le plus d'écart, 13.5% de plus que la nage libre, 19.7% de plus que la brasse et 25.4% de plus que le dos, qui est la nage créant le moins d'écart.

Pour les 4 courses, c'est toujours le départ qui crée le plus d'écarts entre les nageurs, cela peut s'expliquer par une différente morphologie des athlètes, les plus légers ont une mise en nage plus rapide, un temps de réaction plus rapide alors que les nageurs plus massifs mettent plus de temps à se mouvoir et donc à s'élancer.

Malgré le gradient de couleur, on peut distinguer 2 parties :

- La partie nagée : 15-25m ; 25-45m ; 65-75m ; 75-95m.
Qui correspond aux portions de course où seule la nage est prise en compte
- La partie non-nagée (ou partie technique) : 0-15m ; 45-50m ; 50-65m ; 95-100m.
Qui correspond aux portions de course où la technique entre en compte (le départ, la deuxième coulée et l'approche des murs du 50 et 100m).

On voit clairement la différence entre ces deux parties, sur la partie nagée les cases sont plutôt claires, ça veut dire que les athlètes vont globalement tous à la même vitesse alors que sur la partie technique la couleur des cases est plus foncée, ce qui équivaut à une moyenne des écarts élevée. Pour le voir, prenons l'exemple de la nage libre :

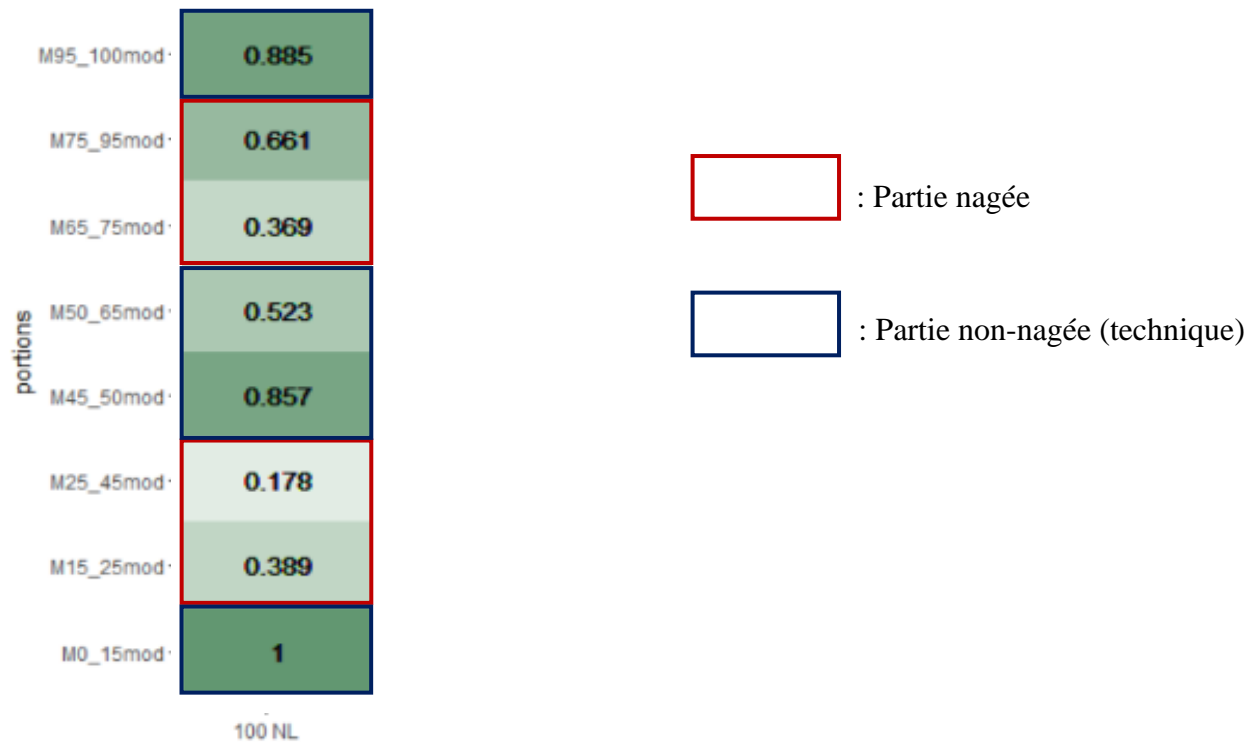


Figure 4 : Exemple Nage libre

Sur cet exemple on remarque que la portion créant le plus d'écart sur la partie nagée (0.661 pour 65-75m) en crée à peine plus que la dernière de la partie technique (0.523 pour 50-65m). De plus, en faisant le calcul sur les pourcentages, on voit que la partie technique crée 104.5% d'écarts en plus, soit plus du double alors qu'elle ne représente que 40% de la course !

En moyenne la partie technique crée 77% d'écarts en plus que la partie nagée :

- 104.5% pour la nage libre
- 44.9% pour le dos
- 104.5% pour la brasse
- 65% pour le papillon

Il y a donc une énorme différence de création d'écarts entre les parties nagées et non-nagées chez les hommes.

Pour les femmes :

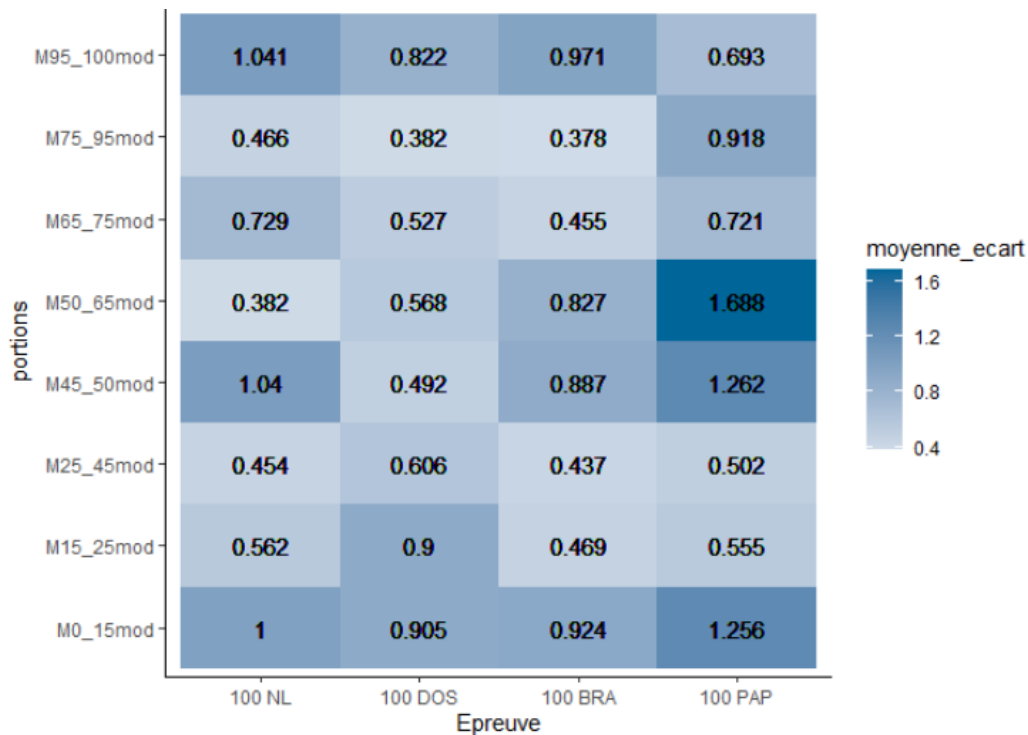


Figure 5: Ecart de course chez les femmes
(en pourcentage)

La portion qui crée le plus d'écart entre les nageuses est en papillon comme chez les hommes, mais sur la deuxième coulée (70% de plus que sur le départ du 100m nage libre). Comme pour les hommes, c'est le papillon qui crée le plus d'écarts, précisément 33.9% de plus que la nage libre, 42.2% de plus que la brasse et 46% de plus que le dos qui est là aussi la course qui en crée le moins.

En revanche, on ne voit que très peu de cases très claires (0.378 au plus bas pour le 75-95m en brasse), ce qui veut dire qu'entre elles, les femmes créent plus d'écarts que les hommes entre eux. La valeur de référence n'étant pas la même sur les deux graphiques, il est impossible de comparer les deux sexes.

Malgré l'absence de case claires, on peut distinguer le contraste de couleur entre les parties nagées et non-nagées, principalement sur la brasse :

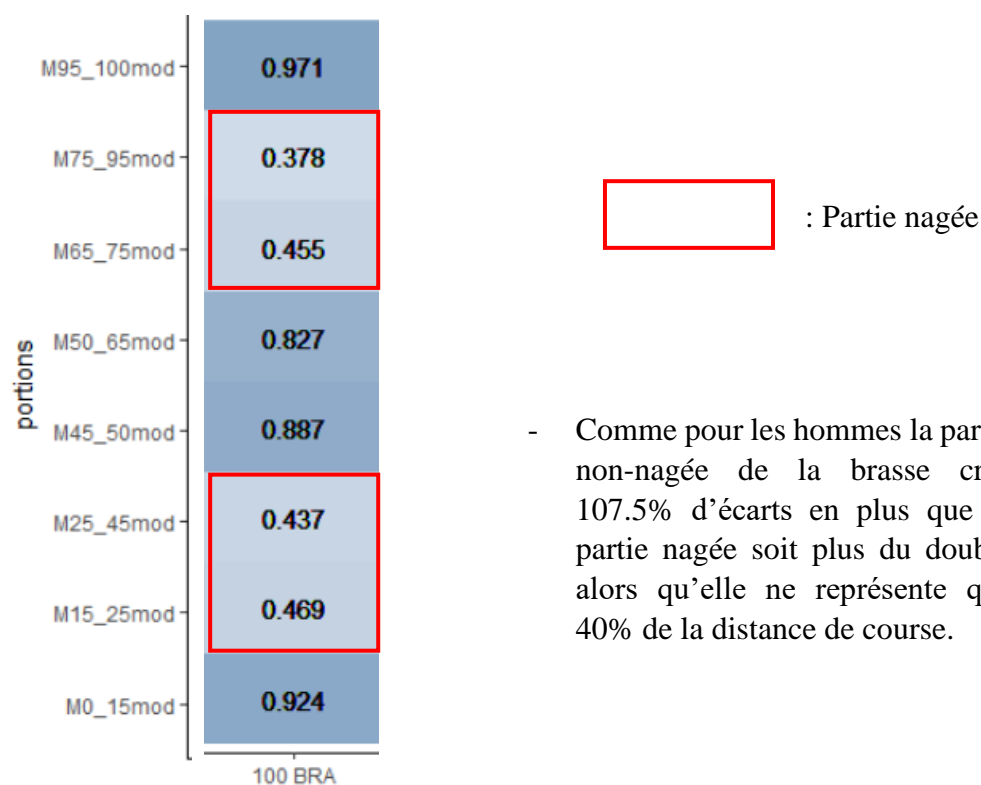


Figure 6: Exemple brasse

- Comme pour les hommes la partie non-nagée de la brasse crée 107.5% d'écarts en plus que la partie nagée soit plus du double alors qu'elle ne représente que 40% de la distance de course.

Que ce soit pour les hommes ou pour les femmes, on peut voir que la dernière portion (95-100m) crée en moyenne 15,2% des écarts de la course (un peu plus pour la nage libre et la brasse) alors que la distance de cette portion ne représente que 5% de la distance totale. Cela peut s'expliquer par la difficulté, en particulier pour la nage libre et la brasse, d'arriver au niveau du mur sur la fin d'un cycle de bras (ou coup de bras pour la nage libre et le dos). Les nageurs ayant chacun leur propre méthode de nage (une grosse fréquence pour une petite distance par cycle ou inversement), il est difficile pour eux d'arriver au 50 ou au 100m en finissant leur cycle, ils doivent donc allonger leur dernier coup de bras ou en commencer un nouveau pour toucher le mur final. C'est ce qui rend cette portion plutôt « créatrice d'écart » chez les hommes et chez les femmes.

Pour palier au problème de la comparaison entre sexe, il suffit de mettre une seule valeur seuil pour les deux tableaux : dans notre cas, le 100m Nage libre Homme.

Le graphique des hommes restera donc le même, seul celui des femmes changera, cela servira à déterminer si les femmes créent plus ou moins d'écarts que les hommes.



Figure 7: Comparaison des écarts entre les hommes et les femmes

: Valeur comparative pour les deux tableaux

En mettant les deux sexes en comparaison on voit bien que les hommes créent plus d'écarts que les femmes, 40% de plus en moyenne. La portion qui créait le plus d'écarts chez les hommes (0-15m papillon) en crée deux fois plus que cette même partie chez les femmes.

Les hommes sont donc ceux qui créent le plus d'écarts en moyenne, mais ce sont aussi eux qui en créent le moins. En effet, le plus petit pourcentage d'écart créé chez les femmes est de 22.1% (50-65 nage libre ; 75-95 dos) alors que chez les hommes c'est le 25-45m en nage libre avec 17.8%. On peut dire que les hommes sont moins réguliers que les femmes.

Le dernier point à apporter sur cette partie est au niveau du 100m dos. Contrairement aux autres nages où l'on voit des changements nets de couleurs des cases, il y en a très peu pour le dos, à part pour le départ des hommes. En réalité dans différentes compétitions, les vainqueurs de cette nage sont souvent premiers tout au long de la course, idem pour les derniers, il y a très peu de changement de position ou de retournement de situation.

Conclusion :

Chez les hommes et chez les femmes, les parties qui créent le plus d'écarts sont les parties non-nagées : le départ, la deuxième coulée et l'approche des murs du 50 et du 100m. Ce qui différencie les deux sexes c'est le pourcentage d'écarts créés : Les femmes créent plus d'écarts entre elles que les hommes entre eux, même si elles en créent moins que les hommes en général.

La nage libre et la brasse sont les deux nages qui se ressemblent le plus en termes d'écarts créés. On peut le voir grâce à la couleur des cases qui est dans les mêmes tons pour les deux nages. Le dos est une nage très régulière, il y a très peu d'écarts créés donc très peu de changements lors d'une course.

Au niveau des portions, c'est le départ (0-15m) qui crée le plus d'écart en moyenne (tout le temps chez les hommes et une fois chez les femmes). Les 5 derniers mètres de la course en créent aussi beaucoup à cause de la difficulté d'arriver sur une fin de cycle pour toucher le mur d'arrivée.

En prenant du recul, on se rend compte qu'il n'est pas facile de gagner du temps sur une partie précise, ou du moins qu'il est plus facile de gagner du temps sur certaines parties plutôt que d'autres. Bien que les portions non-nagée créent plus d'écart que les portions nagées, il ne faut pas que les nageurs ne s'entraînent qu'à la technique et délaissent la nage sur laquelle ils seront les seuls à prendre du retard sur la partie nagée, qui représente quand même 60% de la course.

Problèmes rencontrés :

Dans cette première partie, le principal problème que j'ai rencontré était au niveau de mon code R, il ne faisait pas moins de 800 lignes avec à l'intérieur une centaine de variables. Sachant que mon code pourrait être repris par mon encadrant ou par moi-même pour d'autres projets, j'ai essayé de le rendre plus « accessible » ou du moins plus clair. Pour cela j'ai utilisé des fonctions, elles permettent d'écrire un morceau de code de manière générale, avec des arguments qui leur sont propres, puis de les utiliser autant de fois que nécessaire dans le code en lui-même. Une fois à l'aise avec cette méthode, j'ai pu utiliser des fonctions imbriquées, ce sont des fonctions qui en appellent d'autres tout en gardant les bons paramètres d'entrée. Avec ces fonctions j'ai fait passer mon code de 800 à 85 lignes, ce qui facilite grandement son utilisation.³

³ Code des fonctions en annexe

2.2 – LA VITESSE

2.2.1 – Un outil purement déterministe

2.2.1.1 – Introduction

Un des deux objectifs fixés au début du stage était de créer des outils simples d'utilisation pour le terrain. Pour la variable vitesse, nous avons décidé de faire un simple modèle déterministe : le temps final en fonction des vitesses sur chaque partie de course.

La seule compétence statistique nécessaire était la création d'un modèle avec en variables d'entrée les différentes vitesses sur les portions de course et en variable de sortie la vitesse moyenne. En revanche j'ai développé une nouvelle compétence technique : la création d'une application Shiny. Shiny est un package très populaire chez les utilisateurs de R, il permet de créer des applications dynamiques sur le web. Cette application est composée de 2 parties :

- UI (User Interface) : qui contiendra tout ce que l'utilisateur pourra voir, à savoir les éléments de mise en page et les composantes de l'application. Ces composantes sont des variables/objets qui pourront communiquer avec la partie « Serveur »

```
ui <- fluidPage(  
  navbarPage(title = "INDICATEURS PERFORMANCE",  
    tabPanel("Modélisation part.1",  
      titlePanel("1 - Un outil pour le terrain"),  
      selectInput("Nage", "Choix de la nage", c("Nage libre", "Dos", "Brasse", "Papillon")),  
      selectInput("Sexe", "Choix du sexe", c("Homme", "Femme")),
```

Figure 8: Exemple de code de la partie UI

- Server : qui contiendra tout le code R. C'est ici que sont créés les éléments affichés dans l'interface grâce aux composantes créées dans la partie « UI ».

```
server <- function(input, output) {  
  output$coefficient1 <- renderPlot({  
    epreuve = case_when(paste(input$Nage4, input$Sexe4) == "Nage libre Homme" ~ "100 NL H",  
      paste(input$Nage4, input$Sexe4) == "Nage libre Femme" ~ "100 NL F",  
      paste(input$Nage4, input$Sexe4) == "Dos Homme" ~ "100 DOS H",  
      paste(input$Nage4, input$Sexe4) == "Dos Femme" ~ "100 DOS F",  
      paste(input$Nage4, input$Sexe4) == "Brasse Homme" ~ "100 BRA H",  
      paste(input$Nage4, input$Sexe4) == "Brasse Femme" ~ "100 BRA F",  
      paste(input$Nage4, input$Sexe4) == "Papillon Homme" ~ "100 PAP H",  
      paste(input$Nage4, input$Sexe4) == "Papillon Femme" ~ "100 PAP F")
```

Figure 9: Exemple de code de la partie Serveur

2.2.1.2 – Résultat

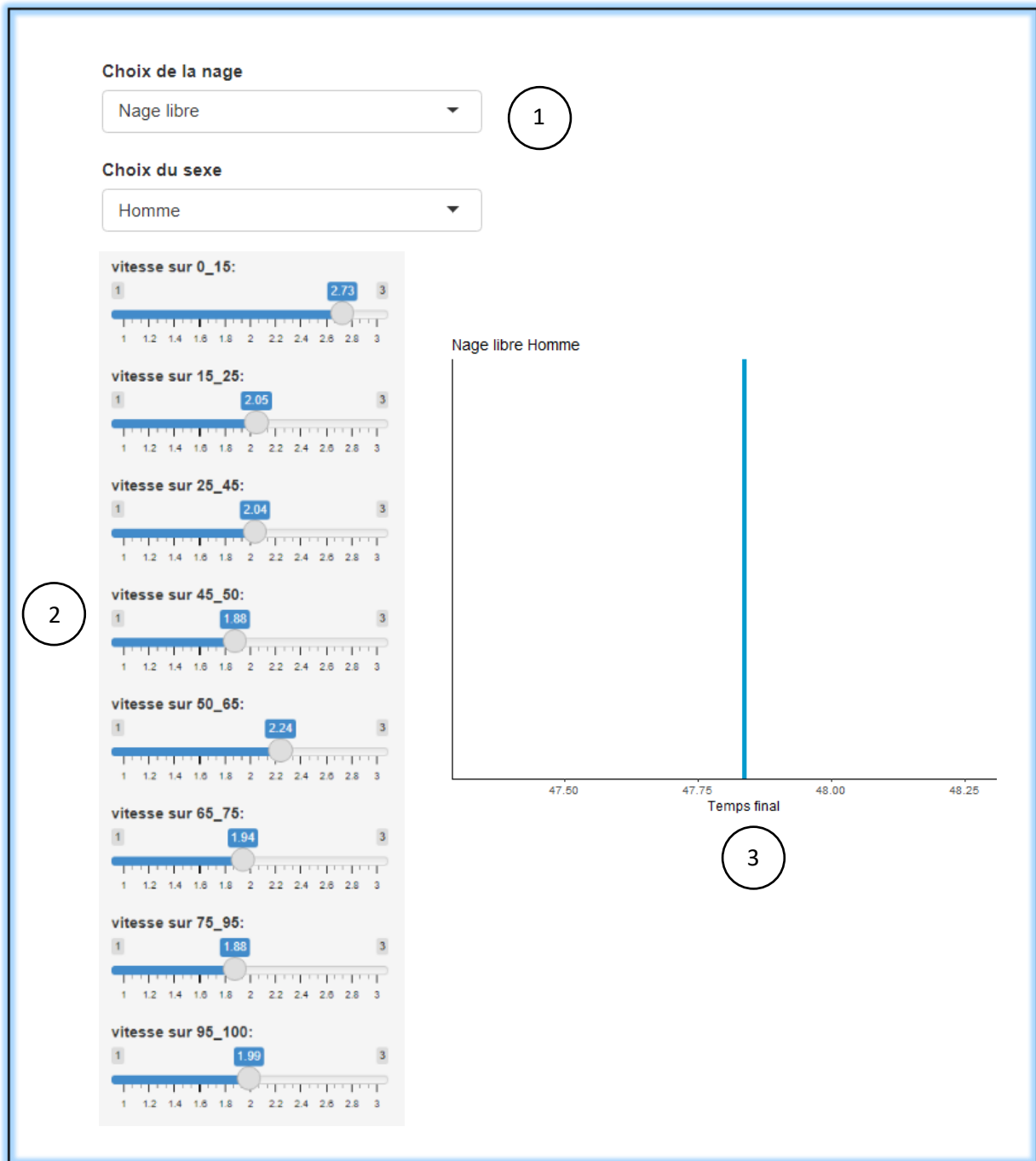


Figure 10 : Mon application Shiny,

Voici le rendu de mon application Shiny, ce que les entraîneurs et/ou conseillers techniques pourront utiliser pendant les entraînements, les phases de réflexion, etc... Cet onglet comporte 3 parties :

1

La première concerne le choix de la nage (nage libre/dos/brasse/papillon) et le choix du sexe (homme ou femme). Comme expliqué dans l'introduction, c'est ici que l'on va créer des « composantes » grâce à la fonction `selectInput` dans la partie UI.

```
selectInput("Nage", "Choix de la nage", c("Nage libre", "Dos", "Brasse", "Papillon")),  
selectInput("Sexe", "Choix du sexe", c("Homme", "Femme")),
```

Figure 11 : Code pour créer les menus déroulant pour la nage et le sexe

C'est ce morceau de code ci-dessus qui donne le choix entre les 4 nages et les 2 sexes :

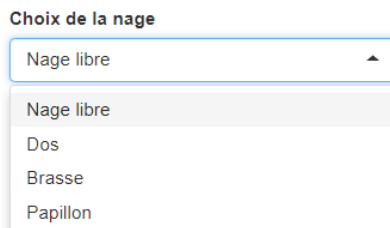


Figure 12 : Menu déroulant pour les nages

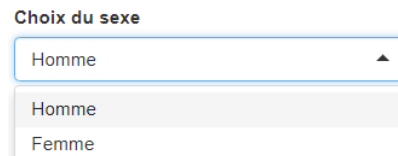


Figure 13 : Menu déroulant pour les sexes

2

Une fois les menus déroulants créés pour discriminer les épreuves, il a fallu créer, dans un deuxième temps, les curseurs de vitesse pour chaque portion⁴. Ils peuvent varier de 1 m.s⁻¹ à 3 m.s⁻¹ et sont prédéfinis comme la moyenne de la vitesse de chaque portion en fonction du sexe et de la nage choisis, ce qui donne, en premier lieu, une idée aux entraîneurs de la vitesse moyenne sur une certaine partie pour une compétition (rappelons que nos données sont uniquement celles des finales du championnat du monde 2019, soit 8 nageurs par épreuve).

Enfin la troisième et dernière partie qu'on retrouve sur l'interface c'est le résultat. Ici un simple calcul :

3

$$temps = \frac{distance}{vitesse}$$

Les valeurs de vitesse choisies par l'utilisateur sont ainsi des composantes que l'on va diviser à la distance de chaque portion pour obtenir le temps sur chacune d'elles. En additionnant le temps des 8 portions, on retrouve le temps final.

⁴ Code pour la création des curseurs en page 34

2.2.2 – Comment répartir sa vitesse ?

2.2.2.1 – Introduction

2.2.2.1.1 – Contexte et objectifs

Après avoir créé un outil permettant de voir le temps final en fonction de la vitesse voulue sur chaque partie de course, nous allons voir l'influence de la vitesse de chaque portion sur le temps final. Autrement dit, le temps final diminue-t-il plus si on gagne 1 m.s^{-1} sur la partie 0-15m ou si on gagne 1 m.s^{-1} sur la partie 50-65 ?

L'objectif principal de cette analyse est d'effectuer un classement des contributions des vitesses sur différentes parties de course, de la plus à la moins importante. Comme pour les écarts, nous allons voir s'il y a des différences et/ou similitudes entre les nages, entre les sexes et entre les parties nagées et non-nagées.

2.2.2.1.2 – Compétence acquise

- Modèle Bayésien

Pour faire ces analyses nous avons créé un modèle bayésien, grâce à la fonction `stan_glm`, avec en variables d'entrée les vitesses sur chaque partie de course, et en variable de sortie le temps final.

```
model_speed <- stan_glm(tps_final ~ M0_15mod + M15_25mod + M25_45mod + M45_50mod + M50_65mod +  
M75_95mod + M95_100mod, data = db)
```

Une fois le modèle exécuté, nous pouvons en extraire les coefficients (ici un exemple pour 0-15m) :

```
posteriors_speed <- insight::get_parameters(model_speed)  
head(posteriors_speed)  
  
## (Intercept) M0_15mod  
## 1 114.6129 -1.479290  
## 2 113.2906 -3.327448  
## 3 112.1820 -2.055504  
## 4 113.8768 -3.289638  
## 5 111.2638 -1.807878  
## 6 114.4329 -2.886745
```

Comme on peut le voir dans cet exemple, l'estimation des paramètres prend une forme de dataframe avec une colonne de plus que le nombre de variable d'entrée. La première correspond à l'intercept et les autres à l'effet des variables d'entrée sur le temps final. Ces colonnes contiennent les distributions à posteriori des paramètres où ces distributions sont approximées à partir d'un échantillon de 4000 tirages pour chaque paramètre. Cette procédure est faite par l'algorithme MCMC (Monte Carlo Markov Chains) qui est un algorithme classique dans le cadre bayésien.

A noter que plus le nombre d'observations sera grand, plus l'affichage du graphique reflètera la réalité, sa courbe sera plus lisse, cependant cela peut demander un peu de temps à l'algorithme.

2.2.2.2 – Analyse

2.2.2.2.1 – Classement des portions

Comme pour l'onglet de l'application Shiny sur la vitesse des nageurs sur différentes parties de course, nous avons ici aussi discriminé les coefficients par sexe et par nage, ce qui nous fait 2 distributions par nage soit 8 distributions au total. Dans cet onglet Shiny, nous avons choisi de mettre au maximum 2 distributions donc 2 variables pour pouvoir les comparer facilement. Voici un exemple :

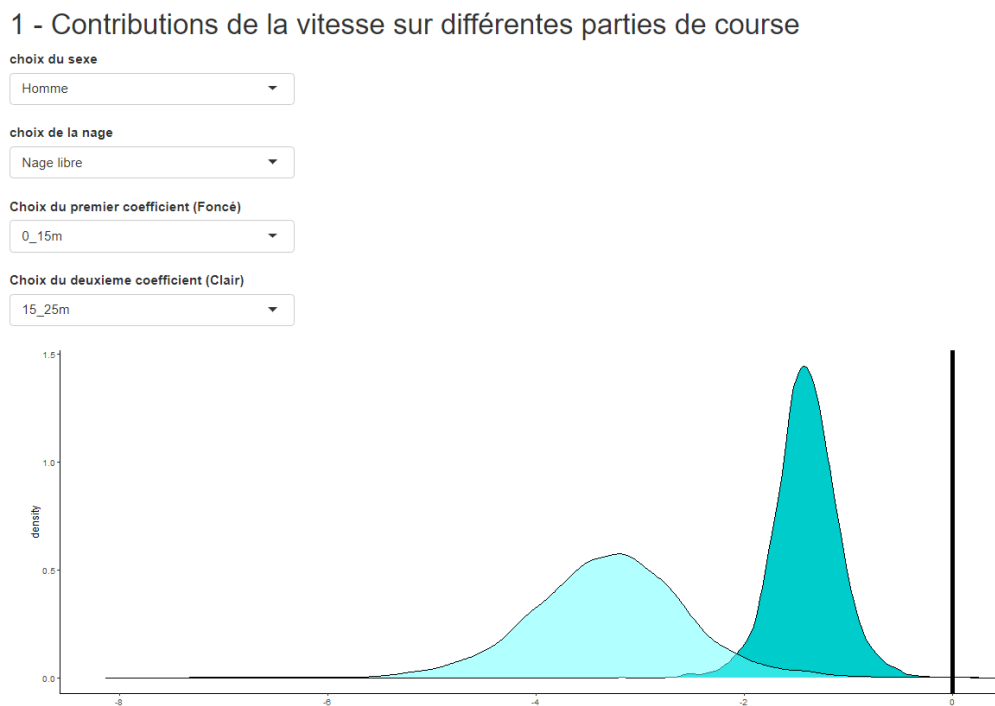


Figure 14 : Rendu de l'analyse des vitesses dans mon application Shiny

Nous pouvons retrouver les mêmes `selectInput` que dans la partie précédente, due à la même discrimination des données. Cependant, deux nouveaux ont été ajoutés, c'est le choix des coefficients.

Pour déterminer quelles sont les portions les plus importantes, j'ai fait un classement de chacune d'elles par nage et par sexe, ainsi il sera facile de déterminer les similitudes et les différences entre les sexes ou les nages.

| | 100m Nage libre | | 100m Dos | | 100m Brasse | | 100m Papillon | |
|----------|-----------------|--------|----------|--------|-------------|--------|---------------|--------|
| Portions | Hommes | Femmes | Hommes | Femmes | Hommes | Femmes | Hommes | Femmes |
| 0-15m | 7 | 7 | 5 | 5 | 6 | 6 | 7 | 7 |
| 15-25m | 4 | 4 | 6 | 6 | 4 | 4 | 5 | 5 |
| 25-45m | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| 45-50m | 6 | 6 | 7 | 7 | 7 | 7 | 6 | 6 |
| 50-65m | 3 | 3 | 4 | 4 | 3 | 3 | 3 | 3 |
| 65-75m | 5 | 5 | 3 | 3 | 5 | 5 | 4 | 4 |
| 75-95m | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 |
| 95-100m | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |

Tableau 1: Classement des portions par nage et par sexe de la contribution de la vitesse sur le temps final




| Portions | classement Hommes | classement Femmes |
|--|-------------------|-------------------|
| 0-15m | 5ème | 5ème |
| 15-25m | 4ème | 4ème |
|  25-45m | 1er | 1er |
| 45-50m | 6ème | 6ème |
| 50-65m | 2ème | 2ème |
| 65-75m | 3ème | 3ème |
|  75-95m | 1er | 1er |
| 95-100m | 7ème | 7ème |

Tableau 2 : Classement des portions selon le sexe


Ce sont les mêmes. Il n'y a aucune différence dans les deux classements ci-dessus, les portions 25-45 et 75-95 sont 1^{er} ex-aequo. La première portion faisant partie de la catégorie « non-nagée » est 2^{ème} (50-65m) dans les deux cas, et c'est pareil pour le reste du classement. On peut le voir dans le Tableau 1 : Pour chaque nage, les colonnes hommes et femmes sont exactement les mêmes. Il n'y a donc aucune différence entre les hommes et les femmes sur ce point-là. Premières au classement, les parties 25-45 et 75-95 seraient celles où il est le plus intéressant d'augmenter sa vitesse, cependant les contributions de vitesse sur ces parties ne sont pas les mêmes pour les hommes ou les femmes, il reste donc à déterminer pour chaque nage à quel point augmenter sa vitesse nous fait gagner du temps, c'est ce que nous verrons dans la partie suivante.

Aucune différence de classement n'est à noter entre les sexes, mais qu'en est-il des classements entre les nages ?




| Portions | classement Nage libre |
|----------|-----------------------|
| 0-15m | 7ème |
| 15-25m | 4ème |
| 25-45m | 1er |
| 45-50m | 6ème |
| 50-65m | 3ème |
| 65-75m | 5ème |
| 75-95m | 2ème |
| 95-100m | 8ème |

Tableau 4 : Classement des portions en nage libre




| Portions | classement Dos |
|----------|----------------|
| 0-15m | 5ème |
| 15-25m | 6ème |
| 25-45m | 1er |
| 45-50m | 7ème |
| 50-65m | 4ème |
| 65-75m | 3ème |
| 75-95m | 2ème |
| 95-100m | 8ème |

Tableau 3 : Classement des portions en dos



| Portions | classement Brasse |
|----------|-------------------|
| 0-15m | 6ème |
| 15-25m | 4ème |
| 25-45m | 2ème |
| 45-50m | 7ème |
| 50-65m | 3ème |
| 65-75m | 5ème |
| 75-95m | 1er |
| 95-100m | 8ème |

Tableau 5 : Classement des portions en brasse



| Portions | classement Papillon |
|----------|---------------------|
| 0-15m | 7ème |
| 15-25m | 5ème |
| 25-45m | 2ème |
| 45-50m | 6ème |
| 50-65m | 3ème |
| 65-75m | 4ème |
| 75-95m | 1er |
| 95-100m | 8ème |

Tableau 6 : Classement des portions en papillon

En 8^{ème} position, on retrouve à chaque fois la dernière partie de course : le 95-100m. La distance étant très courte, améliorer sa vitesse ici n'a pas trop de sens, comme nous l'avons vu dans la partie sur les écarts, les athlètes doivent arriver sur une fin de coup de bras, augmenter leur vitesse ne les aiderait pas forcément, le but étant de se fier à leur repère déterminé à l'entraînement. C'est aussi le cas pour la portion 45-50m qui reste 6^{ème} ou 7^{ème} sur chaque course.

De manière générale, on voit que toutes les portions de la partie non-nagée sont dernières dans le classement, sauf la deuxième coulée (50-65m). Sur cette dernière, les nageurs prennent appui sur le mur et ne partent pas en plongeon comme au départ, en conséquence ils parcourent moins de distance sous l'eau et nagent sur une plus grande partie des 15m que sur la première coulée. Cela rejoint l'idée que les parties nagées sont dans le haut du classement, en effet on a vu que très peu d'écarts se créaient dans ces parties ce qui voulait dire que les nageurs allaient tous à la même vitesse. Par conséquent, si un nageur augmente ou diminue sa vitesse sur une de ces parties, il va respectivement gagner ou perdre beaucoup de temps par rapport aux autres.

On remarque également qu'il y a très peu de différence de classement entre les nages, à part pour le dos où la portion 0-15m (le départ) est un peu mieux classée que pour les autres. Cela

peut s'expliquer par la manière dont commence cette course, contrairement aux trois autres où les athlètes s'élancent du plongeoir, les nageurs de dos partent directement de l'eau et leur temps de vol est proche de 0, ils nagent en conséquence plus que les athlètes des autres courses sur cette partie.

Je tiens à préciser que ces résultats sont à nuancer, les portions qui contribuent le plus sont les plus longues (celles de 20m) alors que les dernières sont globalement les plus courtes (5 à 10m).

2.2.2.2.2 – Contribution des vitesses

Le but de cette partie n'est pas seulement de d'effectuer un classement par nage et par sexe des portions où la vitesse contribue le plus sur le temps final, elle sert également à voir à quel point ces contributions permettent aux nageurs de réduire leur temps final, donc d'améliorer leur performance. Pour ce faire, les entraîneurs et conseillers techniques disposent de l'onglet correspondant dans l'application Shiny que nous avons créée.

Par exemple, en reprenant le classement des portions en brasse, on a vu que la portion où la vitesse contribuait le plus au temps final était le 75-95m, et la deuxième le 25-45m. Mais à quel point ces contributions sont-elles différentes ?

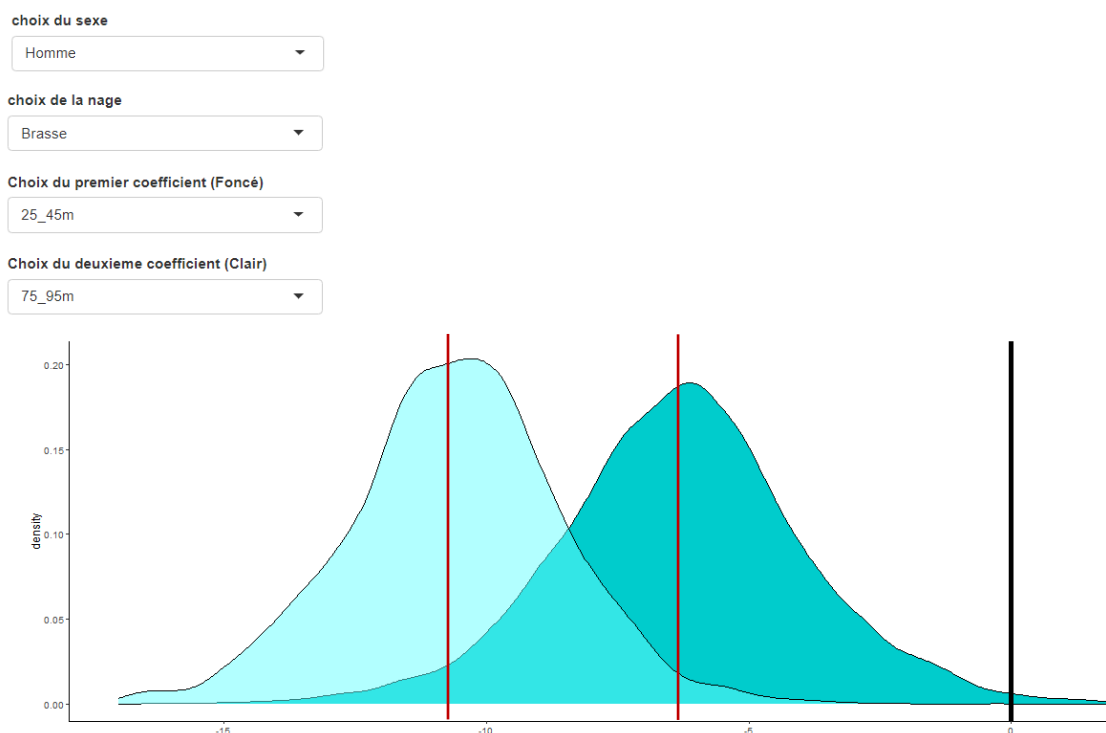


Figure 15 : Exemple de contribution des vitesses pour la brasse

Ce graphique apporte une nouvelle information : si un nageur décide d'aller 1 m.s^{-1} plus vite sur la partie 75-95m, il gagnera plus de 10 secondes en moyenne sur son temps final, alors que s'il décide d'aller 1 m.s^{-1} plus vite sur la partie 25-45, il gagnera en moyenne 6.5 secondes sur son temps final. Pour la brasse chez les hommes, la portion 75-95m contribue 69% de plus que la portion 25-45. En faisant cela pour tous les coefficients de toutes les nages, les

entraîneurs pourront déterminer le temps idéal qu'il faudrait passer à l'entraînement pour chacune des portions.

Si on garde l'exemple de la brasse chez les hommes, ça donne le tableau suivant :

| Portions | Contributions Brasse | Pour 1h d'entraînement |
|----------|----------------------|------------------------|
| 0-15m | 0,32 | 4,84 |
| 15-25m | 0,65 | 9,82 |
| 25-45m | 0,69 | 10,43 |
| 45-50m | 0,18 | 2,72 |
| 50-65m | 0,68 | 10,28 |
| 65-75m | 0,43 | 6,50 |
| 75-95m | 1,00 | 15,11 |
| 95-100m | 0,02 | 0,30 |

Figure 16 : Exemple du temps d'entraînement idéal pour la brasse chez les hommes

Dans le tableau ci-dessus, j'ai mis la contribution de la vitesse de la portion 75-95 en valeur seuil car c'est la plus élevée. Ainsi j'ai pu déterminer quel était le pourcentage de contribution de vitesse des autres portions, puis remettre ces résultats sous forme de minutes dans le cadre d'un entraînement d'une heure.

Dans ce cadre-là, on peut voir que le travail de la vitesse sur la partie 75-95 devrait représenter 15.11 minutes (15 minutes et 7 secondes) de l'entraînement soit plus de 25% de l'entraînement ! En revanche la dernière portion (95-100) ne devrait représenter que 0.3 minute d'un entraînement d'une heure soit 18 secondes...

Conclusion :

Sur cette partie d'analyse de la vitesse regroupant le modèle déterministe et le modèle bayésien, j'ai pu acquérir deux nouvelles compétences : la première pour R-Shiny, j'avais très peu utilisé Shiny pendant les deux ans de DUT, cela m'a donc permis d'en apprendre beaucoup plus sur cette fonctionnalité et d'avoir un certain niveau dans ce domaine. J'ai également pu utiliser les statistiques bayésiennes dans la création du modèle dans la deuxième partie.

Le premier outil créé pour les membres de la FFN est plus un outil pratique que statistique, cependant il n'en sera pas moins utile. Quant au deuxième onglet sur la contribution des vitesses, je pense qu'il sera important d'accorder plus de temps d'entraînement à certaines parties de course tout en gardant une vision professionnelle de la chose. En effet, bien que la vitesse sur la dernière partie de course en brasse chez les hommes ne devrait représenter que 18 secondes d'un entraînement d'une heure, je pense qu'il est nécessaire d'accorder plus de temps notamment pour que les nageurs arrivent au niveau de leur repère sur une fin de cycle.

Problèmes rencontrés :

Cette deuxième partie a été riche en problèmes, notamment car c'est ici que nous avons mis en place l'application Shiny. Au tout début nous avons combiné les deux analyses de cette partie, l'onglet avec les curseurs de vitesse était le même, seulement le modèle qui était à l'intérieur était le modèle bayésien sur les contributions des vitesses. L'objectif de cet onglet était non plus d'additionner le temps de chaque partie (grâce à la vitesse choisie) pour en trouver le temps final, mais de voir à que point le temps final variait en fonction de l'augmentation d'un curseur ou d'un autre selon les mêmes variations. Ce graphique était trompeur, c'est pourquoi on a divisé cette partie en deux analyses, avec un outil centré uniquement sur la vitesse, et un autre sur les contributions.

En créant l'outil du temps final en fonction des vitesses de chaque portion sur R Shiny, j'ai également rencontré quelques problèmes. En arrivant sur l'onglet, l'utilisateur a le choix entre la nage qu'il veut et le sexe qu'il veut mais dans la base de données, aucune variable ne comprend ces deux modalités ensemble, il a fallu créer un objet les comprenant simultanément :

```
epreuve = case_when(paste(input$Nage, input$Sexe) == "Nage libre Homme" ~ "100 NL H",
  paste(input$Nage, input$Sexe) == "Nage libre Femme" ~ "100 NL F",
  paste(input$Nage, input$Sexe) == "Dos Homme" ~ "100 DOS H",
  paste(input$Nage, input$Sexe) == "Dos Femme" ~ "100 DOS F",
  paste(input$Nage, input$Sexe) == "Brasse Homme" ~ "100 BRA H",
  paste(input$Nage, input$Sexe) == "Brasse Femme" ~ "100 BRA F",
  paste(input$Nage, input$Sexe) == "Papillon Homme" ~ "100 PAP H",
  paste(input$Nage, input$Sexe) == "Papillon Femme" ~ "100 PAP F")
```

Après avoir créé cet objet grâce à la fonction `case_when`, j'ai créé tous les sliders (`sliderInput`) en mettant leur valeur initiale à 2 m.s⁻¹. Pour qu'à chaque choix de nage et de sexe de l'utilisateur, les valeurs initiales des curseurs soient égales à la moyenne de vitesse de la portion du choix en question, j'ai dû créer un dataframe des 8 (nombre de portions) x 2 (nombre de sexe) x 4 (nombre de nage) moyennes, soit 64 lignes. De base la création des sliders se fait dans la partie UI, en rentrant les valeurs initiales à la main, mais dans notre cas on les a créés dans la partie server (pour pouvoir affecter les valeurs voulues), pour ensuite les transférer dans la partie UI.

Partie Serveur :

```
output$slider1 <- renderUI({
  epreuve = case_when(paste(input$Nage, input$Sexe) == "Nage libre Homme" ~ "100 NL H",
    paste(input$Nage, input$Sexe) == "Nage libre Femme" ~ "100 NL F",
    paste(input$Nage, input$Sexe) == "Dos Homme" ~ "100 DOS H",
    paste(input$Nage, input$Sexe) == "Dos Femme" ~ "100 DOS F",
    paste(input$Nage, input$Sexe) == "Brasse Homme" ~ "100 BRA H",
    paste(input$Nage, input$Sexe) == "Brasse Femme" ~ "100 BRA F",
    paste(input$Nage, input$Sexe) == "Papillon Homme" ~ "100 PAP H",
    paste(input$Nage, input$Sexe) == "Papillon Femme" ~ "100 PAP F")

  initslider = db_moy1 %>% filter(Epreuve == epreuve, portions == "M0_15mod") %>% pull(mean)

  sliderInput("P1",
    "vitesse sur 0_15:",
    min = 1,
    max = 3,
    value = initslider,
    step = 0.01)})
```

Partie UI :

```
sidebarPanel(
  uiOutput("slider1"),
```

Pour la deuxième analyse, celle des coefficients, aucun réel problème ne s'est présenté.

2.3 – BONUS : LES HEATMAPS

2.3.1 – Introduction

Pour cette dernière partie, qui serait plutôt un bonus, nous avons repris l'idée des heatmaps que mes encadrants Arthur LEROY et Robin PLA avaient déjà utilisées lors de leur article qui traite de l'impact de la morphologie sur la performance en natation de haut-niveau. Ces outils permettent, grâce à un dégradé de couleur, d'attribuer à une relation entre deux variables (ici la fréquence de coups de bras et la distance par cycle) une valeur propre (dans notre cas la vitesse).

Les objectifs étaient donc de faire une data visualisation claire et concise de la relation entre la fréquence de coups de bras et la distance par cycle, de rendre cette data visualisation interactive et d'établir des hypothèses sur des manières de nager idéales en fonction des épreuves. Faut-il avoir une grosse fréquence de coups de bras mais une petite distance par cycle, est ce que c'est totalement l'inverse ou y'a-t-il un juste milieu entre ces deux variables ?

Pour répondre à ces questions, j'ai développé de nouvelles capacités notamment pour la création des heatmaps (grâce aux fonctions `geom_raster` et `scale_fill_gradient`) et leur utilisation.

Pour rendre le graph encore plus intéressant pour les entraîneurs, j'ai eu recours au package `plotly` :



Il permet de réaliser des graphiques en 2D ou en 3D, de tous types (histogramme, nuage de points, etc...). Ce qui est pratique avec ce package, c'est que tous les graphiques réalisés avec `ggplot2` peuvent être rendus interactifs en ajoutant la fonction « `ggplotly` ». En passant la souris sur un graphique, elle devient un curseur qui indique les valeurs prises en ordonnées et en abscisse sur un point précis. Mises en forme dans l'application Shiny, les fonctionnalités de `ggplotly` pourront être utilisées tout en discriminant les données par nage, par sexe, et par quart de course⁵, ainsi les entraîneurs auront à leur disposition un véritable outil statistique.

Au niveau des compétences sollicitées, j'ai trouvé les équations de droites correspondant aux vitesses (selon la relation fréquence ~ Distance par cycle) et les ai intégrées à un graphique `ggplot` grâce à la fonction `geom_line`.

⁵ Ici on ne prend plus 8 portions mais 4 : 0-25m ; 25-50m ; 50-75m ; 75-100m

2.3.2 – Analyses

2.3.2.1 – Création des Heatmaps

Avec la relation : fréquence de coups de bras par minute ~ distance par cycle, trouver la vitesse revient juste à multiplier ces deux variables, puis de diviser par 60 pour revenir en seconde. Soit l'équation :

$$Vitesse = \frac{(Fréquence * Distance \text{ par cycle})}{60}$$

En appliquant en gradient de couleur à toutes ces vitesses, on obtient le graphique suivant :

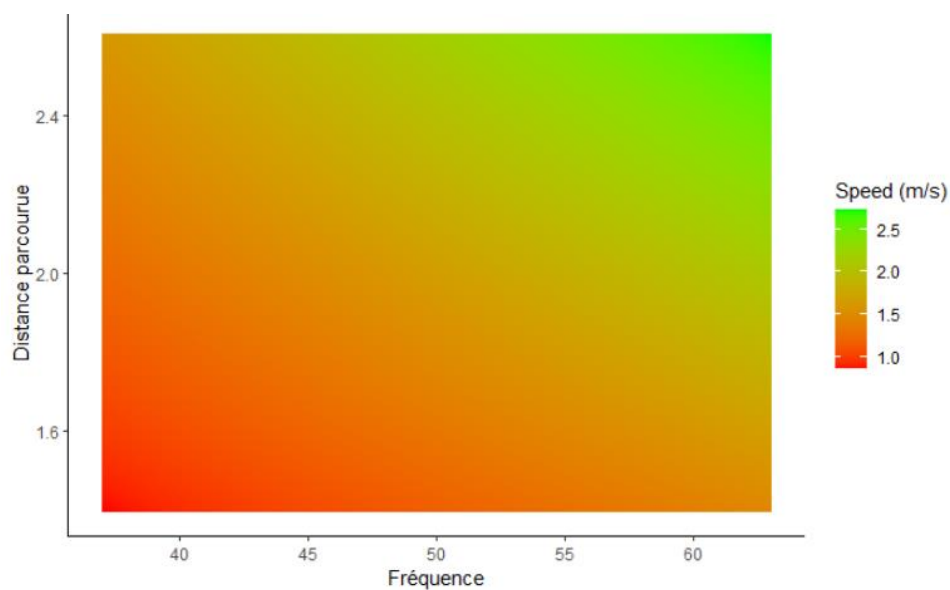


Figure 17 : Création Heat map

Ajoutons maintenant les point de chaque nageurs en discriminant sur une portion de course (ici le premier quart : 0-25m):

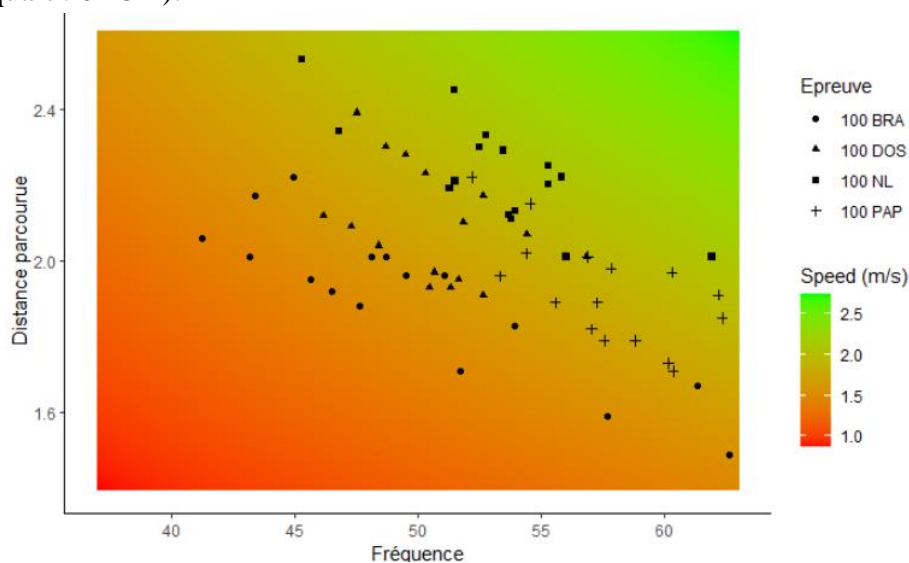


Figure 18 : Rajout des données

Même en discriminant les données par portion et en changeant la forme des points selon l'épreuve, le graphique reste trop chargé, c'est pourquoi nous l'avons intégré à l'application Shiny, ainsi nous aurons pour chaque graphique 8 points correspondant aux 8 nageurs de l'épreuve choisies. Prenons l'exemple du 100m nage libre chez les hommes :

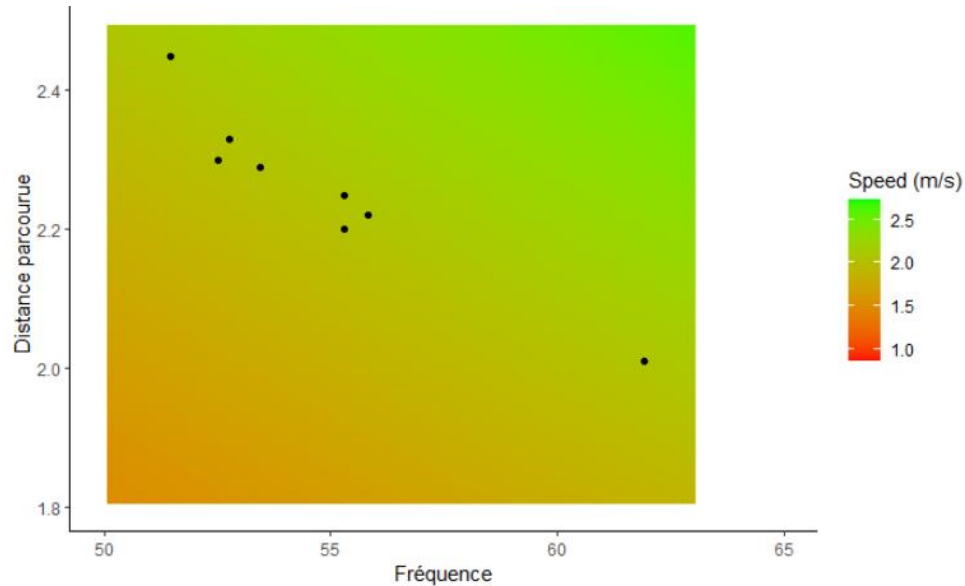


Figure 19 : filtre des données pour rendre le graphique plus lisible

En ajustant l'échelle, on se retrouve avec un gradient de couleur beaucoup moins contrasté, car c'est en fait un zoom du précédent graphique sur la partie souhaitée. A ce graph nous avons rajouté 4 courbes représentant les vitesses (2.25m.s^{-1} ; 2.0m.s^{-1} ; 1.75m.s^{-1} ; 1.5m.s^{-1}), pour faciliter la compréhension et pour discerner les méthodes de nage en fonction des épreuves. Sur nos graphique, l'échelle est très réduite et on pense voir des droites, mais ce sont bien des relations hyperboliques de type $\frac{1}{x}$.

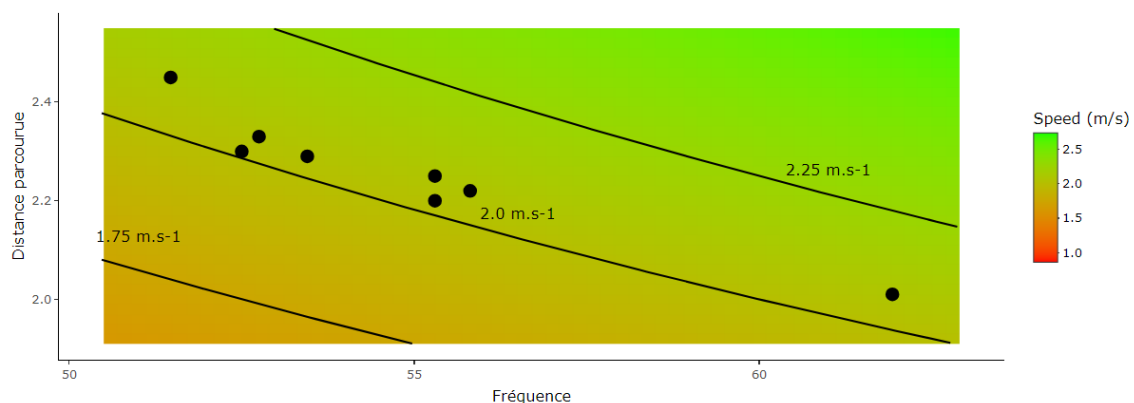


Figure 20 : Ajout des courbes de vitesse

Enfin, en ajoutant la fonction ggplotly, on peut avoir la vitesse pour chaque coordonnée voulue, simplement en pointant le curseur sur ces points.

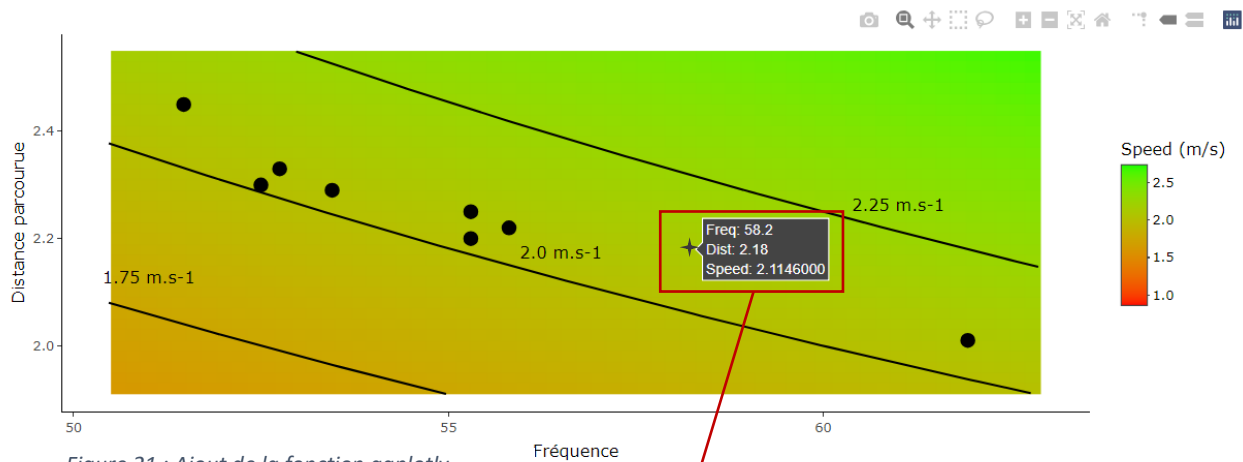


Figure 21 : Ajout de la fonction ggplotly

Exemple de ce que permet de faire la fonction ggplotly sur ce graphique. On retrouve un curseur avec la fréquence associée et la distance parcourue associée ainsi que la vitesse, résultat de l'opération entre ces deux variables

Maintenant que nous avons le bon type de graphique, il faut le mettre dans l'application Shiny afin de rendre son utilisation plus facile. En ajoutant un choix de la nage, un choix de la course et un choix de la partie voulue (1^{er} quart ; 2^{ème} quart ; 3^{ème} quart ; 4^{ème} quart), les entraîneurs et conseillers techniques pourront jongler entre différentes épreuves pour voir si des méthodes de courses idéales se dessinent.

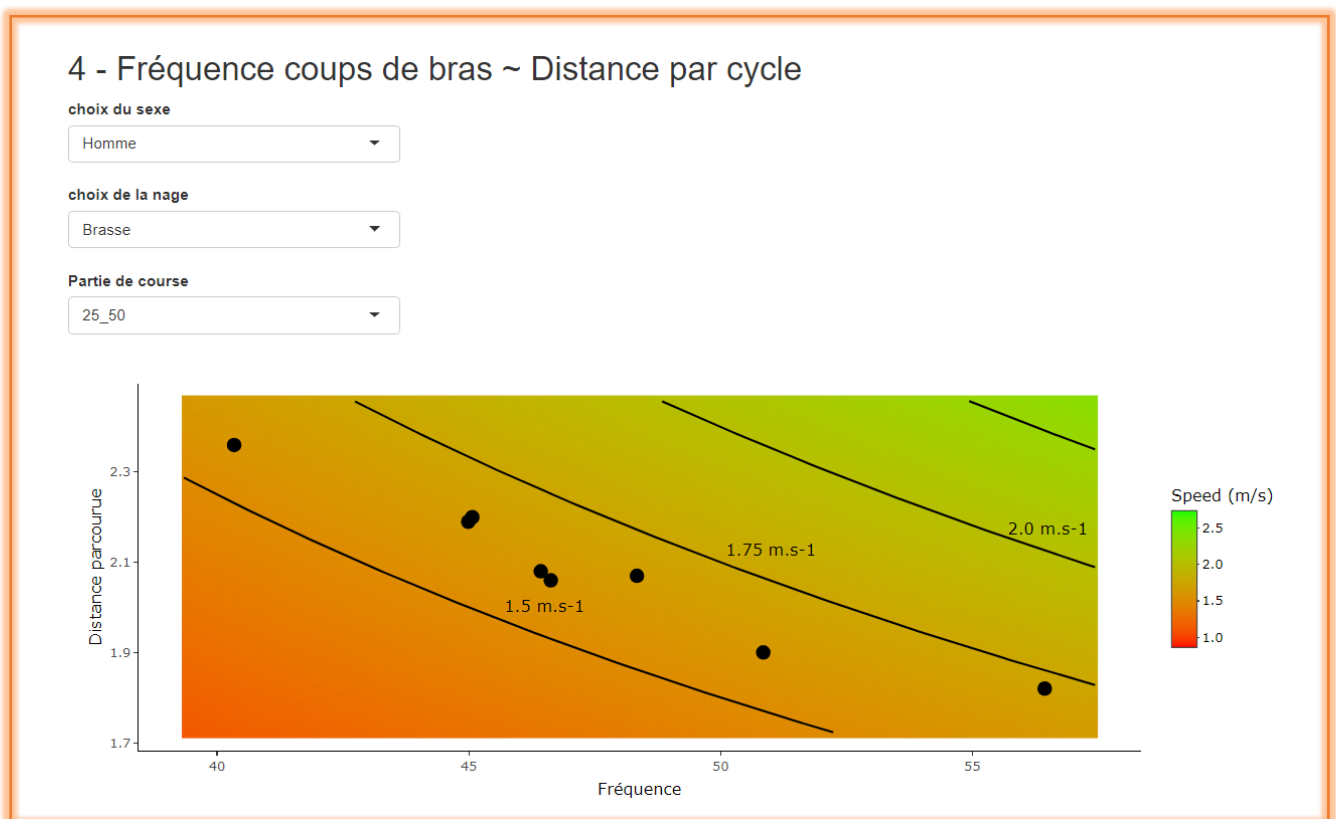


Figure 22 : Rendu des Heatmaps dans mon application Shiny

2.3.2.2 – Résultats

La découverte la plus intéressante que j'ai trouvée se passe au niveau de la brasse (chez les hommes ou les femmes), en analysant les heatmaps j'ai vu qu'une méthode de nage semblait être plus rapide que toutes les autres. Pour le voir, prenons l'exemple du premier quart de course chez les femmes :

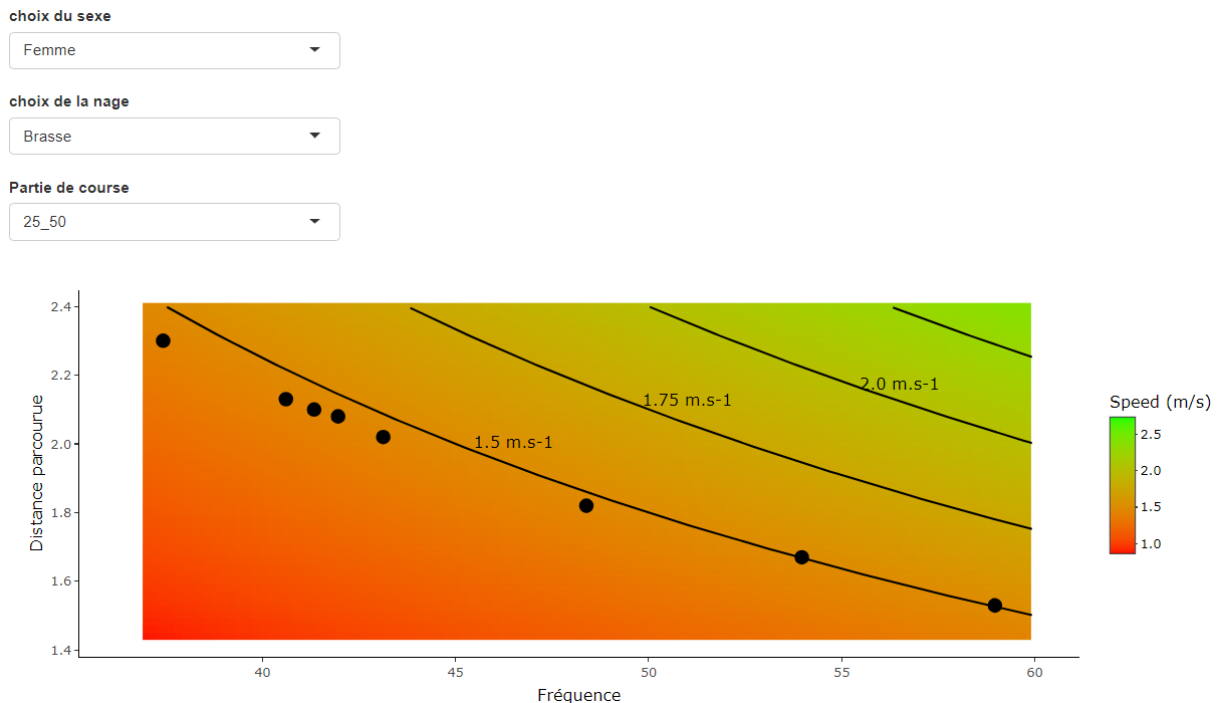


Figure 23 : Exemple sur le 2ème quart de course chez les femmes en brasse

Sur cet exemple on peut voir, grâce à la courbe de vitesse $1.5\text{m}\cdot\text{s}^{-1}$, une nette amélioration de la vitesse des nageuses en fonction de l'augmentation de leur fréquence de coups de bras. En effet, les deux plus rapides sur cette partie sont celles qui possèdent des distances parcourues par coups de bras plutôt faibles mais des fréquences de nage rapides (53.97 et 58.97 coups de bras par minutes), au contraire les plus lentes ont des distances parcourues plus élevées mais des fréquences très faibles. C'est la même chose sur les trois autres quarts de course, encore plus sur le 0-25m :

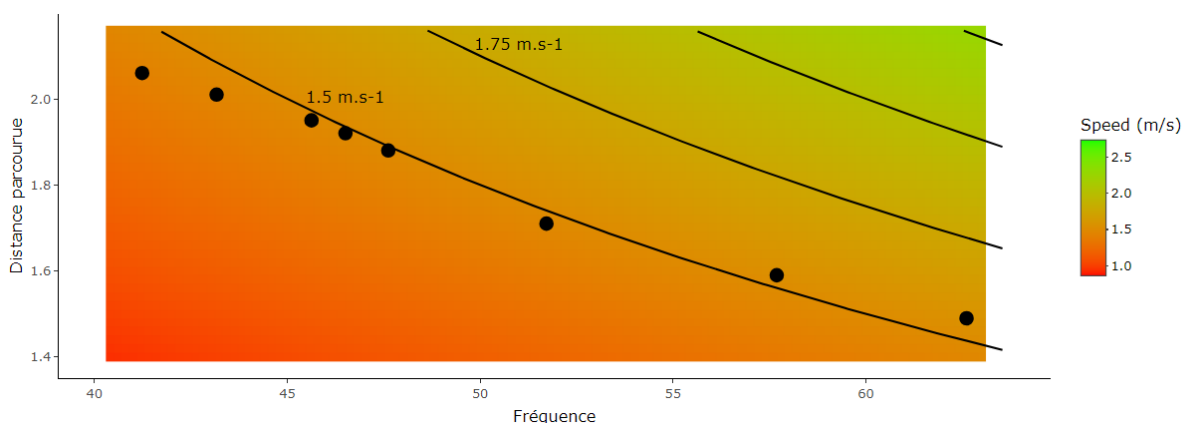


Figure 24 : Exemple brasse femmes 0-25m

Ici, on voit encore mieux à quel point il est important d'avoir une grosse fréquence et pas forcément une grande distance parcourue par cycle de bras.

Pour la nage libre et le dos, il n'y a pas vraiment de méthodes de nage qui se sont dessinées, les nageurs et nageuses ont chacun leur préférence pour ce qui est de la relation fréquence ~ distance parcourue⁶. La seule chose à noter pour la nage libre c'est que plus la course avance plus la différence de vitesse entre les nageurs est importante.

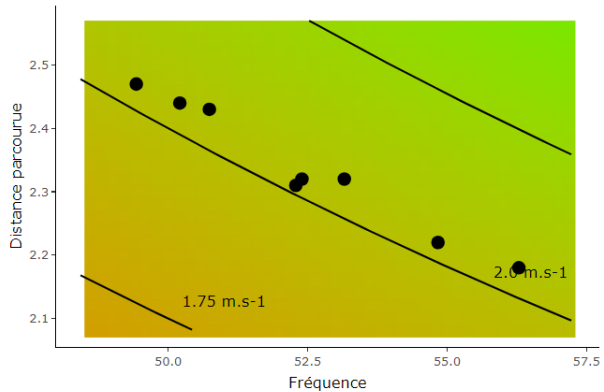


Figure 26 : Nage libre Homme 25-50m

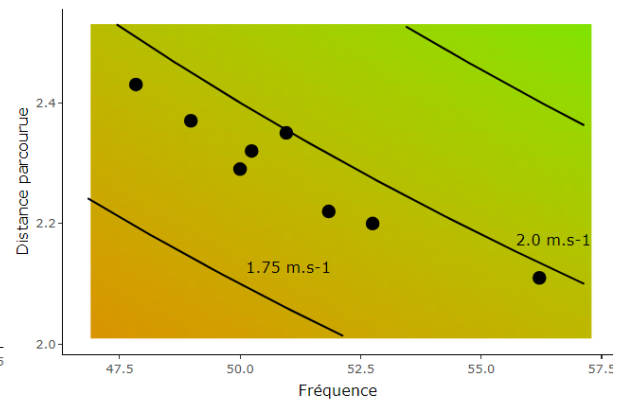


Figure 25 : Nage libre Homme 50-75m

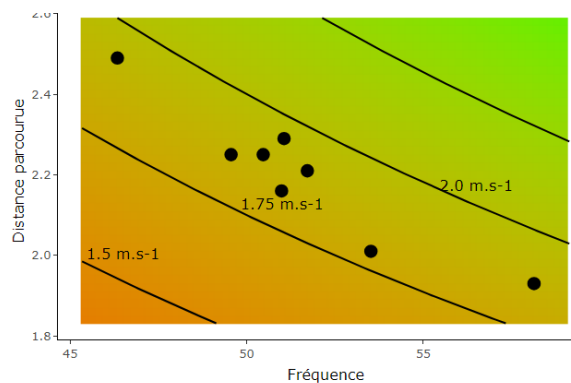


Figure 27 : Nage libre Homme 75-100m

Rien qu'en regardant la couleur des heatmaps, on peut voir que le gradient de couleur s'élargit en fonction de l'avancée de la course, pour le 25-50m le vert est la couleur prédominante et les points sont à peu près tous sur une même ligne de vitesse (juste au-dessus de 2.0 m.s^{-1}), pour le quart suivant on retrouve une teinte orangée, les points prennent la moitié de l'espace entre deux courbes de vitesse. Cette teinte orangée va venir s'assombrir dans le dernier quart, et les points s'écartent, on retrouve un écart de vitesse de presque 0.25 m.s^{-1} entre le plus et le moins rapide.

A défaut de retrouver un style de nage plus efficace comme en brasse, on voit qu'au fur et à mesure que la course avance, les différences de vitesse entre les nageurs sont importantes, cela rejoint la première partie des analyses sur les écarts de course. Les nageurs ont une méthode personnelle de course en fonction de leurs caractéristiques et/ou préférences

⁶Rappel distance parcourue : pour la nage libre et le dos, c'est la distance parcourue en deux coups de bras.

Enfin le dernier point de cette analyse traite du papillon, y-a-t-il une méthode de nage plus intéressante qu'une autre, est-ce qu'elle se voit ou est ce qu'il n'y a aucune nouvelle information ?

En analysant chaque quart de course du papillon chez les hommes et chez les femmes, j'ai émis une hypothèse quant à la façon de nager le deuxième 50m. Voici les graphiques des deuxièmes 50m chez les hommes et chez les femmes :

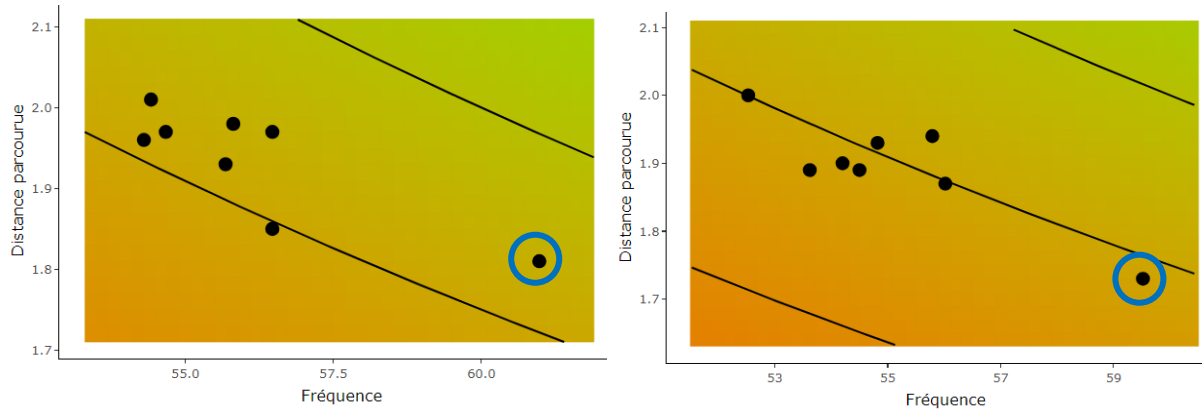


Figure 28: Papillon deuxième 50-75m et 75-100m homme

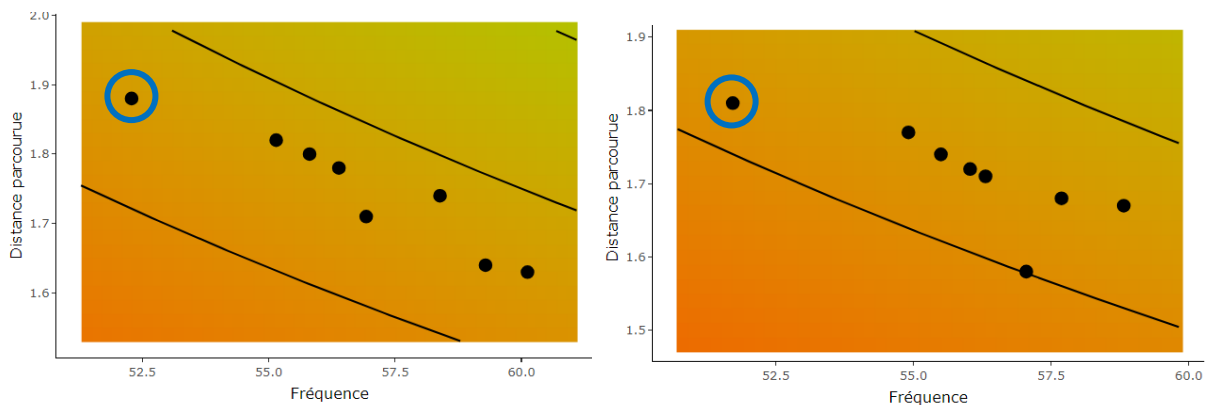


Figure 29 : Papillon deuxième 50-75m et 75-100m femme

Sur les 4 graphiques on retrouve à chaque fois un point qui se sépare du reste. Chez les hommes il est à droite du graphique, il a donc une grosse fréquence et une petite distance par cycle, chez les femmes c'est l'inverse. Le point est à gauche du graphique donc la nageuse a une petite fréquence pour une plus grande distance parcourue. Comme il n'y a à chaque fois qu'un seul point, je suis allé voir quel était le nageur ou la nageuse qui correspondait et leur classement.

Chez les hommes, avec une fréquence de 60.98 coups de bras/minutes (pour le 50-75m) et 59.92 (pour le 75-100m), le nageur James GUY a fini la course à la 7^{ème} place soit avant-dernier.

Chez les femmes, avec une fréquence de 52.29 coups de bras/minutes (pour le 50-75m) et 51.72 (pour le 75-100m), la nageuse Marie WATTEL a fini la course en dernière position.

Au vu de ces deux résultats, on pourrait penser que chez les hommes il serait préférable d'avoir une plus petite fréquence de nage mais une distance par cycle plus longue (l'inverse de James

GUY), et que chez les femmes ce serait l'inverse. Cependant, je ne pense pas que les méthodes de nage entre les hommes et les femmes sur une même épreuve peuvent être diamétralement opposées. Avec la brasse on a vu que la vitesse augmentait en mettant plus de fréquence de coups de bras que de distance par cycle, et ce chez les femmes et chez les hommes. Même si les performances ne sont pas les mêmes, les hommes et les femmes se ressemblent au niveau des caractéristiques. Par conséquent je pense que si ces deux nageurs se retrouvent dans le bas du classement c'est parce que leur ratio fréquence-distance parcourue est trop extrême, il n'est pas assez proche des autres nageurs. Plus généralement, il doit y avoir un certain seuil d'équivalence entre la fréquence et la distance qu'il ne faut pas dépasser, au risque de perdre du temps.

Conclusion :

Cet outil, en plus d'être pratique et facile d'utilisation pour les entraîneurs a été un véritable travail statistique. De la création d'une heatmap vierge à la rendre interactive dans un nouvel onglet de mon application, j'ai acquis et développé de nombreuses compétences, principalement en codage R et R-Shiny. Les résultats quant à eux n'en sont pas moins intéressants, avec notamment une piste très sérieuse sur la méthode de nage idéale pour la brasse. On sait qu'en augmentant la fréquence de nage, la vitesse a tendance à augmenter, mais y'a-t-il un seuil à ne pas franchir pour ce ratio fréquence-distance ? Est-ce qu'avoir une fréquence extrêmement élevée mais une distance parcourue extrêmement faible permet d'être très rapide ou est-ce que cette tendance change de sens à une certaine valeur ?

Problèmes rencontrés :

Comme je n'avais jamais créé de heatmap avant, j'ai mis du temps avant d'être à l'aise avec cet objet. Après avoir compris le principe j'ai commencé par intégrer le code dans l'application Shiny avec les deux menus habituels : choix de la nage et choix du sexe. Pour cet onglet il fallait rajouter un menu déroulant pour la fréquence des quatre quarts de course et la distance par cycle. Pour ne laisser qu'un choix à l'utilisateur (0-25 ;25-50 ;50-75 ;75-100) et non un choix pour la fréquence et un pour la distance parcourue, il a fallu que je reprenne la fonction `case_when`.

```
portion = tibble(freq = case_when(input$Portion == "0_25" ~ db %>% filter(newEpreuve == paste(input$Nage3, input$Sexe3)) %>% pull(FREQ25),
input$Portion == "25_50" ~ db %>% filter(newEpreuve == paste(input$Nage3, input$Sexe3)) %>% pull(FREQ50),
input$Portion == "50_75" ~ db %>% filter(newEpreuve == paste(input$Nage3, input$Sexe3)) %>% pull(FREQ75),
input$Portion == "75_100" ~ db %>% filter(newEpreuve == paste(input$Nage3, input$Sexe3)) %>% pull(FREQ100)),
dc = case_when(input$Portion == "0_25" ~ db %>% filter(newEpreuve == paste(input$Nage3, input$Sexe3)) %>% pull(DC25),
input$Portion == "25_50" ~ db %>% filter(newEpreuve == paste(input$Nage3, input$Sexe3)) %>% pull(DC50),
input$Portion == "50_75" ~ db %>% filter(newEpreuve == paste(input$Nage3, input$Sexe3)) %>% pull(DC75),
input$Portion == "75_100" ~ db %>% filter(newEpreuve == paste(input$Nage3, input$Sexe3)) %>% pull(DC100)))
```

Le deuxième problème rencontré était pour l'affichage des droites de vitesse sur le graphique. Avec l'équation (2), trouver une équation pour une certaine vitesse n'était pas compliqué, il suffisait de multiplier la vitesse par 60 : $freq * dist = 60 * vitesse$

Puis pour chaque vitesse voulue, choisir une abscisse et trouver l'ordonnée correspondante.

Exemple pour 1.5 m.s^{-1} :

$$freq * dist = 60 * 1.5$$

$$freq * dist = 90$$

$$freq = \frac{90}{dist}$$

Pour $dist = 2$,

$$freq = 45$$

Trouver toutes les équations des courbes de vitesse était facile, mais les afficher sur R était beaucoup plus compliqué. Dans ce cas-là je commençais par chercher sur internet s'il y avait la réponse à ma question sur des sites comme stackoverflow, je pouvais également chercher sur les cheatsheets, ce sont des résumés d'un package ou d'une fonction et sont directement sur R dans l'onglet « Help ». Et si je ne trouvais malheureusement pas la réponse, je pouvais demander à mon maître de stage Arthur LEROY qui, grâce à son expérience, avait la réponse à la quasi-totalité de mes questions.

PARTIE 3 : CONCLUSION

3.1 – CONCLUSION DE L'ETUDE

Le but de cette étude, et donc de ce stage était de déterminer les indicateurs de performance en natation de haut niveau, pour cela nous disposions d'une base de données de 64 nageurs des finales du championnats du monde de 2019 à Gwangju, en Corée du sud.

Les objectifs définis étaient bien évidemment de déterminer ces indicateurs de performance, mais également de construire des outils pratiques, compréhensibles et simple d'utilisation pour les membres de la FFN.

Pour répondre à ces objectifs, nous avons analysé les données en trois parties.

La première sur les écarts de course permet de savoir quelles sont les parties de course où les écarts entre les nageurs se créent, donc à quel moment le choix des positions finales se détermine. Les entraîneurs bénéficieront d'un tableau croisant les 4 nages avec les 8 portions de course de la base, qui sera colorié avec un gradient de couleur qui correspond aux valeurs des écarts créés. Pour la deuxième partie, nous avons choisi de traiter la vitesse pure et sa contribution (qui sont sans doute les indicateurs de performance les plus importants).

Dans un premier temps en créant un modèle linéaire classique pour en fabriquer un outil purement déterministe, choisir la vitesse de nage sur chaque portion et avoir le temps final en arrivée, puis dans un deuxième temps un modèle bayésien un peu plus poussé qui apporte la contribution de chaque vitesse sur le temps final. Après avoir fait un classement des contributions, j'ai déterminé le temps d'entraînement idéal qu'il faudrait accorder à la vitesse sur chaque portions (voir tableaux en annexe).

Enfin dans la dernière partie, nous avons réalisé des heatmaps, ce sont des cartes de couleur, dans notre cas de la vitesse, où sont représentés les nageurs. Cela nous a permis de définir des styles de nage bien précis (notamment en brasse) et de trouver qu'il existe une valeur « seuil » du ratio fréquence – distance (en tout cas pour certaines courses) à ne pas dépasser.

Toutes ces analyses sont intégrées à une application Shiny qui les rend interactifs, les entraîneurs pourront l'utiliser en choisissant les nages qu'ils veulent, les portions qu'ils veulent ou le sexe qu'ils veulent.

3.2 – BILAN PERSONNEL ET PROFESSIONNEL

Durant mon stage la première chose que j'ai apprise, c'est que dans le domaine de la recherche, une grosse partie du travail que l'on fait n'aboutira pas à des résultats concluants. A la différence des cours, où pour chaque travail nous avons une ligne directrice pour trouver les résultats attendus, la recherche elle, sert en fait à trouver cette ligne directrice, une problématique.

D'une certaine façon, passer de l'état « appliquer ce que je sais faire » à l'état « qu'est-ce que je veux savoir » a été le plus gros blocage que j'ai rencontré durant ce stage.

Malgré ces quelques blocages j'ai vraiment apprécié faire ce stage. Il m'a appris à voir les choses de différentes manières, aussi bien sur le plan personnel que professionnel, et m'a forcé à passer un nouveau cap, en terme de réflexion (pour les analyses) mais aussi en terme de bagage technique. En travaillant sous R pendant ces 7 semaines, j'ai développé de nouvelles compétences aussi bien en statistique (création des modèles, R Shiny, etc...) qu'en mathématiques (fonctions, algèbre, etc...).

Même si mon projet professionnel n'était pas d'intégrer le domaine de la recherche avant de commencer le stage, ce dernier m'a conforté dans l'idée de devenir un statisticien. J'aime vraiment apporter mes connaissances et mes compétences dans le but de faire découvrir des choses, éclaircir des idées ou créer des outils utiles dans la vie de tous les jours.

3.3 – ANNEXES

Fonctions imbriquées pour les écarts de course :

```
fecarts <- fonction(SexeNage, partiechar, premiernageur, derniernageur){
  a = gather(SexeNage, key = "portions", value = "tps_intermediaire", M0_15mod, M15_25mod, M25_45mod,
M45_50mod, M50_65mod, M65_75mod, M75_95mod, M95_100mod) %>% filter(portions == partiechar)
  for (i in a[premiernageur:derniernageur, 1]){
    return(mean(abs(a$tps_intermediaire - mean(a$tps_intermediaire))))
  }
}

reph <- fonction(Nage, partiechar, premiernageur, derniernageur){
  return(rep(fecarts(Nage, partiechar, premiernageur, derniernageur)/fecarts(HNL, "M0_15mod", 1, 8), 8))
}

repf <- fonction(Nage, partiechar, premiernageur, derniernageur){
  return(rep(fecarts(Nage, partiechar, premiernageur, derniernageur)/fecarts(FNL, "M0_15mod", 9, 16), 8))
}
```

Code pour le tableau des écarts chez les hommes :

```
Ecarts_H = gather(db %>% filter(Sexe == "H"), key = "portions", value = "tps_intermediaire", M0_15mod,
M15_25mod, M25_45mod, M45_50mod, M50_65mod, M65_75mod, M75_95mod, M95_100mod)

vecarts_h = c(reph(HNL, "M0_15mod", 1, 8), reph(HD, "M0_15mod", 17, 24),
reph(HB, "M0_15mod", 33, 40), reph(HP, "M0_15mod", 49, 56),
reph(HNL, "M15_25mod", 1, 8), reph(HD, "M15_25mod", 17, 24),
reph(HB, "M15_25mod", 33, 40), reph(HP, "M15_25mod", 49, 56),
reph(HNL, "M25_45mod", 1, 8), reph(HD, "M25_45mod", 17, 24),
reph(HB, "M25_45mod", 33, 40), reph(HP, "M25_45mod", 49, 56),
reph(HNL, "M45_50mod", 1, 8), reph(HD, "M45_50mod", 17, 24),
reph(HB, "M45_50mod", 33, 40), reph(HP, "M45_50mod", 49, 56),
reph(HNL, "M50_65mod", 1, 8), reph(HD, "M50_65mod", 17, 24),
reph(HB, "M50_65mod", 33, 40), reph(HP, "M50_65mod", 49, 56),
reph(HNL, "M65_75mod", 1, 8), reph(HD, "M65_75mod", 17, 24),
reph(HB, "M65_75mod", 33, 40), reph(HP, "M65_75mod", 49, 56),
reph(HNL, "M75_95mod", 1, 8), reph(HD, "M75_95mod", 17, 24),
reph(HB, "M75_95mod", 33, 40), reph(HP, "M75_95mod", 49, 56),
reph(HNL, "M95_100mod", 1, 8), reph(HD, "M95_100mod", 17, 24),
reph(HB, "M95_100mod", 33, 40), reph(HP, "M95_100mod", 49, 56))

Ecarts_H$moyenne_ecart <- round(vecarts_h, 3)
```

Tableau des temps d'entraînement des vitesses par portions pour toutes les épreuves :

| Portions | 100m Nage libre | 100m Dos | 100m Brasse | 100m Papillon |
|----------|-----------------|----------------|----------------|----------------|
| 0-15m | 3min et 52sec | 6min et 11sec | 4min et 50sec | 3min et 32sec |
| 15-25m | 8 min et 19sec | 4min et 33sec | 9min et 49sec | 7min et 16sec |
| 25-45m | 12min et 35sec | 14min et 44sec | 10min et 26sec | 11min et 55sec |
| 45-50m | 4min et 16sec | 4min | 2min et 43sec | 3min et 32sec |
| 50-65m | 8min et 43sec | 6min | 10min et 17sec | 9min et 19sec |
| 65-75m | 7min et 44sec | 8min et 55sec | 6min et 30sec | 8min et 56sec |
| 75-95m | 12min et 23sec | 13min et 38sec | 15min et 6sec | 13min et 25sec |
| 95-100m | 2min | 2min | 18sec | 2min et 3sec |

3.4 – REFERENCES

Article de l'impact de la morphologie sur les performances en natation de haut-niveau :

<https://hal-insep.archives-ouvertes.fr/hal-02925019/document>

Aide pour R Shiny :

<https://shiny.rstudio.com/tutorial/written-tutorial/lesson1/>

Aide pour créer des heatmaps :

https://ggplot2.tidyverse.org/reference/geom_tile.html

Aide pour le package plotly :

https://stt4230.rbind.io/tutoriels_etudiants/hiver_2018/plotly/

Aide pour les statistiques bayésiennes :

<https://youtu.be/x-2uVNze56s>

<https://easystats.github.io/bayestestR/articles/example1.html>

<https://www.amazon.fr/Choix-bay%C3%A9sien-Christian-P-Robert/dp/2287251731>

Site web du laboratoire MAP5 :

<https://map5.mi.parisdescartes.fr/presentation/>

Site web de la FFN :

<https://www.ffnatation.fr/>