HEC MONTRÉAL

La segmentation par séquence : une application sur l'évolution du style de jeu des joueurs dans un jeu vidéo FPS.

par

Clarisse Bailliart

Sciences de la gestion
(Option Intelligence d'affaires)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences
(M. Sc.)

Août 2018

Table des matières

Liste des tableaux

Table des figures

Sommaire

Le domaine du jeu vidéo connaît une croissance importante depuis la dernière décennie. Petits et grands studios se partagent donc un marché mondial de joueurs de plus en plus exigeants tant qu'à la qualité du jeu qu'au contenu du jeu.

Face à cela, les studios ont bien compris l'importance de l'implication du joueur dès le début du développement d'un jeu vidéo : en faisant tester leur jeu au fur et à mesure de l'avancée du développement, il est possible de corriger la majorité des points de friction qui peuvent apporter de la frustration au joueur et par conséquent assurer une meilleure rétention.

Tout élément du jeu peut déclencher de la frustration : que cela soit à cause d'une mauvaise compréhension des objectifs, de l'histoire ou encore des mécanismes du jeu (que l'on qualifie entre autre de *gameplay*). En plus de pouvoir venir de diverses sources, la frustration peut impacter négativement la rétention des joueurs. En effet, un joueur frustré abandonnera plus vite le jeu. Cette envie d'abandon face à la frustration est encore plus présente lors des premières heures de jeu. Lors des premières heures, le joueur découvre le jeu et se forge ainsi son opinion sur le jeu. Les méthodes UX (c'est à dire toutes méthodes pouvant caractériser l'eXpérience Utilisateur) utilisées sont nombreuses et variées pour analyser au mieux ce phénomène.
L'observation reste l'un des moyens les moins intrusifs pour analyser le comportement du joueur. Les entrevues et questionnaires à des moments spécifiques du jeu permettent d'appuyer les comportements observés.
De plus, les données quantitatives générées au sein du jeu lorsque le joueur joue (plus communément appelé les données de télémétrie) viennent appuyer ou réfuter des sujets abordés ou observés lors du test utilisateur en plus de pouvoir mettre l'emphase sur des actions effectuées de manière inconsciente par le joueur.

Proposer une bonne expérience de jeu au joueur est donc un objectif à retenir tout au long du développement d'un jeu, surtout dans les jeux à très gros budgets (les *Triple A*). Un des moyens d'avoir une bonne expérience de jeu sur un jeu de tir (*FPS*) est la possibilité au joueur de pouvoir jouer comme il le souhaite : c'est ce qu'on nomme le style de jeu.

Ce terme peut se définir comme étant toutes les stratégies que le joueur peut mettre en application au sein du jeu à l'aide des ressources disponibles. En d'autres termes, un joueur qui aime jouer de manière furtive devrait, à l'aide d'armes dans le jeu, de la disposition du décor sur le lieu de combat ainsi que des intelligences artificielles ennemies, pouvoir mettre en oeuvre sa propre stratégie de jeu et surtout pouvoir se sentir maître de son jeu.

Il est donc important que le joueur ait le contrôle sur sa manière de mettre en place ses stratégies. Cependant, il n'existe actuellement que très peu d'études sur le sujet spécifique du style de jeu, laissant l'industrie mettre en place elle-même quelques principaux styles comme : *le type Assaut*, *Approche furtive*, *combat rapproché* et *Sniper*.

Afin de proposer une expérience de jeu optimale au joueur, il est important de comprendre comment les joueurs s'approprient le jeu et plus particulièrement pouvoir identifier de manière simple et efficace les styles de jeu de ces derniers. Comme expliqué ci-dessus, les premières heures de jeu sont cruciales car ce sont elles qui vont motiver ou non le joueur à continuer à jouer. Notre problématique s'intéresse donc à l'identification et l'évolution des styles de jeu des joueurs sur les premières heures de jeu jouées. Le développement d'un modèle de ce genre permettrait, d'une part de mieux appuyer et enrichir les méthodes UX existantes et d'autre part pouvoir aider les équipes de développement à mieux comprendre comment les joueurs s'approprient et jouent au jeu. Enfin, vu qu'il s'agit d'une évolution, un intérêt par-

ticulier s'est porté sur une représentation graphique optimale des résultats.

Il est important de noter que ce mémoire s'est effectué en entreprise, ainsi de nombreux paramètres ont du être respectés :

1. La contrainte d'un échéancier court est très commun en entreprise. Ainsi, le développement de modèles simples à implémenter et à utiliser sont à privilégier afin de respecter le temps alloué ;
2. La possibilité de faire des comparaisons entre les jeux est souvent appréciée. En effet, il n'est pas rare en entreprise de vouloir comparer les produits développés entre eux pour mieux comprendre les points à améliorer : par exemple, comparer son produit au produit phare de l'entreprise est un moyen de comprendre où se situe son produit au sein de l'entreprise. Ainsi, l'utilisation de mêmes procédés statistiques permet de plus facilement appliquer les modèles d'un projet à un autre et donc de faire d'une pierre deux coups une comparaison plus simple des résultats entre les projets ;
3. Enfin, une présentation simple et claire de l'information est essentielle afin de partager les résultats aux équipes impliquées dans le développement du jeu. En effet, ces dernières sont souvent très occupées et ne consacrent donc de temps qu'aux résultats trouvés. Ainsi, une présentation des résultats sous forme graphique peut aider à la compréhension.

De par l'application directe en entreprise, il est important de prendre en considération les paramètres cités ci-dessus afin de développer le modèle le plus adapté.

Grâce à une analyse par regroupement, il a été possible d'identifier, selon le temps de jeu joué, entre 2 et 4 groupes. Malgré le fait que ces derniers soient stables en soit (c'est-à-dire qu'il existe peu d'évolution dans la définition des groupes), les joueurs se voient naviguer d'un groupe à un autre au fur et à mesure du temps passé dans le jeu, ce qui montre une grande adaptation de ces derniers ainsi qu'une aisance

dans toutes les stratégies qu'un jeu de tir peut proposer. Ces groupes se distinguent comme tels :

1. Le joueur assaut qui exploite les ressources du jeu : Un joueur assaut se qualifie comme étant un joueur qui utilise fréquemment des armes à courte portée pour tuer ses ennemis. Par conséquent il n'est pas rare que ce dernier montre des ratios plus élevés de morts enregistrées par ennemis tué. En effet, ce joueur aime le conflit rapproché ce qui l'expose plus à l'ennemi. Enfin, il exploite aussi les activités du *Monde Ouvert* proposées par le jeu, c'est à dire des activités que le joueur est libre de commencer, suspendre ou terminer quand il le souhaite. De plus, ces activités sont souvent annexes dans le sens où elles ne sont pas obligatoires pour terminer le jeu.
2. Le joueur furtif qui exploite les ressources du jeu : À l'inverse du joueur présenté précédemment, ce dernier privilégie les armes de longue portée tel que les *snipers* ou toute arme permettant de tirer de manière précise de loin. De plus, afin de ne pas se faire détecter par l'ennemi, il fait des tirs dans la tête (plus communément appelés *headshots*) qui permettent de tuer l'ennemi en un seul coup et donc de ne pas le mettre en alerte. Enfin, tout comme précédemment, ce dernier exploite les ressources du jeu comme les activités du *Monde Ouvert*.
3. Le joueur assaut qui n'exploite pas les ressources du jeu : Outre le fait qu'il soit un type assaut, ce dernier joue plus au jeu pour la stratégie du jeu plutôt que pour profiter des particularités que le jeu offre. En d'autres termes, il n'exploite pas les attributs qui différencie le jeu d'un autre du même genre.
4. Le joueur polyvalent : Ce joueur maîtrise aussi bien l'approche furtive que le type assaut ce qui lui permet de s'adapter facilement à toutes les situations du jeu. Tout comme le joueur assaut, ce dernier exploite moins les ressources disponibles dans le jeu.

Comme expliqué, au cours du temps, les résultats montrent des groupes relativement stables concernant la définition mais les flux d'individus varient d'un temps à un autre. Afin de répondre à la deuxième partie de la problématique qui est la représentation graphique des résultats, les visualisations choisies représentent le flux migratoire des individus d'un groupe à un autre au cours du temps. Pour cela, deux visualisations ont été utilisées :

— Un diagramme de Sankey qui est une visualisation souvent utilisée pour représenter des flux d'un état à un autre. Ce type de visualisation permet donc aux équipes de mieux comprendre le comportement de jeu de chaque joueur.

— Une représentation des positions sur la carte du jeu qui permet de voir si certains endroits dans le jeu sont plus propices à un style qu'à un autre. En effet, une visualisation plus géographique de nos données permet aux concepteurs de jeu de voir exactement comment les joueurs exploitent Far Cry 5 et peuvent faire les ajustements ciblés sur des parties de la carte afin de réajuster les possibilités de styles de jeu.

Ces deux aspects de la problématique (le développement d'une analyse de segmentation ainsi qu'une représentation efficace des résultats) ont ainsi permis de développer une solution efficace, pratique et réutilisable sur d'autres projets. Ainsi, cette étude a pu être appliquée, pour l'instant, sur un projet au sein d'Ubisoft Montréal et sera notamment proposé pour un futur projet, ce qui prouve que la solution, en plus d'être une réponse aux questions que se posent les équipes de développement de jeu, permet d'être une solution pratique vu qu'elle assimile bien les paramètres de l'application d'outils statistiques que nous pouvons retrouver en entreprise.

Remerciements

Le cheminement mémoire est connu pour être difficile de par la rigueur à adopter, les recherches poussées à effectuer ainsi que l'écriture à faire. Cependant, j'en garde un superbe souvenir grâce au soutien de très nombreuses personnes.

Avant toute chose, j'aimerais remercier mon directeur de mémoire, Laurent Charlin. Développer un mémoire en entreprise n'est pas une mince affaire et il est aujourd'hui possible de qualifier ce travail d'un mémoire grâce à son aide, ses conseils, sa disponibilité et surtout sa qualité de compréhension des enjeux d'entreprise qui impactent le mémoire.

Je tiens à remercier du fond du coeur toute l'équipe du laboratoire en Recherche Utilisateur d'Ubisoft car, sans elle, ce mémoire n'aurait pas vu le jour. La liste est bien longue et il est difficile de retranscrire sur papier la chance que j'ai eu d'être entourée de personnes brillantes, intéressantes et curieuses d'en savoir plus sur mon mémoire.
Isabelle, pour m'avoir donné l'opportunité d'effectuer mon mémoire au sein de son équipe ainsi que m'aider à concilier travail et étude.
Amine, pour son rôle de mentor dans mon mémoire. Il a toujours su se rendre disponible : sans son aide précieuse et ses idées adaptées à Ubisoft, ce mémoire n'aurait jamais été aussi intéressant. Je ne pensais pas rencontrer un mentor aussi impliqué dans l'avancée de mon mémoire, encore moins d'imaginer qu'il porterait un tutu au Zeitgeist pour l'occasion !
Merci à toute l'équipe Far Cry qui a toujours pris très au sérieux mon mémoire en m'aidant par tous les moyens possibles. Louise, qui a pris le temps de le lire et me référer aux bonnes personnes. Stéphane, Sop et Amandine pour le soutien au quotidien, où chaque petite avancée était célébrée.
Je remercie Marc avec qui j'ai pu discuter de mes problématiques plus techniques

lors de nos entraînements en vue de préparation à un semi-marathon.

Même s'il le sait, mes remerciements vont tout particulièrement à Lucas. Je ne le remercierai jamais assez pour sa présence tout au long de mon mémoire, pour sa patience, son soutien, sa gentillesse, son aide et ses encouragements. Sans lui à mes côtés, le mémoire aurait été un parcours du combattant !

Merci à toute ma famille qui a toujours cru en moi et en mes capacités. Merci à mes parents d'avoir rendu mon expatriation au Canada aussi facile. Je leur dois tout ce que j'ai accompli jusqu'à présent. Les mots me manquent pour exprimer toute la gratitude que j'ai envers eux.

L'achèvement de ce mémoire marque la fin de mes études au HEC Montréal. Je remercie tout le corps professoral pour leur passion ainsi que leur qualité d'enseignement qui prépare extrêmement bien à la vie active.

Chapitre 1

Introduction : L'histoire du jeu vidéo et du style FPS

Il est question, dans ce chapitre, d'une mise en contexte du jeu vidéo, son émergence jusqu'à son évolution en industrie telle que nous la connaissons aujourd'hui.

1.1 Naissance du jeu vidéo et focus sur le jeu Triple A

Aussi surprenant que cela puisse paraître pour la jeune génération, le jeu vidéo est né au sein des universités lors de recherches en informatique, dans les années 50. C'est en 1950 que le premier jeu connu, *Bertie the Brain*, est développé par Josef Kates dans le cadre de l'Exposition Nationale Canadienne. Ce jeu permet au public de jouer, avec différents niveaux de difficulté, au *Tic-tac-toe* face à une intelligence artificielle. Cependant, ce jeu requiert une machine spéciale afin de pouvoir y jouer (Lallain (2015)).

Il faut attendre 1962 pour qu'un groupe d'étudiants du MIT programme un jeu nommé *Spacewar !*. Ce dernier met en scène deux vaisseaux spatiaux, chacun contrôlé par un joueur à l'aide de manettes. Il s'agit du premier jeu d'arcade ayant un impact

sur le grand public.

C'est dans les années 70 que le jeu vidéo devient un produit commercialisé en masse, tout d'abord sous la forme de bornes d'arcade, puis sous forme de consoles de salon. En effet, l'entreprise *Atari* est créée en 1972 et développe le jeu *Pong*, premier jeu à devenir populaire. La même année, la compagnie *Magnavox* commercialise la première console de jeu de salon, l'*Odyssey*. Quittant ainsi les arcades pour s'installer confortablement dans les salons, la démocratisation des jeux vidéo ainsi que leur aspect ludique débutent donc.

Face à une popularité croissante des jeux vidéo, le marché de la console en Amérique du Nord se retrouve vite saturé et connaît le *North American video Game Crash* en 1983 (DeMaria and Wilson (2002)). Aussi connu sous le nom de *Atari shock* au Japon (Ernkvist (2008)), les revenus qui culminaient 3.2 milliards de dollars en 1983 sont tombés à environ 100 millions de dollars en 1985 (soit une baisse de près de 97%). Plusieurs facteurs étaient en cause :

1. Création de nombreuses consoles de salon, chacune ayant sa propre bibliothèque de jeux produite par le fabricant de la console et des développeurs tiers spécifiques. Cette surproduction de jeux met en avant la quantité plus que la qualité des jeux produits (Jones (1982)).

2. Compétition croissante des PC qui sont populaires dans les foyers. De par leurs meilleurs graphismes, sons, plus grande mémoire et processeurs plus rapides que les consoles de jeux, les PC permettent le développement de jeux plus sophistiqués.

Après cette crise, les compagnies de jeu vidéo ont décidé de mettre en place un terme standard afin d'aider l'audience à distinguer les titres de qualité parmi tous les jeux développés. Par exemple, Nintendo développe son propre système avec son label *Nintendo Seal of Quality*, indiquant que les jeux ont été correctement testés et

approuvés (nintendo.com (2018)).

Il faudra attendre la fin des années 90, lors d'une convention du jeu vidéo aux Etats-Unis, pour que certains studios emploient le terme AAA, basé sur l'échelle de notation du système académique aux Etats-Unis, où les jeux peuvent être comparés aux films blockbusters (Steinberg (2007)) : les jeux Triple-A voient le jour.
De nos jours, un jeu Triple-A est considéré comme un jeu à gros budget marketing, de grande qualité (camadegames.com (2014)).

1.2 Histoire du jeu FPS

De nombreux genres de jeux existent comme les jeux d'action, de combat, de plateforme, d'aventure, d'horreur, de stratégie mais aussi des jeux de réflexion, jeux de rôle ou encore jeux de tirs. Ces termes sont utilisés afin de pouvoir mieux catégoriser les jeux. Cependant, ces genres sont définis selon un *gameplay*. Olivier Lejade (2013) caractérise ce terme comme étant : les éléments que vit le joueur lors d'une expérience vidéoludique.

Parmi les genres mentionnés figurent les jeux de tir qui regroupent les Jeux de Tirs à la Première Personne (plus connu sous l'acronyme *FPS (First Person Shooter)*). Il est possible de le définir comme étant un jeu de tir où le joueur incarne un personnage. Ainsi, les visées et les déplacements sont en vue subjective. Cela a pour but de rendre l'expérience du joueur plus immersive, à la place du personnage, tenant l'arme et voyant exactement ce que le personnage voit. Cette perspective génère une forte identification accentuée par des graphismes en 3D.

Ce genre de jeu est apparu tardivement, essentiellement par nécessité de la 3D disponible en temps réel. Le premier jeu répertorié de ce genre est *Maze War* développé par un ingénieur de la Nasa en 1974. Avec 2 ordinateurs reliés, deux joueurs

se baladent en vue subjective dans un labyrinthe case par case et le but est simple : éliminer son adversaire en lui tirant dessus. Ce jeu marque le début de la vue subjective ainsi que l'intérêt du monde 3D dans le développement d'un *FPS*.

Nombreux sont ceux qui affirment que le réel premier succès commercial du jeu du FPS est *Wolfenstein 3D* qui sort en 1992. Le joueur est prisonnier dans un château Nazi dans lequel il doit s'échapper : il lui faut donc tuer les soldats Nazis qu'il rencontre sur son chemin lors de sa fuite. Le moteur 3D est fluide, les textures détaillées et variées et le jeu met en place la caractéristique principale que nous retrouvons encore de nos jours dans les *FPS* : la vue de l'arme à feu sur le bas de l'écran. Aussi, la notion de gestion de l'inventaire (des armes, des munitions selon l'arme utilisée) est mise au grand jour avec ce jeu. Enfin, malgré sa popularité, les premières critiques, que nous retrouvons encore aujourd'hui dans ce genre de jeu, sur la violence du jeu apparaissent (Ez' (2015a)).

Un an plus tard, le jeu *Doom* sort. Ce dernier plonge le joueur dans un univers futuriste où le design du jeu est beaucoup plus travaillé : possibilité de monter des étages, d'interagir avec l'environnement du jeu (par exemple faire exploser des bidons d'essence à l'aide de balles). Ce jeu marquera mondialement le genre par ses graphismes et son ambiance et représente un des plus grands succès de l'histoire du *FPS* (Ez' (2015b)).

Grâce à ces deux succès sur PC, les consoles commencent à voir l'intérêt de ce genre de jeu et décident de l'exploiter à partir du milieu des années 90. Cependant, le succès ne sera pas aussi important que sur PC de par la difficulté de viser avec une manette plutôt qu'une souris d'ordinateur.

Longtemps, le scénario des *FPS* demeure secondaire, mettant plutôt en avant la jouabilité, la violence du jeu et la beauté des graphismes. Par exemple, *Half-Life* est

un des jeux importés sur console et qui est modifié en 1999 par les joueurs à travers son mode multijoueur pour déboucher sur le jeu *FPS* le plus populaire : *Counter Strike*. Depuis maintenant presque 20 ans, il s'agit toujours du jeu *FPS* en ligne le plus joué et le plus populaire.

Il faut attendre l'arrivée du nouveau millénaire pour que de nombreux jeux se démarquent grâce à leur *gameplay* remarquable, leurs graphismes ou scénarios de plus en plus proche d'oeuvres cinématographiques.

1.3 Les jeux vidéo aujourd'hui

Avec un chiffre d'affaire global de 108.9 milliards de dollars en 2017, soit une augmentation de 7.8 milliards de dollars ou de 7.8% comparé à l'année précédente (McDonald (2017)), le domaine du jeu vidéo s'impose largement face à des industries comme le cinéma ou la musique : en effet ce dernier possède un marché 5 fois plus gros que celui de la musique et 1.5 fois plus gros que celui du cinéma (Taylor (2016)).

Il s'agit donc d'un domaine en pleine expansion qui touche de plus en plus de consommateurs. Les grosses compagnies de jeux vidéo apportent une dimension cinématographique à leurs jeux en investissant dans des moteurs de jeu performants afin de produire des graphismes de plus en plus réalistes. Ils investissent aussi dans le domaine narratif pour mettre sur pied des histoires qui seront les fils conducteurs de leurs jeux ainsi que dans l'audio pour donner une identité à leurs jeux grâce à des trames sonores uniques.

Avec la popularité des jeux de tir à la première personne, ces derniers n'échappent pas à ces gros studios qui proposent désormais des jeux d'une qualité irréprochable avec des scénarios qui permettent aux joueurs de suivre l'histoire de leur personnage sur plusieurs dizaines d'heures.

1.4 L'importance du style de jeu

La recherche en expérience utilisateur (plus communément appelé *UX Research*) a récemment gagné beaucoup de terrain dans le développement des jeux. Il est devenu évident à quel point il est bénéfique de connaître le public ciblé par son produit. De plus en plus, les studios intègrent des méthodes centrées sur le joueur afin d'améliorer le processus de développement d'un jeu. De par le fait que la *UX Research* soit un champ spécifiquement orienté sur l'utilisateur (son expérience, son jeu, son amusement, sa frustration, etc) et que l'utilisateur représente aussi un client, il va donc de soit que la *UX Research* aide à améliorer les jeux et par conséquent les ventes. Ainsi, ce domaine connaît un investissement constant et considérable afin d'augmenter le plaisir et la rétention des joueurs (Drachen et al. (2013)).

Le mot *gameplay* (ou l'ensemble des mécanismes de jeu) est souvent utilisé pour juger le ressenti des joueurs : un « bon » *gameplay* tient le joueur en éveil tandis qu'un « mauvais » *gameplay* le frustrera ou l'ennuiera. D'après Duflo (1997), maître de conférence en philosophie, le *gameplay* forme un ensemble de règles « constitutives » et de règles « régulatives ».

Les règles « constitutives » représentent les choix de *game design* (ou conception de jeux), c'est-à-dire le processus de création et de mise au point des règles et autres éléments qui rendent l'expérience ludique : règles, ergonomie, difficulté, etc. L'un des aspects importants de l'élaboration du système de jeu est le personnage, c'est-à-dire la définition des actions que le joueur pourra entreprendre à travers son personnage.

Les règles « régulatives » sont toutes les stratégies mises en place par le joueur qui lui permettent d'atteindre de la meilleure façon possible les objectifs définis par le jeu.

Ainsi, les règles « régulatives » découlent des règles « constitutives ». Les tests utilisateurs ont donc pour but de s'assurer que le *game design* créé par les concepteurs du jeu est compris et se marie bien avec les différentes stratégies de jeu employées par les joueurs, soit les règles « régulatives ». En effet, la richesse d'un jeu se voit dans les innombrables possibilités que le joueur a pour atteindre l'objectif déterminé par le jeu. En d'autres termes, dans un *FPS*, la multitude d'armes mises à disposition, l'environnement où le joueur évolue ainsi que les types d'ennemis permettent au joueur d'aborder l'objectif sous plusieurs angles. Pour arriver à mieux cerner ces stratégies, les tests utilisateurs utilisent de nombreux outils comme l'observation directe d'un joueur en train de jouer, des entrevues pour aller chercher le ressenti des joueurs concernant le jeu ou encore l'utilisation de questionnaires afin de noter l'appréciation des contenus effectués.

De par l'avancée technologique, les jeux développés de nos jours possèdent bien plus de contenus jouables et demandent très souvent au minimum une dizaine d'heures pour terminer l'histoire principale voire grimper à une centaine d'heures si le joueur cherche à terminer à 100% le jeu (c'est-à-dire être capable de terminer toutes les quêtes secondaires du jeu, récolter toutes les récompenses possibles à travers les défis proposés, etc). Ainsi, l'importance d'un « bon » *gameplay* est primordial pour limiter les frustrations du joueur, augmenter son plaisir et donc lui donner envie de jouer le plus longtemps possible. De ce fait, tout l'intérêt de la suite de ce mémoire réside dans l'exploitation de méthodes d'apprentissage non supervisés combinés à de la visualisation simple des données afin de faire ressortir les styles de jeu des joueurs et donc en faire un outil pour les développeurs de jeu.

1.5 La collecte de données

L'une des méthodes les plus répandues pour collecter de manière non-intrusive les données d'un grand nombre d'utilisateurs est la télémétrie. En effet, de nombreux

domaines y ont recours pour récolter des données sur :

— **des patients** : Dans le domaine médical, des études ont utilisé la télémétrie afin de pouvoir localiser les instruments médicaux dans le corps des patients ainsi que la réponse enregistrée par les corps en contact de ces instruments Ferre (1995) ;

— **des clients** : La célèbre plateforme de films et séries, *Netflix*, utilise la télémétrie pour analyser les données de visionnement de films afin de mettre en place des systèmes de recommandations. De plus, une grande portion des sites web utilisent la télémétrie dans le but de mieux comprendre le trafic de leur clientèle sur leurs sites ;

— **des joueurs** : Un des buts de la collecte de données des joueurs peut être d'améliorer le jeu en étudiant par exemple les schémas d'apprentissage de compétences des joueurs.

De manière plus générale, la télémétrie est une technologie qui permet de mesurer une quantité à distance. Le concepteur de la télémétrie est donc libre de choisir quel(s) élément(s) pour lequel (lesquels) il souhaite avoir de la donnée. La mesure est effectuée puis transmise ailleurs pour y être stockée et éventuellement traitée. Dans le cas des jeux vidéo, de nombreuses mesures quantitatives spécifiques à l'environnement du jeu à un temps précis peuvent être mises en place :

— le nombre de morts,

— les positions des joueurs dans le jeu,

— le nom de l'arme utilisée pour tuer un ennemi,

— les objets ramassés par le joueur,

— etc.

Cette méthode de collecte a de nombreux avantages puisqu'elle permet de récolter facilement et rapidement des ensembles de données de chaque joueur qui sont ensuite compilés, triés et analysés à un seul et même endroit. Dans le cadre plus spécifique

du jeu vidéo, les données récoltées sont majoritairement des données *ingame*, c'est-à-dire des données exclusives du jeu joué. Ces données-ci peuvent donc facilement appuyer ou réfuter des observations ou des hypothèses faites en test utilisateur : par exemple, lorsqu'un joueur se plaint que le jeu est trop difficile, il est fréquent d'analyser le nombre de morts du joueur et à quel(s) emplacement(s) afin de mieux comprendre cette difficulté et donc pouvoir améliorer le jeu.

1.6 Problématique

La première expérience du joueur avec le jeu est cruciale pour garantir son retour dans le jeu. En effet, la première expérience se base sur les premières heures de jeu, il est donc primordial d'y porter une attention toute particulière en essayant de minimiser la frustration qui représente la raison numéro 1 d'une mauvaise expérience de jeu.

Pour cela, beaucoup d'aspects sont pris en compte, comme un bon équilibre de l'économie à l'interne (par exemple donner suffisamment de ressources au joueur au début du jeu pour qu'il puisse s'acheter de l'équipement, sans pour autant trop en donner pour ne pas l'ennuyer rapidement) ou encore un bon calibrage de la difficulté des ennemis.

Cependant, un élément crucial est à prendre en considération dans un jeu *FPS* : le style de jeu. Comme nous l'avons développé dans l'introduction, si le joueur ne peut utiliser ses propres stratégies de jeu, ce dernier peut vite se sentir frustrer et donc quitter le jeu.

Or, outre les techniques utilisées lors des tests utilisateurs comme l'observation, les questionnaires ou entrevues, il est difficile de pouvoir donner une vision générale du style de jeu que peut avoir un joueur lors d'un test utilisateur. En effet,

l'observation est un processus lourd pour les modérateurs et demander une réponse directement de la part des joueurs représente une approche obstructive qui le sort de son expérience de jeu. De plus, il est commun que les déclarations faites par le joueur soient fausses : il n'est pas rare que le joueur mette la faute sur le jeu plutôt que sur sa manière de jouer qui n'est pas adéquate (par exemple un joueur se plaint de ne pas pouvoir jouer de manière furtive alors que ce dernier essaye une approche furtive juste devant l'ennemi qui le détecte facilement). Ainsi, il n'existe actuellement pas de manière efficace pour représenter le style d'un joueur ainsi que de voir son évolution dans le temps.

Pourtant, la télémétrie est un outil non obstructif très efficace pour mieux comprendre le comportement du joueur dans le jeu et peut donc être utilisé pour développer des outils statistiques adéquats qui peuvent aider à mieux caractériser le style de jeu d'un joueur. Grâce aux très nombreuses données *ingame* récoltées pour chaque joueur, il est possible de dresser un portrait assez détaillé de l'expérience de jeu du joueur (à travers des questions telles que : est-ce qu'il meurt souvent dans le jeu ? Est-ce qu'il prend son temps pour effectuer les missions ? Est-ce que ce dernier effectue les missions secondaires qui ne sont pas obligatoires dans l'avancée du jeu ? Aime-t-il découvrir l'environnement mis à sa disposition ? etc...). Cette richesse d'informations disponible grâce à la télémétrie, aussi bien à travers la quantité d'informations envoyée que le nombre d'éléments qu'il est possible de traquer, cela permet de développer de nombreux outils statistiques selon les questions de recherche abordées.

De par l'abondance de données disponibles à analyser grâce à la télémétrie ainsi que le peu de documentation disponible sur le style de jeu dans un jeu *FPS*, ce mémoire propose une approche d'apprentissage non-supervisé pour représenter le style de jeu des joueurs lors des tests utilisateurs ainsi que son évolution au cours du temps. L'approche non-supervisée a été privilégiée car il est difficile de clairement

définir le nombre de styles de jeu dans un jeu *FPS* : en effet, chaque joueur est libre de développer ses propres stratégies de jeu, donc définir un nombre précis de groupes risquerait de ne pas découvrir certains groupes à motifs cachés. En d'autres termes, « laisser parler » les données permettrait de faire ressortir des groupes auxquels nous n'aurions pas pensé.

Cette approche a pour but d'être appliquée directement en entreprise, ainsi l'utilisation d'un modèle simple à expliquer et à comprendre ainsi qu'une représentation graphique adéquate sont à privilégier. En effet, ces modèles doivent pouvoir facilement être expliqués à des équipes business ou de designers qui ne sont pas familiers avec les statistiques.

La question à laquelle nous voulons répondre à travers ce mémoire est donc la suivante :

Dans un contexte de test utilisateur, comment, dans un premier temps, peut-on représenter l'évolution du style de jeu des joueurs à travers un modèle statistique facilement implémentable, explicable et transférable d'un projet à un autre ? Dans un deuxième temps, quel(s) type(s) de représentation(s) graphique(s) correspondraient le mieux pour représenter l'évolution des styles de jeu au cours du temps ?

1.7 Contribution

Comme nous le verrons tout au long de ce mémoire, notre contribution se matérialise à travers la proposition d'un procédé applicable en entreprise. L'analyse par regroupement est très utilisée en entreprise car en plus d'être peu contraignante à mettre en place, elle est aussi simple à expliquer. De plus, il existe un nombre exhaustif de revues de littérature traitant ce sujet-là. Enfin, il existe de très nombreuses méthodes au sein même de l'analyse par regroupement, ce qui permet de

trouver, pour chaque spécificité que la base de données peut offrir, une méthode adéquate. Il sera ainsi possible de comparer la performance de chaque méthode et choisir celle qui est la plus adéquate. Le choix des méthodes réside essentiellement sur la facilité de l'implantation de ces dernières d'un projet à un autre ainsi que la rapidité de calcul.

Enfin, deux visualisations seront proposées afin de mettre en avant l'évolution des groupes au cours du temps. Ces visualisations choisies ont pour but principal d'être simples à comprendre ainsi qu'à être exploitées d'un projet à un autre.

Chapitre 2

Revue de littérature

Dans le cadre de notre étude, il a été décidé de faire une analyse par regroupement afin d'identifier les groupes de styles de jeu ainsi que l'évolution dans le temps. Cette revue de littérature a donc pour but d'identifier les méthodes et techniques déjà utilisées pour traiter l'analyse par regroupement. De manière complémentaire, cette section portera également sur les outils nécessaires pour sélectionner l'algorithme de regroupement le plus adapté à notre cas. De plus, nous ferons un tour d'horizon des approches existantes pour répondre à la question de l'évolution des groupes dans le temps. Enfin, nous porterons un regard sur les autres travaux qui existent sur la modélisation des styles de jeu. Cet exercice nous permettra d'identifier les bonnes techniques à appliquer dans notre cas et donc de mettre sur pied un procédé ré-adaptable pour des jeux de données similaires.

2.1 L'analyse par regroupement

L'analyse par regroupement (ou *clustering*), est une méthode d'analyse de données et est plus particulièrement une méthode de classification non supervisée de modèles (qui peuvent être des joueurs comme dans notre cas) en groupes (ou *clusters*). Elle vise à diviser un ensemble de données en groupes homogènes, c'est à dire que les données existantes dans chaque sous-ensemble partagent des caractéristiques

similaires, qui correspondent le plus souvent à des critères de proximité. Pour obtenir un bon partitionnement, il convient à la fois :

— que les individus *intra-groupe* soient les plus *homogènes* possibles

— que les individus *inter-groupe* soient les plus *hétérogènes* possibles

Pour résumer plus facilement ces deux points, Everitt et al. (2001) définit un *cluster* comme étant :

1. « Un *cluster* regroupe un ensemble d'entités qui se *ressemblent* mais les entités de différents *clusters* ne se ressemblent pas. »
2. « Un *cluster* est une agrégation de points dans l'espace de telle sorte que la distance entre deux points quelconques du *cluster* soit inférieure à la distance entre n'importe quel point du *cluster* avec tout point qui n'appartient pas au *cluster*. »
3. « Un *cluster* peut être décrit comme une régions connectée dans un espace multidimensionnel contenant une densité relativement élevée de points, séparé des autres régions similaires par une région contenant une densité relativement faible de points. »

D'un point de vue plus mathématique, en nous basant sur p variables $(X_1,X_2,...,X_p)$ qui caractérisent nos individus, nous recherchons à créer des groupes d'individus homogènes. Comme mentionné plus haut, le but est de former et de regrouper les individus en groupes homogènes, c'est à dire que ces derniers soient le plus semblables possibles par rapport à certaines caractéristiques tout en ayant des groupes les plus différents possibles.

Bien qu'il soit facile de donner une définition fonctionnelle claire de ce qu'est un groupe, il est très difficile de donner une définition opérationnelle d'un groupe. Comme le mentionne toujours Everitt et al. (2001), les points peuvent être regroupés en groupe dépendamment de nos objectifs de recherche. Ce principe est illustré dans le graphique 2.1.

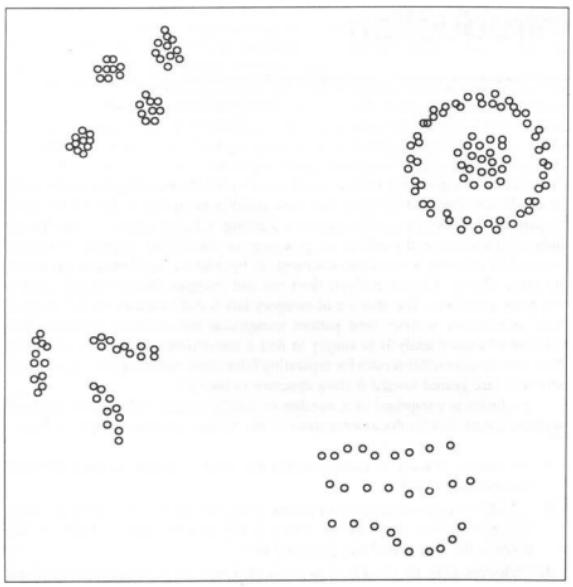


FIGURE 2.1 – Exemple de groupes de points dans un espace à deux dimensions

Nous remarquons dans la figure 2.1 ci-dessus que selon le niveau de similarité utilisé, les résultats de l'étude ne seront pas les mêmes. En effet, à un haut niveau ou plus grand niveau de similarité nous apercevons quatre groupes mais à un niveau plus local ou à un niveau plus bas de similitude, nous remarquons douze groupes. Il est donc crucial d'étudier un même cas à plusieurs niveaux de similarités dans les données selon nos objectifs de recherche.

La méthode non-hiérarchique

La méthode non-hiérarchique se différencie de la méthode hiérarchique par le fait que le nombre de groupes ainsi que les « centres » de ces derniers sont à spécifier au départ. De ce fait, l'algorithme cherche à partir d'une solution initiale, la meilleure distribution des sujets à travers le nombre de groupes prédéterminés d'une manière itérative. Ainsi, d'une itération à une autre, l'assignation d'un sujet à un groupe peut changer. Le problème majeur de cette méthode est la spécification de nombre de groupes ainsi que des « centres » au départ. En effet, la détermination

des centres initiaux est très sensible car d'un changement de centre à un autre cela peut grandement influencer le groupement final et donc l'interprétation des groupes.

L'algorithme des K-moyennes

De part l'application facile ainsi que l'interprétation simple des résultats, l'algorithme de K-moyennes est de loin l'une des méthodes non-hiérarchique les plus utilisées dans le domaine de l'apprentissage non-supervisé appliqué à l'analyse de regroupement. En effet, pour ne citer que quelques exemples de l'application de cet algorithme dans de nombreux domaines :

- En génétique : application dans l'analyse de l'expression des gènes (Lu et al. (2004)) ;
- En milieu universitaire : application pour la prédiction de la performance universitaire des étudiants (Oyelade et al. (2010)) ;
- En marketing : à travers la segmentation du marché pour découvrir des groupes de clients distincts à partir de base de données d'achats (Punj and Stewart (1983)) ;
- En environnement : catégorisation des risques dans le domaine de la chimie industrielle (Shi and Zeng (2014)) ;
- En assurance : identification des comportements de fraude dans le domaine de l'assurance automobile (Ghorbani and Farzai (2017)).

Le but premier de cet algorithme est de créer des groupes stables. Comme le montre la figure 2.2, l'algorithme fonctionne comme-suit :

1. On définit le nombre de groupes souhaités ainsi que les centres initiaux de chaque groupe, nommés dans l'illustration C_i avec $i \in \{1, 2, 3, 4\}$.
2. Constitution de groupes autour des centres C_i. Chaque groupe formé montre les points les plus proches du centre C_i relié à ce groupe en particulier ;

3. Calcul de nouveaux centres de gravité des groupes formés qui sont aussi nommés dans l'illustration C_i avec $i \in \{1, 2, 3, 4\}$. De nouveaux groupes sont formés autour de ces nouveaux centres.

4. On continue l'étape précédente jusqu'à temps que les centres de gravité des groupes formés ne bougent plus. Ainsi, l'assignation finale met en évidence des groupes stables.

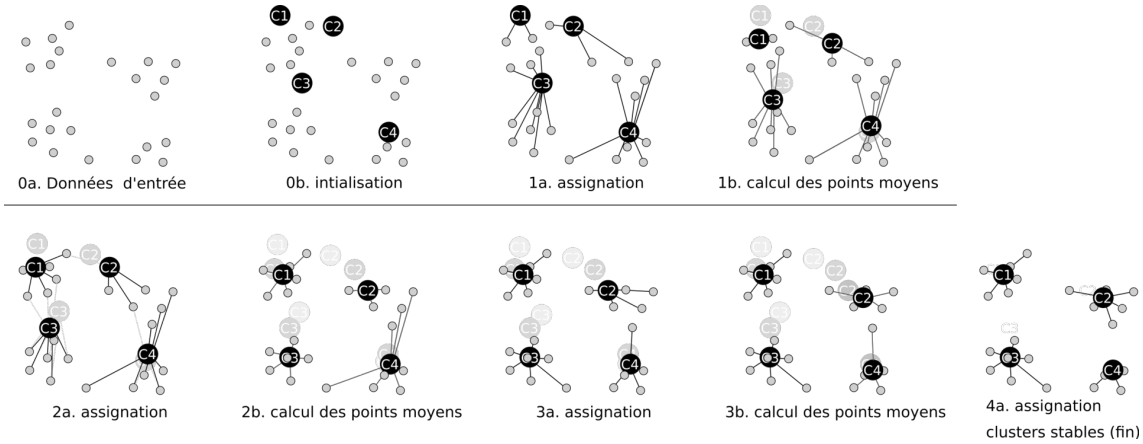


FIGURE 2.2 – Illustration du déroulement de l'algorithme des K-moyennes (Mquantin (2017))

L'algorithme des K-moyennes nous donne un ensemble de groupes qui n'ont ni structure particulière inter-groupe ni structure particulière intra-groupe. Cependant, il se pourrait facilement que certains groupes soient étroitement liés à d'autres groupes et plus éloignés avec d'autres. De manière plus concrète, si nous effectuons un traitement d'images, il est possible que nous ne voulions pas qu'un groupe de fleurs mais aussi avoir des roses et des lilas à l'intérieur de ce dernier. Ou alors, si nous regroupons les dossiers médicaux des patients, il est utile d'avoir un « super-groupe » *Plaintes Respiratoires* qui regroupe les *Pneumonies, Influenza et SRAS*. Dans le cas de notre étude, il serait intéressant d'avoir des super-groupes qui peuvent regrouper des ensembles intéressants de styles de jeu. Pour cela, une méthode hiérarchique peut-être plus adéquate (Shalizy (2009)).

Les méthodes hiérarchiques

Fonctionnement de l'algorithme de la méthode ascendante hiérarchique

Tout comme la méthode des K-moyennes, la méthode hiérarchique ascendante est utilisée dans de nombreux domaines :

— Le domaine universitaire : afin d'évaluer la performance universitaire des élèves d'un Institut (Rana and Garg (2016)) ;

— Le domaine médical : afin d'identifier les modèles de changement de coût des patients atteints d'insuffisance rénale terminale (ESRD) (Sousa and Gama (2014)) ;

— Le domaine environnemental : afin de décrire la trajectoire du vent de basse altitude au dessus de la région où se trouve la rivière La Plata en Amérique du Sud (Ratto et al. (2014)).

Ainsi, nous remarquons que cette méthode-ci est utilisée dans autant de domaines variés que la méthode des K-moyennes.

L'algorithme de la méthode hiérarchique ascendante est simple. En effet, Murtagh (1983) explique qu'au départ, chaque sujet, nommé C_i avec $i \in \{1, 2, ..., N\}$, est dans un groupe. Il y a donc autant de groupes que de sujets, soit N groupes.
Larocque (2016) explique aussi que la distance entre chaque paire de groupes est calculée et les deux groupes ayant la plus petite distance qui les sépare sont regroupés pour ne former qu'un seul groupe. Ainsi, nous nous retrouvons avec *(N-1)* groupe.
La plus petite distance est calculée entre chaque paire parmi les *(N-1)* groupes afin de donner *(N-2)* groupes. Cette même étape se reproduit jusqu'à ce que tous les sujets soient regroupés dans un seul et même groupe. Cet algorithme particulier est nommé méthode ascendante hiérarchique. Il existe aussi la méthode descendante hiérarchique qui débute par un seul groupe qui se compose tous les sujets pour en arriver à un groupe par sujet.

La représentation graphique de l'algorithme peut être vue grâce au dendrogramme formé, comme le montre notre exemple présenté dans la figure 2.3. Il faut le lire de bas en haut, il présente l'historique de la méthode hiérarchique. Les sujets sont alignés sur l'axe des abscisses et l'axe des ordonnées donne la perte d'inertie interclasse nommée le SPRSQ.

Stéphane (2012) définit le SPRSQ (ou Semi-partial R-squared), de notation

$$SPRSQ = \frac{\Delta I_B}{I},$$

représente le pourcentage de la diminution de la perte d'inertie interclasse provoquée en regroupant 2 classes (notée ΔI_B) par rapport à l'inertie minimum de toutes les observations regroupées (notée I). Le but étant d'avoir une inertie interclasse la plus grande possible, on recherche un pic pour k classes et un creux pour k+1 classes. Ainsi, cela indique une bonne classification en k+1 classes. L'exemple présenté dans la figure 2.4 illustre bien les propos précédents.

Le dendrogramme se divise en partitions et peut donc aider à déterminer le nombre de groupes à retenir.

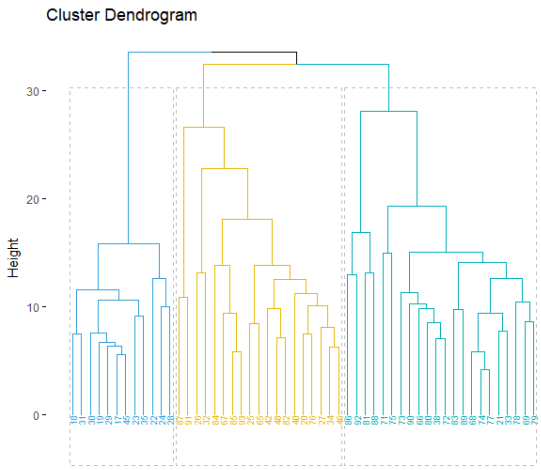


FIGURE 2.3 – Exemple d'un dendrogramme qui illustre une procédure de regroupement hiérarchique sur un jeu de données de points en 2D.

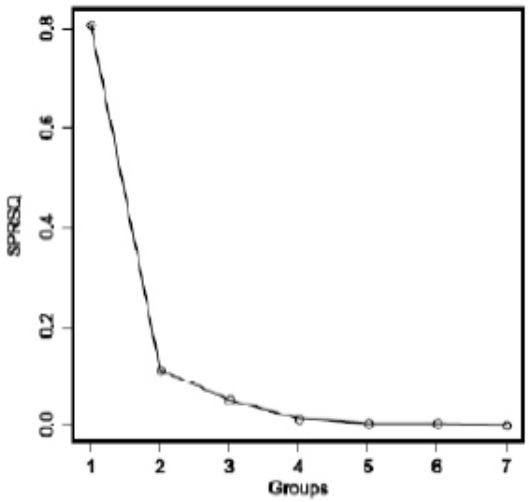


FIGURE 2.4 – Exemple d'un graphique qui présente la perte d'inertie interclasse (ou de distance) nommée SPRSQ selon le nombre de groupes.

L'émergence de la structure du regroupement dépend de plusieurs choix : la représentation et la normalisation des données, le choix d'une mesure de dissemblance et enfin le choix de l'algorithme de regroupement (donc dans notre cas la méthode hiérarchique ou la méthode des K-moyennes).

La matrice de dissemblance

Tout algorithme de regroupement utilise une mesure de dissemblance ou de similarité. Une matrice de similarité donne un score élevé aux individus « similaires » comme le fait la Corrélation de Pearson par exemple, de formule $\rho = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y}$ pour les variables x et y.

Une matrice de dissemblance sert donc à quantifier la « distance » reflétée entre deux sujets en se basant sur leurs variables X_1,...,X_p.

Plus cette mesure est petite, plus les sujets sont rapprochés, semblables, similaires. D'après Larocque (2016) la majorité des mesures de dissemblances, d, respectent les critères suivants :

1. $d(S_i, S_j) \geq 0$
2. $d(S_i, S_i) = 0$
3. $d(S_i, S_j) = d(S_j, S_i)$
4. $d(S_i, S_j)$ qui augmente au fur et à mesure que les deux sujets deviennent différents.

Avec les observations suivantes :

— $(X_{11}, X_{12}, ..., X_{1p})$: valeurs des p variables pour le sujet 1, S_1,

— $(X_{21}, X_{22}, ..., X_{2p})$: valeurs des p variables pour le sujet 2, S_2,

— ...

— $(X_{n1}, X_{n2}, ..., X_{np})$: valeurs des p variables pour le sujet n, S_n.

Ainsi, selon la notation ci-dessus, X_{nj} est donc la valeur de la j$^{\text{ème}}$ variable pour le n$^{\text{ème}}$ sujet.

Pour les variables quantitatives, 2 principales mesures de dissemblance existent :

1. **La distance euclidienne ou distance euclidienne au carré** : La distance euclidienne entre les sujets i et j est :

$$d(S_i, S_j) = \sqrt{(X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2 + ... + (X_{ip} - X_{jp})^2}$$

avec p dimensions possibles. En d'autres termes, il s'agit de la distance calculée de la ligne droite entre deux points dans l'espace euclidien. Cette distance prend en compte toutes les variables et n'élimine pas la redondance de l'information, en d'autres termes, si 3 variables expliquent la même chose (qui sont donc corrélées), alors le poids accordé à cet effet sera 3 fois plus élevé.

2. **La distance "City-block" ou Manhattan** : la distance entre les sujets S_i,S_j est :

$$d(S_i, S_j) = |X_{i1} - X_{j1}| + |X_{i2} - X_{j2}| + ... + |X_{ip} - X_{jp}|$$

Mesure de dissemblance entre deux groupes (inter-classe)

Comme expliqué, les groupes avec la plus faible distance sont combinés à chaque itération. Il faut donc être en mesure de pouvoir calculer la distance entre ces deux groupes.

La dissimilarité de deux groupes C_1=x, C_2=y contenant chacune un individu notés x et y (par fin de simplification de formules) se définit par la dissimilarité entre ses individus : *dissim*(C_1,C_2)=*dissim*(x,y).
Rencher Rencher (2002) résume les méthodes de dissemblance suivantes pour calculer la distance :

— **Mesure de dissemblance maximale** (plus communément appelé *Maximum ou Complete-linkage clustering*) : il s'agit de la dissimilarité entre les individus de C_1 et C_2 les plus éloignés :

$$dissim(C_1, C_2) = \max_{x \in C_1, y \in C_2} (dissim(x, y)).$$

— **Mesure de dissemblance minimale** (ou *Minimum ou Single-linkage clustering*) : il s'agit de la distance minimale entre les individus C_1 et C_2

$$dissim(C_1, C_2) = \min_{x \in C_1, y \in C_2} (dissim(x, y)).$$

— **Mesure de dissemblance moyenne** (ou *Mean ou Average Linkage Clustering, aka UPGMA*) : celle-ci calcule la moyenne des distances entre les individus C_1 et C_2

$$dissim(C_1, C_2) = \bar{X}_{x \in C_1, y \in C_2}(dissim(x, y)).$$

— **La méthode centroïde** (ou *Centroid linkage clustering, aka UPGMC*) : celle-ci joint les groupes ayant la plus petite distance en remplaçant tous les individus du groupe par le centroïde du groupe noté c. Ce centroïde est considéré comme un objet unique à la prochaine étape de regroupement :
$\|c_s - c_t\|$, avec c_s et c_t les centroïdes des groupes s et t.

— **La méthode de Ward** : Enfin, cette méthode vise à maximiser l'inertie inter-groupe, c'est à dire la séparation entre les groupes :

$$dissim(C_1, C_2) = \frac{n_1 n_2}{n_1 + n_2} dissim(G_x, G_y),$$

avec n_1 et n_2 les effectifs des deux groupes, G_x et G_y leurs centres de gravité respectifs. G_x est le point de coordonnées pour l'individu x ($\bar{x}_{1,1}$, ..., $\bar{x}_{n,x}$), où $\bar{x}_{j,x}$ (pour tout $j \in (1, ..., n)$) désigne la moyenne des valeurs observées de X_j sur les n_x individus du groupe x. De même pour G_y.

La méthode de Ward a été introduite par Ward Jr (1963) et prend en considération non seulement la distance intra-groupe lors de la formation des groupes mais aussi la distance inter-groupe. Ward stipule que non seulement les distances entre les groupes devraient être maximisées, mais que les distances entre les individus à l'intérieur de chaque groupe devraient également être réduites au minimum.
Dans ce sens, Jain and Dubes (1988) proposent tout un raisonnement pour démontrer que la méthode de Ward surpasse les autres méthodes présentées ci-dessus en terme de bonne représentation des groupes formés avec une méthode hiérarchique.

Détermination du nombre de groupes

Sugar and James (2003) commencent leur article avec cette phrase :

> Un problème fondamental et largement non résolu dans l'analyse de groupes est la détermination du « vrai » nombre de groupes dans un ensemble de données.*(traduction libre)*

De nombreuses approches ont abordé ce problème. Milligan and Cooper (1985) et Hardy (1996) ont mis en place une documentation détaillée des différentes références possibles pour répondre à ce problème. Des exemples dans la littérature statistique incluent l'index Caliński and Harabasz (1974), la règle Hartigan (1975), le test Krzanowski and Lai (1988), la statistique Silhouette (Kaufman and Rousseeuw (2009)) pour n'en citer que quelques-uns.

Cependant, comme le souligne Sugar and James (2003), bon nombre des approches qui ont été suggérées pour choisir le nombre de groupes ont été élaborées pour des problèmes spécifiques. Les méthodes généralement applicables ont tendance à être fondées sur des modèles précis et nécessitent donc des hypothèses paramétriques fortes et/ou beaucoup de calculs.

En d'autres termes, tous les index proposés dans la littérature doivent servir de guide à l'analyste. Ce dernier doit s'assurer que le nombre proposé par ces index donne des groupes interprétables et il est libre de modifier ce nombre selon ce qui a le plus de sens pour l'étude.

Limites des méthodes hiérarchique et non-hiérarchique

Comme l'expliquent Gil-Garcia et al. (2006), la méthode hiérarchique a comme particularité d'offrir une vision des données à différents niveaux d'abstraction, ce qui la rend idéale pour visualiser et explorer de manière interactive notre jeu de données.

La méthode hiérarchique ascendante permet de donner des groupes de meilleure qualité et plus détaillés que ceux produits avec une méthode non-hiérarchique. En

effet, Sousa and Gama (2014) affirment qu'en analyse de données, l'algorithme de hiérarchie ascendante est un outil puissant pour identifier des groupes sans besoin d'avoir de l'information préalable concernant la structure des données, ce qui lui permet de s'adapter aussi bien aux petits qu'aux gros jeux de données. De plus, cet algorithme est souvent utilisé car il fournit une représentation graphique des partitions résultantes, présentée par le dendrogramme, à l'inverse des méthodes non hiérarchiques qui renvoient qu'une partition unique.
De plus, comme nous l'avons vu dans la présentation de la méthode hiérarchique, le nombre de groupes n'est pas à spécifier au départ. En d'autres termes, la découpe du dendrogramme à différents niveaux de la hiérarchie produit différentes partitions et donc différents groupes.

Cependant, face à de larges bases de données, le temps de calcul de la méthode hiérarchique croît de manière quadratique : il faut calculer au moins n x n coefficients de dissemblances (n étant le nombre total de sujets à regrouper) en plus de les mettre à jour lors du processus de regroupement.

A l'inverse, la réputation de la méthode des K-moyennes vient essentiellement de sa simplicité à mettre en place ainsi qu'un temps de calcul relativement plus faible que celui de la méthode hiérarchique. Ainsi, vu la tendance actuelle du *Data Mining* elle est souvent la première méthode testée lors de ces études.
Cependant, comme nous l'avons vu dans la section 2.2.2, *L'algorithme des K-moyennes*, les groupes trouvés sont très sensibles à la détermination des centres initiaux au départ de l'algorithme. Ainsi, l'interprétation finale des groupes peut grandement différer d'un choix de centres initiaux à un autre.

Une approche hybride K-moyennes avec la méthode ascendante hiérarchique

Steinbach et al. (2000) affirment que la méthode hiérarchique est souvent décrite comme une approche de meilleure qualité mais limitée en raison de sa complexité quadratique. À l'inverse, la méthode des K-moyennes est plus rapide a exécuter mais offre des groupes de moins bonne qualité que la méthode hiérarchique à cause de l'importante sensibilité des groupes face au choix des centres initiaux. En effet, pour appuyer ce dernier point, si les centres initiaux changent à chaque fois que l'on roule l'algorithme, les groupes trouvés par la méthode des K-moyennes diffèrent d'une fois à l'autre.

Chen et al. (2005) proposera un nouvel algorithme : *Hybrid Hierarchical K-Means* (HHK). Ce dernier le décrit comme suit :
Tout d'abord, nous effectuons un regroupement hiérarchique sur un pourcentage de la table de données. Nous définissons ce pourcentage en nous basant sur l'ensemble du processus de regroupement effectué par la méthode hiérarchique. Par exemple, nous pouvons effectuer une méthode hiérarchique sur 70% des données. Les groupes générés à partir de la méthode hiérarchique servent à calculer la valeur moyenne de chaque groupe qui représente le centre initial de chaque groupe pour la méthode des K-moyennes.
En outre, le nombre de groupes générés à partir de la méthode hiérarchique sert de valeur de départ pour le nombre de groupes de la méthode des K-moyennes.
Ensuite, nous travaillons sur la méthode des K-moyennes sur l'ensemble des données, où chaque groupe doit au moins contenir les mêmes sujets générés qu'avec la méthode hiérarchique. En effet, la méthode hiérarchique place les sujets très proches des uns des autres dans des groupes, et le but de la formation des groupes avec la méthode des K-moyennes est de regrouper des objets proches de la même manière que la méthode hiérarchique. Par conséquent, les groupes finaux obtenus avec la

méthode des K-moyennes sont relativement proches de ce que nous pourrions avoir avec la méthode hiérarchique.

Grâce à cela, nous évitons à la fois le problème du temps de calcul qui peut être assez long si le jeu de données est grand ainsi que les problèmes de la détermination au départ du nombre de groupes voulus, la mise en place aléatoire des centres initiaux pour l'algorithme des K-moyennes ainsi que d'assurer une bonne qualité des groupes trouvés (Kassambara (2017)).

Cette approche est surtout utilisée dans le domaine de la classification des documents où le document est classé selon les mots qui le constitue (Cutting et al. (2017)).

2.2 Réduction de la dimensionnalité

Il n'est pas rare qu'une technique de réduction de dimensionnalité soit employée avant de procéder à l'algorithme de regroupement.

Comme nous le verrons dans le chapitre 3 à la section 3.2, *Catégorisation des actions prises par le joueur*, nous avons une base de données qui possède de très nombreuses variables. Ainsi, il est difficile d'appréhender globalement l'information contenue ou encore d'en déduire les relations statistiques entre diverses caractéristiques. Par exemple, une liste non-exhaustive des variables présentes dans notre base de données :

— le nombre de morts,

— les positions des joueurs dans le jeu,

— le nom de l'arme utilisée pour tuer un ennemi,

— les objets ramassés par le joueur,

— le ratio d'ennemis tués en fonction des morts du joueurs,

— le temps passé dans les quêtes,

— le nombre de mercenaires utilisés,

— etc.

L'introduction de très nombreuses variables dans une base de données amène un problème assez important dans l'analyse par regroupement : travailler avec une grande dimensionnalité. Ce problème rend difficile l'interprétation graphique des résultats d'analyse de regroupement.

L'Analyse par Composantes Principales (ACP)

L'ACP est une méthode d'analyse exploratoire des données. Elle est généralement utilisée lorsqu'on essaye d'interpréter de gros jeux de données qui possèdent de nombreux sujets ainsi que de très nombreuses variables. En effet, elle produit des facteurs (ou axes principaux) qui sont des combinaisons linéaires des variables de base. Chaque axe produit est indépendant des autres et explique une partie de la variance totale de l'information. Généralement, les premiers axes concentrent l'essentiel de l'information, ce qui facilite l'analyse.
Dans l'approche exploratoire, l'ACP est principalement utilisée dans deux cas :

— **La réduction de dimentionnalité** : Hair and Black (1998) affirment qu'elle permet de réduire un grand nombre de variables en un jeu de données plus petit et plus maniable ;

— **L'exploration des données** : Watters (2008) indique qu'elle peut identifier les variables latentes dans les grands ensembles de données représentés par des variables d'entrée fortement corrélées.

Fonctionnement de l'ACP

L'ACP fonctionne tel quel : Si on dispose de p variables

$$X_1, X_2, ..., X_p,$$

l'ACP cherche à définir p nouvelles variables, qui sont des combinaisons linéaires de p variables originales et qui feront perdre le moins d'informations possible :

$$C_1 = w_{11}X_1 + w_{12}X_2 + ... + w_{1p}X_p$$

$$C_2 = w_{21}X_1 + w_{22}X_2 + ... + w_{2p}X_p$$

$$...$$

$$C_p = w_{n1}X_1 + w_{n2}X_2 + ... + w_{np}X_p.$$

Où la somme des variances des p composantes principales est égale à la somme des variances des p variables de base.

Les deux principales propriétés des composantes principales sont :

- La variance d'une composante principale est égale à l'inertie portée par l'axe principal qui lui est associé ;
- Les composantes principales sont non corrélées deux à deux, car ce sont des informations et des organisations spatiales de nature bien différentes que l'analyse fait ressortir pour chaque axes.

La première composante principale C_1 correspond à la composante qui capte la plus grande partie de la variance totale. La deuxième composante principale possède la plus grande variance parmi toutes les combinaisons linéaires qui ne sont pas corrélées à C_1.

Ces deux premiers axes sont donc ceux qui restituent le maximum d'informations, ils sont identifiés automatiquement par les logiciels d'analyse de données.

Ainsi, tout l'intérêt de l'ACP se fait voir : comme l'explique Larocque (2016), les composantes principales forment un nouvel ensemble de variables non corrélées entre

elles qui récupèrent en ordre décroissant le plus de variance possible des variables originales. En d'autres termes, un petit nombre de composantes principales arrive à expliquer la plus grande partie de la variance totale.
Un aspect important de l'ACP a considérer est la rotation utilisée. Comme nous l'avons vu, l'ACP permet de regrouper les variables en dimensions (ou facteurs) afin de pouvoir expliquer la plus grande partie de la variance totale. La rotation consiste à faire pivoter virtuellement les axes des facteurs autour du point d'origine dans le but de redistribuer plus équitablement la variance à expliquer.
Cependant, pour mieux interpréter la structure reliant les groupes trouvés à l'aide des composantes, la rotation montre son utilité : le but ultime de la rotation est toujours de simplifier la lecture des poids des variables sur les facteurs (voir 2.5 ci-dessous).
En effet, comme les facteurs sont d'abord extraits selon leur importance, l'ACP tend à produire un premier facteur général difficile à interpréter puisqu'il regroupe un grand nombre d'items (Pett et al. (2003)). Par conséquent, le chercheur a généralement recours à une rotation des axes afin de faciliter l'interprétation de la solution factorielle (Kieffer (1998)). Il s'agit alors de faire pivoter les axes autour de l'origine afin d'obtenir un ajustement optimal de la distribution empirique des données.

La rotation Varimax est l'une des rotations les plus utilisées. Elle consiste à associer chacune des variables à un nombre réduit de facteurs et à représenter chaque facteur par un nombre limité de variables. Visuellement, les variables sont rapprochées des axes auxquels elles contribuent de manière à en faciliter l'interprétation.
Ainsi, l'application de la rotation Varimax aide à identifier la contribution des variables à la formation des axes factoriels. Ceci permet de tirer, d'une manière rapide et synthétique, des conclusions sur les dimensionnalités des variables, évitant tout biais lié à la qualité de la projection et à la synthèse des données.

Un autre aspect important est le nombre de composantes principales à retenir. En effet, Ferre (1995) insiste sur le fait que si le bon nombre n'est pas retenu pour l'ana-

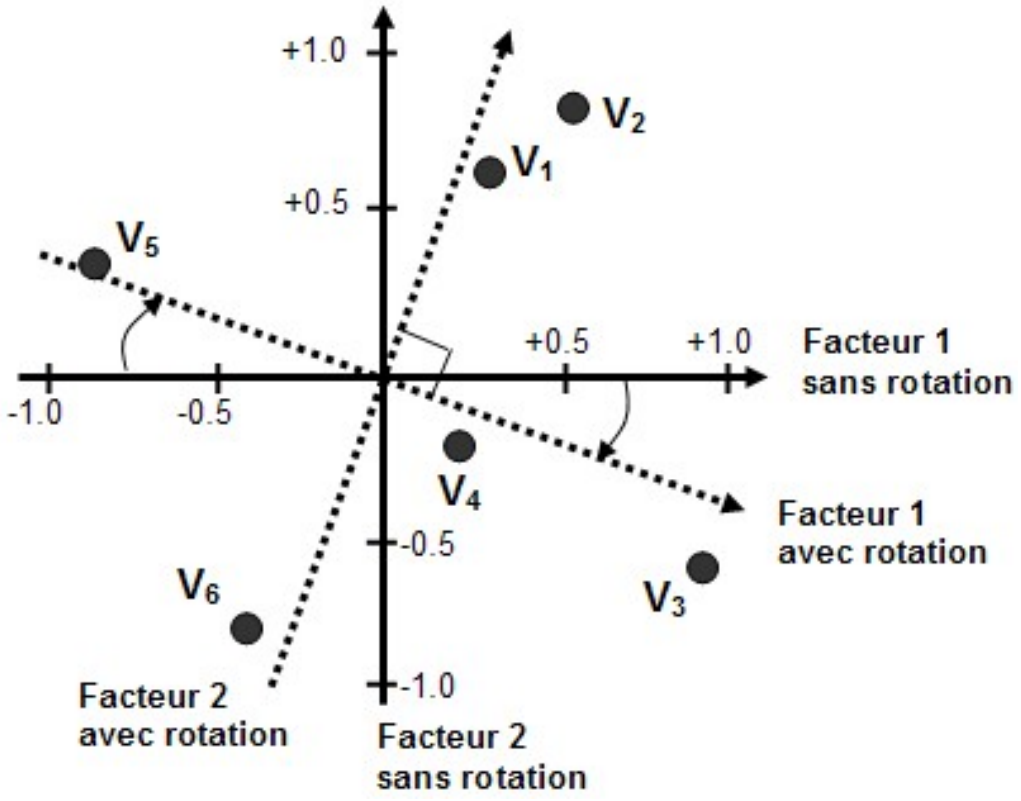


FIGURE 2.5 – Exemple d'une représentation d'une rotation orthogonale à 6 variables notées *V*. Comme les variables sont rapprochées des axes, il est plus facile de les interpréter en évitant notamment tout biais lié à la qualité de projection (Eric Yergeau et Martine Poirier).

lyse subséquente, soit l'information pertinente est perdue (sous-estimation), soit le bruit est inclus (surestimation), causant une distorsion dans les modèles sous-jacents de variation/covariation. Ainsi, depuis la fin des années 70, de nombreux auteurs comme Pimentel (1979), Karr (1981), Stauffer (1985), Jackson (1991) ou encore Jolliffe (2002) s'entendent pour dire qu'un des problème de longue date dans la littérature biologique et statistique est le suivant : la détermination du nombre idéal de composantes principales reste un des plus grands défis pour fournir une interprétation significative des données. En pratique, il est généralement accepté que les dimensions qui ont une valeur propre supérieure à 1 sont à retenir dans notre modèle.

2.3 Modèles de regroupement par séquence

Pour rappel, l'un des objectifs de notre problématique est de connaître l'évolution du style de jeu des joueurs au cours du temps. En effet, cela permettrait de mieux comprendre comment les joueurs s'approprient le jeu et jouent le contenu disponible. Notre problématique aborde donc l'aspect temporel afin de voir une évolution éventuelle du style des joueurs au cours du temps. Ainsi, la notion d'analyse par séquence est à prendre en considération dans notre modèle afin de pouvoir noter l'évolution du style de jeu.

Le modèle de Markov caché (HMMs), développé par Baum and Petrie (1966) a comme principal avantage de pouvoir modéliser l'évolution des observations à travers le temps. Smyth (1997) utilisera ce procédé dans le but de construire un modèle descriptif plutôt que prédictif des données dans le domaine de la génétique. Cependant, l'utilisation du procédé de Smyth (1997) demande de déterminer au départ le nombre K groupes que nous souhaitons avoir. Or, dans le cadre de notre étude, nous ne connaissons pas le nombre de styles de jeu qui paraissent dans notre jeu, ce qui veut dire que le nombre d'états est inconnu. De plus, chaque état peut être associé à plusieurs observations. Ainsi, un modèle HMMs ne serait pas la méthode la plus adaptée dans notre cas.

2.4 La classification des joueurs dans les jeux vidéo

Plus spécifiquement à notre étude, des travaux essentiellement basés sur des modèles de psychologie et des questionnaires ont permis de modéliser des styles de jeu.

Bontchev and Georgieva (2018) proposent une manière de reconnaître le style de jeu dans un jeu vidéo spécifique, *Honey and Mumford*. Leur modèle se base sur la théorie de l'apprentissage expérientiel (ELT) proposée par Kolb (2012), qui focalise

sur l'expérience, la perception, la cognition et le comportement dans le jeu. Pour former les groupes de style d'apprentissages des joueurs, Kolb (2012) se base d'abord sur les 4 groupes :

— Les accommodateurs : ces joueurs aiment essayer des choses et expérimenter. Ils vont probablement développer une perception intuitive de ce qui est juste et l'utiliser comme base pour leurs décisions.

— Les assimilateurs : ces joueurs aiment comprendre. Ils vont souvent relier l'expérience à des expériences passées.

— Les divergeants : ces joueurs aiment regarder. Ils vont surtout regarder des expériences et les internaliser ou passer du temps à réfléchir pour analyser l'expérience regardée.

— Les convergeants : ces joueurs aiment monter des modèles. Ils formeront un modèle du problème dans leur tête et l'expérimenter jusqu'à ce qu'ils trouvent la bonne solution.

Après avoir observé les joueurs pour les catégoriser parmi l'un des quatre groupes, un test sur papier (nommé LSI) est donné aux joueurs pour qu'ils évaluent leur style d'apprentissage. Cela produit des scores distincts, des agrégations de scores qui vont être comparés aux observations faites précédemment et donc déterminer leur style d'apprentissage.

Ainsi, Bontchev and Georgieva (2018) utilisent cette théorie pour l'adapter aux styles de jeu. En effet, au lieu de parler de 4 styles d'apprentissage, 4 styles de jeu ont été identifiés par les auteurs : les joueurs compétiteurs, rêveurs, logiques et stratégiques. Grâce à l'observation, des données de performance concernant chaque type de groupes ont été donnés ainsi qu'à travers des coefficients trouvés de manière heuristique.
Cooley (2015) applique aussi la théorie de Kolb pour détecter les styles d'apprentissages dans les jeux vidéo.

Un des rares ouvrages concernant une étude approfondie des styles de jeu est celui de Bartle (1996) pour le jeu MUD. Le joueur incarne un personnage et voit des descriptions textuelles de salles, d'objets ou d'autres personnages dans un monde virtuel. Les joueurs peuvent interagir entre eux et avec l'environnement en tapant des commandes qui ressemblent au langage courant. Les groupes identifiés relèvent donc des discussions qui font partie du style de jeu du joueur. L'approche consiste à examiner en détail les discussions ainsi que les interactions avec l'environnement pour pouvoir répartir les joueurs en 4 groupes :

— Les tueurs : ils représentent les joueurs qui sont plus agressifs, qui attaquent d'autres personnages dans le but de les tuer.

— Les sociables : ces joueurs s'intéressent aux gens et à ce qu'ils ont à dire. Les relations inter-joueurs sont importantes pour eux : notamment l'empathie, la sympathie, la rigolade et le divertissement entre joueurs.

— Les explorateurs : ces joueurs aiment découvrir et inspecter le monde du jeu, trouver des éventuels *bugs*, des zones où peu de joueurs vont ainsi que savoir comment le jeu fonctionne en fonction de ses actions.

— Les compléteurs : le but principal de ces joueurs est de gagner de l'expérience et des niveaux dans le jeu. Toutes les actions qu'ils mèneront auront comme but de rapporter le plus de points ou d'expérience possible.

Le graphique 2.6 ci-dessous explique la répartition de ces joueurs selon les 4 groupes énoncés.

Toutes ces études abordées utilisent donc l'observation accrue du comportement des joueurs, l'utilisation de questionnaires et/ou d'entrevues adressés à ces derniers pour pouvoir déterminer les styles de jeu.
Notre approche diffère de ces études car nous utilisons exclusivement des données *in-game* fournies par la télémétrie. Cette méthode permet de ne pas nuire à l'expérience

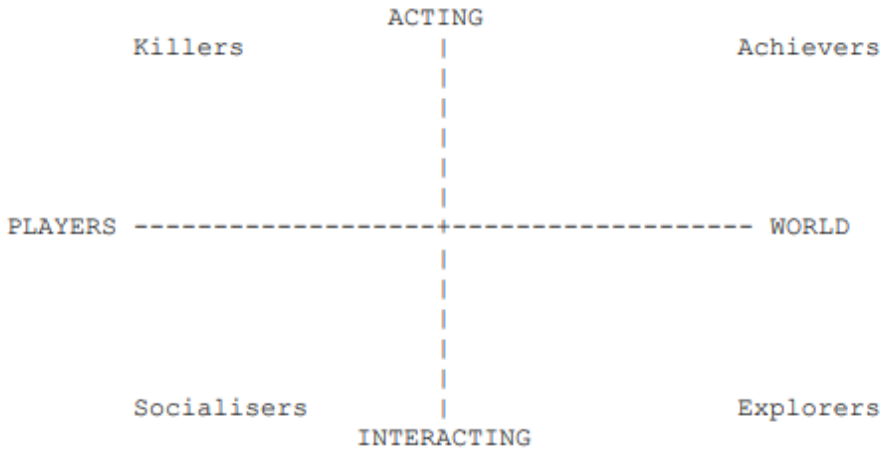


FIGURE 2.6 – Représentation du graphique de classification des joueurs de MUD (Bartle (1996)).

du joueur et surtout de pouvoir étudier de manière très détaillée son comportement dans le jeu.

2.5 Synthèse de la Revue de Littérature

Comme nous venons de le voir, la littérature concernant les styles de jeu dans les jeux vidéo est peu exhaustive. En effet, la plupart des études vues intègrent l'observation ainsi que la participation active du joueur dans l'étude. Or, dans le cadre de tests utilisateurs, il est important que l'analyste soit le plus passif possible ainsi que de minimiser l'interaction avec le joueur lorsqu'il joue afin de ne pas nuire à son expérience de jeu.

La littérature est globalement riche en méthodes de regroupement et ce, dans des domaines très variés. En effet, l'analyse par regroupement peut aussi bien servir comme méthode préliminaire, lors du travail de préparation de données dans le cadre de prédiction, ou encore comme objectif final si le mieux est de mieux comprendre sa base de données actuelle.

Nous avons vu dans l'apprentissage non-supervisé, deux principales méthodes de regroupement : les méthodes hiérarchiques et non-hiérarchiques.

L'algorithme le plus en vogue dans la littérature pour la méthode non-hiérarchique est la méthode des K-moyennes. En effet, comme déjà expliqué, à l'air du *Big Data* et du *Data Mining*, sa facilité à mettre en place ainsi que sa rapidité d'exécution en fait un algorithme extrêmement intéressant à utiliser à titre exploratoire et donne des groupes souvent interprétables : il n'y a presque pas de paramètres à mettre en place si ce n'est déterminer les centres initiaux ainsi que le nombre de groupes voulus. Ces deux derniers points peuvent être vus comme des handicaps : la détermination des centres initiaux reste une étape délicate à traiter car c'est de là que tous les groupes vont se former. Ainsi, les résultats sont sensibles au choix des centres initiaux et peuvent changer d'un choix à un autre. De plus, la détermination du nombre de groupes voulus au départ demande à l'analyste une certaine connaissance de ses joueurs et des groupes non identifiés peuvent ne pas être pris en considération à cause du nombre prédéfini.

Concernant les méthode hiérarchiques, la méthode ascendante hiérarchique est intéressante à exploiter lorsque nous recherchons plus de finesse dans notre analyse, plus de détail sur les groupes trouvés. En effet, de part la composition de l'algorithme qui débute avec autant de groupes qu'il y a d'individus, permet de donner des groupes de plus grande qualité.
Cependant, cela a un coût considérable qui est le temps de calcul de l'algorithme. Ainsi, il est difficile de bien l'implémenter sur de gros jeux de données tout en étant très performant sur les petites bases de données.

Face à ces avantages intéressants et désavantages contraignants que ces algorithmes offrent, la littérature s'est tournée vers la proposition d'un algorithme hybride qui

propose d'effectuer dans un premier temps une méthode hiérarchique ascendante sur la moitié des données et utiliser ces résultats pour générer les centres initiaux pour la deuxième partie des données qui est faite avec une méthode des K-moyennes. Cette technique permet donc un temps de calcul moindre tout en assurant une bonne qualité des groupes donnés.

Toutes les études vues ne montraient pas d'application concrète en entreprise. Or, la recherche en entreprise montre des réalités bien différentes de celles rencontrées dans le cadre de la recherche universitaire. Par exemple, en entreprise, il est préférable d'avoir des modèles faciles à implémenter afin de pouvoir les utiliser sur d'autres projets. Enfin, il est important de pouvoir vulgariser ses résultats afin de pouvoir divulguer l'information pertinente à des équipes non habituées aux statistiques. Les représentations graphiques des résultats restent encore aujourd'hui un moyen non négligeable pour véhiculer de la bonne manière l'information aux équipes business.

Chapitre 3

Description des bases de données utilisées

Cette section se concentre sur la base de données à exploiter dans le cadre de l'étude. Cette dernière se base sur le jeu nouvellement produit par Ubisoft Montréal, *Far Cry 5*. L'histoire de ce jeu *FPS* est comme suit : Une secte apocalyptique fanatique du nom d'Eden's Gate s'est développée à Hope County dans le Montana. Le joueur doit donc s'élever contre le meneur de la secte en promouvant un mouvement de résistance afin de libérer la communauté assiégée. De nombreuses options sont proposées dans le jeu afin d'aider le joueur à vaincre le Père. La plus importante option étant le recrutement de *Guns For Hire* (sous l'acronyme *GFH*), ou mercenaires, ces derniers ont pour but de répondre aux ordres du joueur pour l'aider lors de ses missions.

Les données utilisées sont envoyées sur les serveurs d'Ubisoft toutes les 20 minutes de jeu. En d'autres termes, toutes les 20 minutes, nous recevons de nombreuses données quantitatives qui s'incrémentent dans le temps. Les données représentent des actions effectuées dans le jeu par le joueur.

3.1 Extraction et informations générales de la base de donnée

Les données concernent 4 tests utilisateurs réalisés dans les locaux d'Ubisoft avant la sortie du jeu *Far Cry 5*. Les objectifs de ces tests sont identiques : essayer de terminer le jeu tout essayant de reproduire son attitude de jeu à la maison.

Particularité de ces tests utilisateurs : Tests d'une durée variant entre 6 à 7 jours. Aucune contrainte imposée au participant concernant le jeu si ce n'est essayer d'aller le plus loin possible dans le jeu. Le but de ces tests était de s'assurer que l'expérience de jeu est optimale, c'est à dire avec le moins de points de friction possible (mauvaise compréhension de l'histoire, frustration dans les mécanismes de jeu, économie interne mal ajustée). Un joueur a terminé le test s'il a terminé le jeu ainsi que répondu à tous les questionnaires et entrevues demandés par l'analyste en Recherche Utilisateur.

Régions : Canada et France

Échantillon total : 60 participants.

Type de données : Données quantitatives qui s'incrémentent toutes les 20 minutes. Les données récoltées représentent des actions précises du joueur.

Caractérisation des données : Chaque joueur est identifié par un ID qui permet de le retrouver au cours du temps dans la base de données. En effet, cela nous permettra de pouvoir identifier l'évolution de ses actions dans le temps. Cet ID ne permet pas de fournir des informations personnelles sur le participant. Ainsi, aucun nom ou signe distinctif ne rattache cet ID au participant : il est donc impossible de savoir qui a généré cette donnée en jouant. La base de donnée a été retravaillée afin

d'être exploitée pour cette étude, cela sera abordé dans le Chapitre 4 section 4.2. Enfin, chaque variable représente un décompte d'une action précise du joueur (qui est incrémentée toute les 20 minutes) qui permet de répondre à des questions sur :

— L'économie au sein du jeu : Avec par exemple, le nombre de fois que le joueur a acheté une arme ou encore le montant d'argent qu'il possède dans le jeu.

— La difficulté au sein du jeu : Grâce au nombre de morts du joueur enregistrées, le temps passé dans les missions, le nombre de fois qu'un joueur n'a pas réussi à terminer une mission.

— Les activités effectuées : Par exemple le nombre de missions secondaires, le nombre de poissons pêchés, le nombre d'endroits découverts dans le jeu.

— L'utilisation des armes et des véhicules : Le nombre d'armes utilisées par catégorie, le nombre de voitures, quad, bateaux, avions, motos, hélicoptères utilisés.

— La progression dans le jeu : L'utilisation des points de compétence ou encore l'avancée dans la quête principale

En d'autres termes, les données disponibles dans notre base de données représentent 2 principales caractéristiques :

— **Les aspects propres au combat** : deux principaux styles reviennent : l'assaut et le style furtif. Un joueur de type assaut va préférer les armes à courte distance, c'est à dire l'utilisation des pistolets, des fusils. De plus, de part la courte distance, les armes de mêlée sont grandement utilisées comme les poings américains. A l'inverse, le joueur style furtif privilégiera les approches lointaines afin de ne pas se faire repérer, en utilisant des armes à longue portée comme les snipers ou toute arme suffisamment précise pour faire des ravages à longue distance. Aussi, ce type de joueur apprécie les approches furtives proches des ennemis en les tuant par surprise : les *takedowns*.

— **Les aspects propres à Far Cry 5** : le jeu a des particularités que nous ne retrouvons pas dans d'autres jeux. L'aspect le plus important est le rôle des missionnaires : les *Guns for Hire* (ou mercenaires). Ceux-ci ont pour but

d'assister le joueur. Le joueur les contrôle en donnant des ordres tels que : suivre le joueur, aller à un endroit précis, ranimer le joueur si ce dernier n'a plus de vie ou encore attaquer un ennemi. Cette diversification d'actions permet au joueur d'enrichir ses stratégies de style de jeu. Au delà des mercenaires, tout l'environnement de Far Cry 5 permet au joueur de mieux spécifier ses stratégies de jeu. En effet, si ce dernier effectue des challenges, il pourra ainsi acquérir plus rapidement des points qu'il pourra échanger contre des compétences comme devenir plus silencieux proche des ennemis (bon pour les types style furtif), infliger plus de dégâts lors d'un combat (intéressant pour le joueur de type assaut) ou encore améliorer les compétences de ses mercenaires. L'exploitation du jeu passe aussi par l'exploitation du Monde ouvert que Far Cry 5 offre au joueur. En effet, l'histoire permet de :

— Sauver des otages : des civils que le Culte force à se convertir,

— Libérer des *Outposts* : libérer des camps détenus par le Culte. Les acquérir permet de réduire la présence du culte,

— Effectuer des *Treasure Hunts* : il s'agit de missions annexes que les NPC donnent au joueur. Il s'agit souvent d'une *chasse au trésor* où le joueur doit trouver un objet en particulier dans un lieu plus ou moins précis,

— Exploration du Monde Ouvert grâce à la pêche, la récupération d'objets, etc.

3.2 Catégorisation des actions prises par le joueur

La base de donnée se compose comme-ci :

Nombre de participants	60 participants
Nombre de variables de base	85 variables
Nombre de variables transformées	43 variables
Nombre total de variables	128 variables

Chaque ligne correspond au total des actions effectuées par le joueur lors des 20 dernières minutes de jeu. A chaque fois que l'évènement est envoyé, les données sont incrémentées au cours du temps. Comme expliqué dans la section *Extraction et Informations générales de la base de données (section 3.1)*, toutes les variables sont des données quantitatives de type dénombrable.

Voici un exemple ci-dessous de l'incrémentation à 20 minutes d'écart des données quantitatives pour un même joueur sur cette table :

ID	XTime	X1	X2	...	X85	...	X128
47	5700	23	221	...	12	...	2
47	6900	157	234	...	38	...	15

Dans le cadre de notre étude, nous souhaitons voir l'évolution du style de jeu du joueur au cours du temps et de pouvoir facilement la représenter graphiquement. La méthode la plus efficace pour étudier l'évolution au cours du temps est de diviser la base de données en intervalles de temps. Comme les données de notre table sont incrémentées automatiquement toutes les 20 minutes, il a été décidé Ainsi, la base de données a été découpée en 6 :

— **temps de jeu inférieur à 20 minutes de jeu :** Il est important de noter que nous avons délibérément choisi de ne pas inclure les 20 premières minutes de jeu. En effet, lors des 20 premières minutes de jeu, le joueur se trouve dans le tutoriel dynamique, c'est-à-dire qu'il est fortement guidé dans ses actions afin de pouvoir apprendre les mécanismes de jeu de *Far Cry 5* telles que l'utilité des contrôles de la manette, ce qu'il peut retrouver dans les différents menus disponibles, etc.

— **temps de jeu au bout de 40 minutes :** Il s'agit de la première véritable entrée de données dans notre table. De plus, lors de cette période de temps spécifique, le joueur est dans la partie du tutoriel dynamique, c'est à dire qu'il apprend le fonctionnement du *gameplay*.

— **temps de jeu au bout de 60 minutes :** Le joueur est censé avoir compris le fonctionnement du jeu et sort du tutoriel dynamique et se retrouve libre dans le Monde Ouvert. Ainsi, le joueur devient maître de son propre jeu.

A partir de 1 heure de jeu, la découpe de la base de données se fait heure par heure jusqu'à atteindre les 5 heures de jeu. En d'autres termes la découpe de la base de données prend cette allure :

— **temps de jeu au bout de 2 heures**

— **temps de jeu au bout de 3 heures**

— **temps de jeu au bout de 4 heures**

— **temps de jeu au bout de 5 heures**

Ainsi, nous devons effectuer une analyse de regroupement sur les 6 nouvelles bases de données qui se composent de cette manière-ci :

Capture des données dans le temps (minutes)	Nombre de participants	Nombre total de lignes
]20 ; 40]	45	45
]40 ; 60]	53	53
]60 ; 120]	60	60
]120 ; 180]	60	60
]180 ; 240]	60	60
]240 ; 300]	60	60

Le fait que moins de participants figurent dans la première heure de jeu est dû à l'absence de télémétrie pour ces joueurs.

3.3 Choix de la découpe des bases de données

Avant tout, il est important de noter que le découpage proposé est un choix arbitraire pour les raisons énoncées ci-dessous. Comme mentionné plus haut, les données de notre table sont incrémentées toutes les 20 minutes de jeu. Dans *Far Cry 5*, les vingt premières minutes de jeu servent à aider le joueur à prendre le jeu en main, c'est-à-dire de l'aider à comprendre les contrôles de la manette, les particularités du jeu (tel que l'utilisation des mercenaires qui est un élément très spécifique à *Far Cry 5*) et à naviguer dans les menus proposés. En plus de cela, les joueurs doivent suivre un chemin défini par le jeu afin de les obliger à passer par le tutoriel, ainsi les armes disponibles lors de cette période de temps sont identiques pour tous les joueurs. En outre, un tutoriel de combat est fait, donc tous les joueurs se voient jouer de la même manière afin de mieux comprendre les mécanismes de jeu. Enfin, pour la plupart des joueurs de notre étude, une très grande portion des variables étaient toujours nulles au bout des 20 premières minutes de jeu : par exemple, il est rare que le joueur rencontre un mercenaire à recruter dans cette période-ci, ou encore, à cause du peu d'armes rencontrées, ce dernier n'utilise pas certains types d'armes importants pour la détermination du style de jeu. Comme le joueur n'est pas entièrement libre dans ses actions au cours des 20 premières minutes en plus de faire peu d'actions spécifiques pour déterminer un style de jeu, il a été décidé de ne pas s'intéresser à cette période de temps.

Comme les données incrémentent toutes les 20 minutes, le premier découpage commence avec les données récoltées au bout de 40 minutes. Dans le jeu *Far Cry 5*,

les joueurs les plus rapides peuvent terminer le tutoriel dynamique au bout de 40 minutes. Même si le tutoriel force les joueurs à jouer d'une certaine manière pour qu'ils apprennent les mécanismes du jeu, nous nous sommes tout de même intéressé aux styles de jeu qui ont prédominés lors du tutoriel.

Il était aussi intéressant d'étudier les éventuels styles de jeu au bout d'une heure de jeu. En effet au bout d'une heure, tous les joueurs doivent avoir terminé le tutoriel proposé. En d'autres termes, il s'agit du moment où les joueurs se retrouvent libres de jouer de la manière dont il leur plaît ainsi que de pouvoir profiter du Monde Ouvert.

Comme le joueur est libre de faire ce qu'il souhaite au bout d'une heure de jeu, il est difficile d'avoir un point de repère fixe à tous les joueurs pour pouvoir étudier leur style de jeu. Par exemple, un joueur peut décider d'effectuer tout de suite la quête principale alors qu'une autre sera intéressé à faire plutôt une quête annexe. Ainsi, le temps devient le seul point de repère qui est identique d'un joueur à un autre. En se basant sur un intervalle d'une heure de jeu, cela permet d'être certain que chaque joueur a eu le temps de faire suffisamment de contenu intéressant pour étudier une éventuelle évolution du style de son jeu.

Nous arrêtons notre analyse au bout de 5 heures de jeu car le joueur a eu le temps de s'approprier correctement les mécanismes de jeu proposés par le jeu, de rentrer dans l'histoire ainsi que de développer son style de jeu grâce aux nombreux combats qui lui sont proposés tout au long de ces 5 heures. De plus, le but de l'étude est aussi de se concentrer sur la première expérience de jeu du joueur qui est cruciale pour l'envie du joueur de continuer à jouer ou non.

Chapitre 4

Descriptif de la méthode d'analyse

Cette section est consacrée à la description des étapes de notre analyse.

4.1 Environnement de programmation

Tout le processus a été développé sous *R*, plus particulièrement dans *RStudio* version 1.1.383. Toute la documentation des libraires utilisées ci-dessous existent sur *RDocumentation*.

Les principales librairies au code source ouvert utilisées pour cette étude sont :
FactoMineR : Il s'agit d'un package R complet et très utilisé notamment dans l'analyse exploratoire multidimensionnelle de données. De part les nombreuses possibilités d'analyses qu'offre ce package, il est donc possible de combiner des méthodes ensembles sans problème. Il a été développé par des français sur le campus d'Agrocampus Rennes. Nous l'avons utilisé à différentes fins :

— l'ACP : la fonction *PCA()* permet de mettre en place un ACP complète avec une présentation simple des dimensions obtenues, la représentation sur les axes ainsi que les valeurs propres assignées à chaque dimension.

— la méthode hybride impliquant une méthode ascendante hiérarchique suivi d'une méthode des K-Moyennes : la fonction HCPC() (Classification Hiérarchique sur Composantes Principales) permet de réaliser une classification non supervisée des individus. L'algorithme fonctionne de même : l'utilisateur choisi le nombre de composantes principales à retenir dans le modèle après avoir effectué dans un premier temps une ACP. Ensuite, une classification hiérarchique est effectuée en utilisant la méthode de Ward sur les composantes principales retenues. Ward est utilisé car c'est basé sur la variance multidimensionnelle tout comme l'ACP. Par la suite, l'utilisateur choisi le nombre de groupes à retenir basé sur le dendrogramme formé. Enfin, une méthode des K-Moyennes est effectuée pour améliorer le partitionnement initial trouvé avec la méthode hiérarchique. Le partitionnement final obtenu après consolidation avec la méthode des K-Moyennes peut être un peu différent de celui trouvé avec la méthode hiérarchique.

— la méthode ascendante hiérarchique : la librairie générale **stats** contient de nombreuses fonctions pour des calculs statistiques, elle permet notamment de faire une méthode hiérarchique ascendante classique. Ainsi la fonction **hclust()** a été utilisée pour mettre en place la deuxième méthode développée dans ce mémoire : la méthode hiérarchique ascendante.

corrplot : Cette librairie permet de faire des représentations graphiques de la matrice de corrélation, la rendant ainsi plus facile à interpréter.

NbClust : Ce package permet de déterminer le nombre idéal de groupes à retenir. Cette librairie se base sur 30 indices pour déterminer le nombre de groupes et propose à l'utilisateur le meilleur schéma de regroupement à partir des résultats des indices obtenus en faisant varier toutes les combinaisons de nombre de groupes, de mesures de distances et de méthode de regroupement possibles.

networkD3 : Cette librairie a été développée par Christopher Gandrud dans

le but de créer des réseaux 3D. Nous avons utilisé cela dans le but de faire un diagramme de Sankey avec la fonction **sankeyNetwork**. Cette représentation est très bien adaptée pour voir l'évolution du style du joueur au cours du temps.

4.2 Travail préliminaire sur les bases de données

Matrice de corrélation

Chaque base de données créée dans la section *3.2 Catégorisation des actions prises par le joueur* représente environ 60 individus et chacun d'entre eux sont caractérisés par 85 variables de base ainsi que 43 transformées que nous expliquerons un peu plus loin dans ce chapitre.
Dans le cadre de notre étude, comme le montre la figure 4.1, la plupart des variables de base sont fortement corrélées positivement avec le temps d'avancement au sein du jeu. En d'autres termes, plus on avance dans le temps, plus les variables positivement corrélées avec le temps augmentent de manière linéaire. Nous avons volontairement renommé les variables à des fins de confidentialité de données.

Si plusieurs variables sont corrélées entre elles, cela peut rendre difficile, voire impossible, l'interprétation des effets individuels de ces variables.

Afin de réduire la corrélation existante et pour faire ressortir les informations voulues dans le cadre de notre étude, de nombreuses transformations de variables ont été effectuées. Nous avons donc rajouté 29 variables transformées. Ces dernières représentent des ratios. De manière non exhaustive voici la liste de ces variables :

— Ratios de morts infligées par rapport aux morts du joueur : ce ratio permet de savoir si le joueur meurt plus qu'il ne tue d'ennemis.

— Ratios d'ennemis tués avec une arme spécifique (arme d'assaut, de type furtif ou de combat rapproché) par rapport au nombre total de victimes : Ce ra-

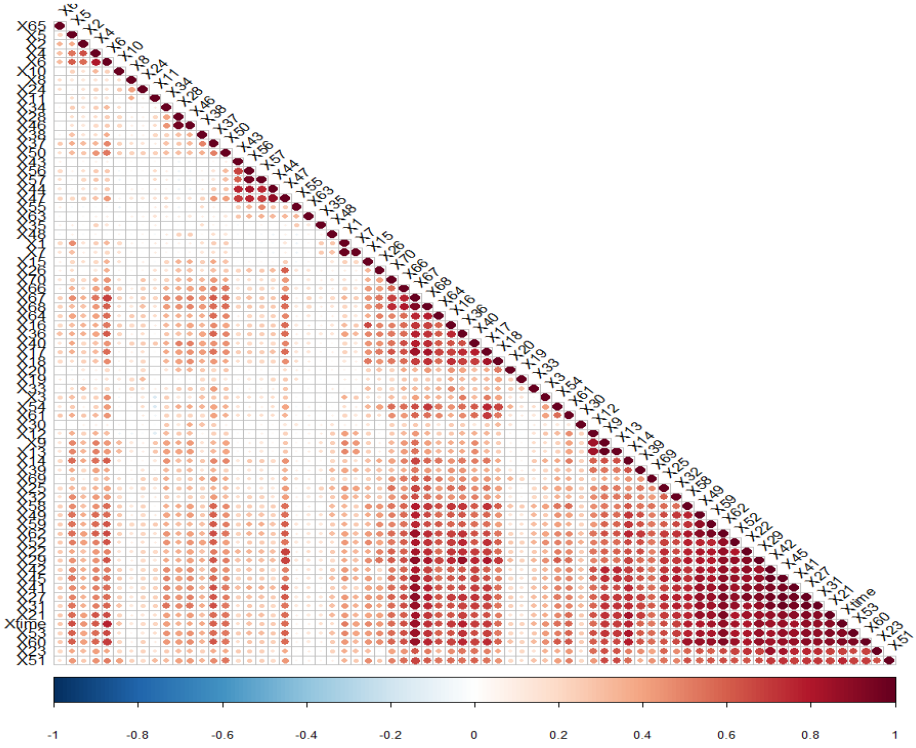


FIGURE 4.1 – Visualisation de la matrice de corrélation entre les variables de base

tio permet d'avoir une idée quel type d'arme le joueur préfère pour tuer les ennemis.

— Ratios d'ennemis tués par rapport au nombre d'ennemis tués par les mercenaires : Ce ratio donne plus d'informations sur l'utilité des mercenaires dans les stratégies de jeu du joueur. Par exemple un bon ratio montrerait que le joueur utilise ces mercenaires dans sa stratégie de jeu.

— Ratios de temps : Par exemple, le ratio de temps passé avec les mercenaires par rapport au temps total passé dans le jeu, cela permet ainsi d'avoir une idée de l'utilité des mercenaires pour le joueur. Un autre exemple serait le temps passé dans les activités de *Monde Ouvert* par rapport au temps total passé dans le jeu : cela permet de savoir à quel point le joueur passe du temps dans des activités spécifiques à *Far Cry 5* ou non.

La figure 4.2 montre la matrice de corrélation avec les nouvelles variables ajoutées.

Nous avons volontairement laissé la variable de temps *XTime*. Nous remarquons que la plupart des variables sont bien moins corrélées avec le temps de jeu. Les principales zones de variables fortement corrélées entre elles s'expliquent par le fait que les variables impliquées représentent presque la même information étudiée. En effet, en plus d'étudier les ratios spécifiques de chaque type d'armes disponibles dans le jeu, des ratios de regroupements de plusieurs type d'armes ont été faits afin d'avoir des ratios qui se démarquent plus par rapport au nombre de victimes faites au sein du jeu.

A l'inverse, les corrélations négatives montrent que plus on avance dans le temps, moins le joueur meurt dans le jeu. Ces corrélations sont logiques puisque plus le joueur joue, plus ce dernier maîtrise le jeu et par conséquent meurt moins.

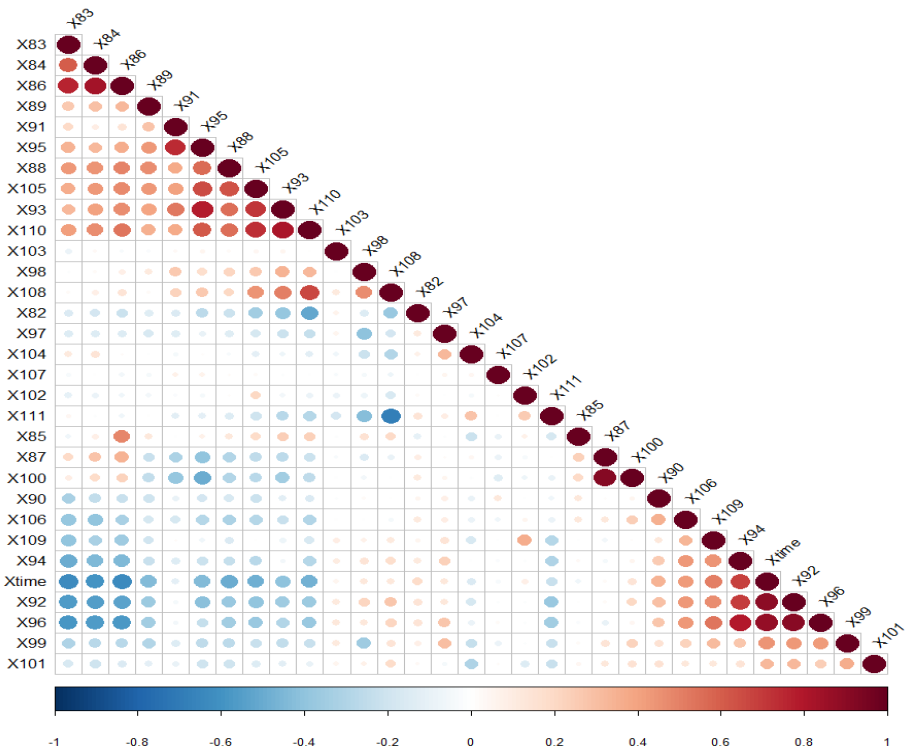


FIGURE 4.2 – Visualisation de la matrice de corrélation entre les variables transformées

4.3 Mise en place de l'ACP

Notre jeu de données retrace les activités effectuées par des joueurs lors de tests utilisateurs. En d'autres termes, il s'agit d'un nombre restreint de joueurs (60 joueurs) avec peu de lignes (60 lignes par base de données créée) mais énormément de variables disponibles (un total de 128 variables).

Comme Watters (2008) le mentionne , l'ACP peut aider à identifier les variables latentes dans les grands ensembles de données représentés par des variables fortement corrélées. De plus, J-M. Bouroche et G. Saporta (1985) expliquent aussi que l'ACP permet de réduire un grand nombre de variables en un jeu de données plus petit et plus maniable . De par ces 2 principales caractéristiques de l'ACP, cette

méthode sera utilisée comme méthode de réduction de la dimensionnalité de notre base de données et servira aussi de moyen de réduction du nombre de variables : en d'autres termes, on espère qu'un petit nombre de composantes principales réussira à expliquer la plus grande partie de la variance totale.

Comme nous pouvons le voir dans le tableau 4.1, la première composante n'arrive à capter que 12.9% de la variance totale, alors que cette dernière représente la composante qui doit saisir la plus grande variance expliquée.

De plus, un des moyens pour savoir combien de composantes retenir est de conserver toutes les composantes dont la valeur propre est supérieure à 1. Dans le cas de notre étude, il faut attendre la 28ème composante pour capter 91.8% de la variance totale.

TABLE 4.1 – Résultat des composantes de l'ACP

Composante	Valeur propre	% de la variance expliquée	% cumulé de la variance expliquée
comp 1	16.05801042	12.95000840	12.95001
comp 2	14.05735401	11.33657582	24.28658
comp 3	10.58058738	8.53273176	32.81932
comp 4	9.02611431	7.27912444	40.09844
comp 5	6.16340978	4.97049176	45.06893
comp 6	5.51754089	4.44962975	49.51856
comp 7	4.86157417	3.92062433	53.43919
comp 8	4.50483987	3.63293538	57.07212
comp 9	4.16265632	3.35698090	60.42910
comp 10	3.67519305	2.96386536	63.39297
...	...	...	...
comp 28	1.02178308	0.82401861	91.06926
comp 29	0.97607687	0.78715877	91.85642

Comme nous l'avons expliqué, notre base de données ne possède pas beaucoup d'individus mais énormément de variables. Cela augmente donc la dimensionnalité de notre base de données. Or, notamment dans l'analyse par regroupement, la visualisation de nos groupes est intéressant. Toute représentation graphique lisible et interprétable à l'oeil nu serait une représentation en 2D ou maximum 3D. Au delà, il devient difficile d'effectuer une analyse simple et claire.
Comme expliqué dans la revue de littérature, l'ACP est souvent utilisée pour : réduire la dimensionnalité mais aussi peut servir de méthode de réduction du nombre de variables de notre modèle. Cependant, malgré le fait que dans la littérature, le réflexe d'effectuer une ACP est devenue une norme, dans le cas de notre petit jeu de données, nous avons vu que l'ACP n'avait que très peu d'intérêt vu que la première composante n'arrive à capter que 12.9% de la variance totale. C'est pour cela que notre méthode hiérarchique ascendante sera faite sans une ACP au préalable, afin de voir si la prédiction des groupes est meilleure sans ACP.

4.4 Rappel des algorithmes de regroupement utilisés

Deux méthodes ont été utilisées afin de faire une meilleure comparaison de nos modèles.
La première est une méthode hybride qui utilise la méthode des K-Moyennes (méthode notamment proposée par Chen et al. (2005)) ainsi que la méthode hiérarchique ascendante pour les raisons avancées par Sousa and Gama (2014) dans le chapitre 2, section 2.2.3 *La méthode hiérarchique*. Cet algorithme procède en 3 étapes :

1. L'utilisation d'une ACP sur nos données d'origine afin de réduire le nombre de variables ainsi que la dimensionnalité : ainsi la réduction de bruit des variables devrait permettre une meilleure classification.

2. Effectuer une méthode ascendante hiérarchique sur les dimensions retenues de l'ACP afin de trouver le nombre k de groupes à utiliser
3. Le partitionnement est effectué à l'aide de la méthode non-hiérarchique des K-moyennes afin de consolider les groupes initialement trouvés à l'étape 2.

La deuxième méthode est une méthode hiérarchique ascendante sans ACP.

4.5 La mise en place de la méthode hiérarchique ascendante

Comme expliqué précédemment dans la revue de littérature (Chapitre 2), les étapes qui précèdent l'algorithme de regroupement sont les choix de la matrice de dissemblance ainsi que la distance entre deux groupes.
Ainsi, nous avons testé plusieurs méthodes afin de trouver la meilleure pour notre base de données.

4.6 La matrice de dissemblance

Comme vu, les deux principales matrices de dissemblances applicables sur des données de type continue sont :

— la distance Euclidienne

— la distance dite « Manhattan »

Ces dernières ont pour but de quantifier la « distance » séparant deux points. Ainsi, plus cette dernière est petite, plus les points sont semblables et peuvent donc appartenir à un même groupe.

La distance « Manhattan » calcule une distance rectiligne, c'est à dire qu'elle est contrainte à calculer le chemin le plus court où il n'est que possible de se déplacer verticalement ou horizontalement. À l'inverse, la distance Euclidienne agit dans un

plan euclidien, ce qui lui permet une plus grande liberté dans le calcul de la plus courte distance. Dans notre travail, nous utiliserons la distance Euclidienne.
Comme la distance Euclidienne est élevée au carré, nous avons dû effectuer une standardisation des variables, c'est-à-dire qu'on soustrait à chaque valeur une valeur de référence (classiquement une moyenne d'échantillon) et en la divisant par l'écart-type (typiquement un écart-type d'échantillon). Cette transformation rend toutes les valeurs (indifféremment de leurs distributions et unités de mesures originales) en unités compatibles avec distribution de moyenne 0 et d'écart-type 1. Ainsi, il est possible de comparer plus facilement des distributions de valeurs entre les variables et/ou sous-ensembles.

4.7 Distance entre 2 groupes

Dans l'algorithme de la méthode ascendante hiérarchique, des paires de groupes se forment afin de donner, itération par itération, plus qu'un seul et même groupe. Le principe reste le même que celui de la matrice de dissemblance : les groupes ayant la plus petite distance les séparant sont regroupés. De nombreuses méthodes existent comme expliqué dans la revue de littérature. Comme il s'agit d'une étape importante pour trouver le groupe le plus efficace pour notre étude, un des critères pour choisir est le coefficient d'agglomération qui permet de déterminer la stabilité d'un groupe. En effet, plus un groupe est stable, plus les sujets choisis dans ce groupe font vraiment partie de ce groupe en particulier. Ce coefficient a été calculé avec 4 méthodes différentes : Ward, Single Linkage, Complete Linkage, Weight et Average (voir tableau 4.2).

TABLE 4.2 – Comparaison des coefficients d'agglomération

Ward	**Average**	**Single**	**Complete**	**Weight**
0.6145824	0.4498748	0.3620486	0.5265511	0.5009976

Ces calculs ont été possibles grâce à la fonction **Agnes()** disponible dans R qui permet de pouvoir facilement calculer ces coefficients selon la méthode de notre choix.
La méthode Ward montre un coefficient d'agglomération stable de 0.61 , bien supérieur aux autres méthodes. Cette dernière serait donc la plus adéquate pour notre étude. De plus, comme expliqué dans la revue de littérature, il s'agit de la méthode la plus utilisée en recherche.

4.8 La détermination du nombre de groupes optimal

Comme montré dans la littérature, il n'existe pas de méthode statistique unique pour déterminer le nombre de groupes à retenir. En effet, il existe plusieurs indices possibles qui peuvent permettre de déterminer le nombre de groupes à retenir. Le meilleur moyen est de comparer les recommandations données par les indices disponibles pour trouver le nombre de groupes optimal. Le logiciel *R* permet de pouvoir facilement obtenir ce résultat de comparaison grâce à sa fonction *nbclust()* qui s'appuie sur de très nombreux indices comme :

— l'indice de Krzanowski et Lai (1988)

— l'indice CH développé par Calinski et Harabasz (1974)

— ou encore l'indice GAP développé par Tibshirani et Al (2001)

4.9 Évaluation de la qualité de prédiction

Nous terminons cette description de l'analyse par un point central du domaine statistique : la qualité du modèle utilisé.

En effet, de nombreuses méthodes de regroupement existent et certaines ont ici été appliquées. Le choix des méthodes statistiques passe avant tout par le type de base de données que nous avons : un jeu de données regroupant des données qualitatives, nominales et quantitatives ne sera pas traité de la même manière qu'un jeu possédant que des données quantitatives. Comme nous l'avons vu dans la description de nos jeux de données, ces derniers regroupent exclusivement des données quantitatives, ce qui permet déjà d'affiner le nombre de méthodes de regroupement possibles.

Afin d'être sûr de choisir la plus appropriée à notre situation, un deuxième critère de choix de modèle repose sur la qualité de prédiction. En d'autres termes, dans notre cas, une bonne prédiction représenterait une bonne classification de nos individus, c'est à dire que chaque individu est bien représentatif du groupe dans lequel il a été assigné.

Dans le cadre de notre étude, à cause de la petite quantité de données disponible (60 individus par base de données produite), il a été décidé de procéder à une classification manuelle de 14 individus, soit 23.3% de notre base de données.

La procédure de la classification manuelle se fait en plusieurs étapes :

— Afin de faire l'évaluation de la qualité de prédiction, nous nous sommes intéressés à la base de données au bout de 4 heures de jeu. Cette base de données possède 60 lignes, soit une ligne par joueur. Nous avons décidé d'utiliser celle-ci en particulier car il s'agit d'une heure avancée dans le jeu, les joueurs ont tous générés des données pour chaque attribut disponible (en d'autres termes ils

ont tous fait au moins acheté une arme ou objet dans le jeu, ont tous fait au moins une victime avec chaque type d'armes disponible, etc). De plus, comme la base de données est riche en données, il sera plus facile de faire une description détaillée et qui identifie de manière unique chaque groupe trouvé lors de la classification manuelle.

— Une fois la base de données choisie pour faire la prédiction, nous avons roulé nos deux modèles (soit la méthode hiérarchique sans ACP et la méthode hybride avec ACP) sur ce jeu de données afin de trouver le nombre de groupes à étudier dans le cadre de notre classification manuelle. Les 2 méthodes (hybride et hiérarchique ascendante) proposent 4 groupes à retenir au bout de 4 heures de jeu. Une étude approfondie de ces groupes trouvés pour chacune de ces méthodes (à travers notamment l'analyse des moyennes obtenues pour chaque groupe) nous avons remarqué que les 2 méthodes proposent 4 groupes similaires dans la définition : les joueurs Assaut, les joueurs assauts qui exploitent Far Cry 5, les joueurs type furtifs qui exploitent Far Cry 5 ainsi que les joueurs polyvalents. Ainsi, pour la suite de la procédure, notre classification manuelle se fera sur la définition de ces 4 groupes.

— Il a été décidé de classer manuellement un peu plus de 20% des 60 joueurs de la base de données. Nous nous sommes arrêtés de manière arbitraire à 14 joueurs. Le but par la suite a été de classer ces 14 individus selon les définitions des 4 groupes trouvés.

— Les individus ont été choisis afin qu'il y ait au minimum 3 et un maximum 4 individus par groupe. Le principal critère pour la sélection de ces 3 ou 4 individus par groupe se base sur la représentativité de l'individu du groupe. En d'autres termes, sur les 3 ou 4 individus, le premier ou (les deux premiers) représente parfaitement le groupe en question (par exemple, un individu qui utilise de manière très prononcée des armes de type furtif par rapport à tous les autres individus sera classé manuellement dans le groupe des joueurs de

type furtif), le deuxième est toujours évident à classer dans un groupe même si ce dernier partage certaines caractéristiques d'un autre groupe (par exemple un joueur qui utilise beaucoup plus les armes de type furtif mais qui utilise aussi un peu des armes de type assaut), enfin le troisième individu du groupe est plus ambigu à classer car il peut partager beaucoup plus de caractéristiques avec d'autres groupes aussi tout en montrant une prédominance de caractéristiques du groupe dans lequel il est classifié manuellement (par exemple, un joueur peut utiliser beaucoup les armes de type assaut tout en utilisant davantage des armes de type furtif. Malgré l'ambiguïté, ce dernier est classifié manuellement comme un joueur type furtif car il possède tout de même plus de caractéristiques de type furtif que d'assaut)

— Une fois les 3 individus par groupe trouvés ainsi que classifiés de manière manuelle en 4 groupes différents, nous regardons les groupes trouvés à l'aide des deux méthodes statistiques retenues. À l'aide de dendogrammes générés lors de l'utilisation de la méthode statistique pour classifier les individus, nous regardons comment les méthodes ont classé les 14 individus en particulier.

— Si la méthode statistique classifie l'individu de la même manière que la classification manuelle, alors cela est considéré comme une bonne classification, à l'inverse cela sera considéré comme une mauvaise classification de la part du modèle.

— Nous effectuons l'exercice vu dans le point juste avant sur les 14 individus. La méthode retenue sera celle qui classifie le plus de la même manière que la méthode de classification manuelle.

Les résultats de la classification manuelle se trouvent ci-dessous (voir tableau 4.3) :

TABLE 4.3 – Classification manuelle des 14 joueurs

Groupe	Description	ID
Le joueur furtif qui exploite FC5	Il s'agit d'un joueur qui fait en moyenne le plus de headshots et arrive à tuer le plus d'ennemis à une distance supérieure à 20 mètres ainsi qu'avec des armes privilégiant le style furtif (armes avec silencieux ou à longue portée par exemple). De plus, ce joueur est celui qui exploite le mieux le jeu FC5 en effectuant le plus d'activités du Monde Ouvert ainsi qu'en utilisant les mercenaires	65 77 78
Le joueur type Assaut qui exploite FC5	Ce joueur peut se résumer de la même manière que le joueur ci-dessus si ce n'est que ce dernier privilégie les armes de type assaut (fusils à pompe, pistolet, etc) ainsi que les armes de combat rapproché. Il exploite les ressources du jeu FC5 en y effectuant des activités du Monde Ouvert ainsi qu'en utilisant les mercenaires	92 87 36 38
Le joueur type Assaut	A la différence du joueur présenté juste au dessus, ce dernier est celui qui effectue le moins d'activités en Monde Ouvert. Étonnamment, il est à la fois le joueur qui débloque le plus de mercenaires et qui les utilise le moins	20 21 31 48
Le joueur polyvalent	Ce joueur est celui qui tue le moins d'ennemis. Il joue les deux styles de jeu : il privilégie les armes de type assaut tout en maîtrisant les headshots et les takedown, des stratégies plus furtives qu'assaut. Ses mercenaires font partie intégrante de sa stratégie de jeu en leur donnant le plus d'ordres.	72 73 39

Comme expliqué dans la procédure, la classification manuelle agit comme un outil de performance de prédiction de nos modèles. En d'autres termes, le modèle que nous choisirons sera celui qui classe les 14 individus de la même manière que la classification manuelle.

D'une part, les dendrogrammes fournis pour chaque méthode développée nous serviront à savoir si les individus sont bien répartis entre les 4 groupes proposés. Par exemple, la classification manuelle classe dans le même groupe les sujets 65, 77 et 78, donc si le dendrogramme d'une des méthodes ne les affiche pas dans le même groupe, cela nous donne un premier indice sur la qualité de la prédiction. Ci-dessous, les individus que nous avons manuellement classé ont été surlignés dans les dendrogrammes.
D'autre part, une analyse plus poussée est effectuée en analysant les données de chacun de ces 14 individus pour comprendre comment le modèle les a classé.

Donc, pour la méthode de la hiérarchie ascendante (voir figure 4.3) nous remarquons que les individus ont été classifiés selon 4 groupes :

Pour la méthode hiérarchique ascendante sans ACP :

- **Bleu** : Le joueur assaut qui exploite le moins les ressources de FC5. Débloque le plus de mercenaires sans pour autant les utiliser (joueurs 48, 21, 36, 20, 31, 39).
- **Jaune** : Le joueur polyvalent qui a plus une attirance pour le style furtif. Il exploite le Monde Ouvert ainsi que les mercenaires (joueurs 73 et 72).
- **Turquoise** : Le joueur assaut qui exploite FC5 (joueurs 38 et 92).
- **Rouge** : Joueur type furtif qui exploite FC5 (joueurs 78, 65, 77 et 87).

Pour la méthode hybride avec ACP (voir ci-dessous la figure 4.4), les individus ont été classés de même :

Pour la méthode hybride avec ACP :

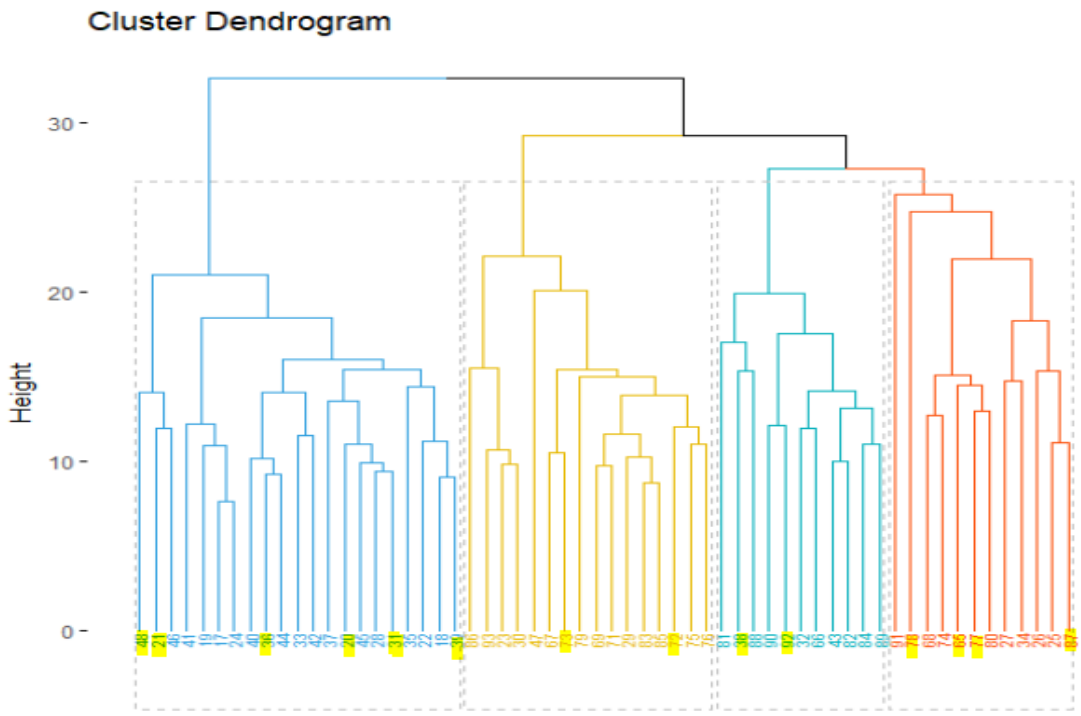


FIGURE 4.3 – Dendrogramme de la méthode hiérarchique ascendante au bout de 4 heures de jeu. Les individus surlignés correspondent aux joueurs sélectionnés pour notre classification manuelle et servent donc de point de comparaison pour la performance du modèle.

— **Rouge** : Joueur type furtif qui exploite FC5 (joueurs 36, 73, 72).

— **Noir** : Joueur Assaut. Il est le joueur qui débloque le plus de mercenaires tout en étant celui qui les utilise le moins (joueurs 20, 31 et 39).

— **Bleu** : Le joueur polyvalent qui a plus une attirance pour le style furtif. Il exploite le Monde Ouvert ainsi que les mercenaires (joueurs 78, 65 et 77).

— **Vert** : Le joueur assaut qui exploite FC5 (92, 38, 21, 48 et 87).

Un code couleur spécifique a été utilisé afin de mieux cerner les classifications :

La dernière étape de la procédure consiste à faire la comparaison entre les résultats obtenus pour ces 14 joueurs pour les méthodes hybride et de hiérarchie descendante avec la classification manuelle. Un code couleur a été utilisé afin de mettre en évidence les bonnes et mauvaises prédictions dans le tableau 4.6 :

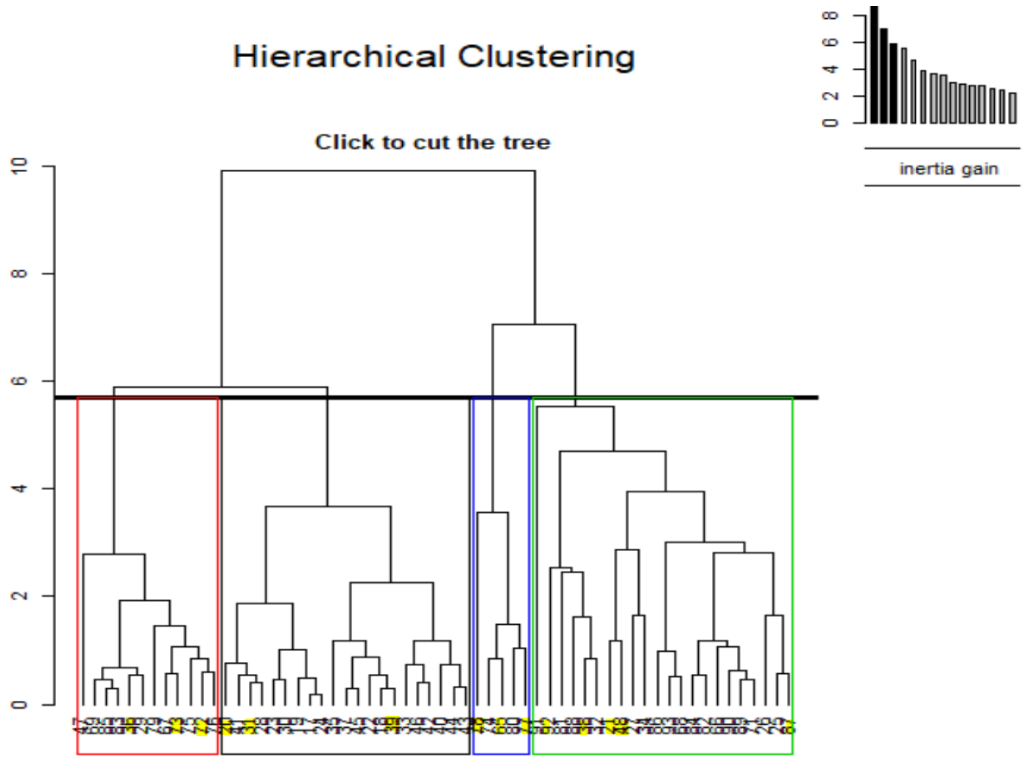


FIGURE 4.4 – Dendrogramme de la méthode hybride avec ACP au bout de 4 heures de jeu. Les individus surlignés correspondent aux joueurs sélectionnés pour notre classification manuelle et servent donc de point de comparaison pour la performance du modèle.

FIGURE 4.5 – Classification des individus pour les 2 méthodes utilisées

— **Rouge** : mauvaise classification de l'individu. Par exemple si la bonne classification est "Joueur Assaut" et que notre méthode le classifie comme étant un "Joueur type furtif" alors la classification est mauvaise car il s'agit de deux styles de jeu très différents.

— **Vert** : Bonne classification de l'individu.

Nous remarquons que la méthode hiérarchique ascendante sans ACP classifie mieux que la méthode hybride. En effet, 6 individus sur 14 ont été mal classifiés avec la méthode hybride contrairement à 4 pour la méthode hiérarchique. Afin de développer nos résultats, nous avons décidé d'utiliser la méthode classifiant le mieux, c'est à dire la méthode hiérarchique ascendante.

Joueur ID	Méthode de Classification		
	Classification Manuelle	Méthode hybride	Méthode hiérarchique ascendante
20	Joueur Assaut	Joueur Type Furtif qui exploite FC5	Joueur Assaut
21	Joueur Assaut	Joueur Assaut qui exploite FC5	Joueur Assaut
31	Joueur Assaut	Joueur Type Furtif qui exploite FC5	Joueur Assaut
48	Joueur Assaut	Joueur Assaut qui exploite FC5	Joueur Assaut
72	Joueur Polyvalent	Joueur Polyvalent	Joueur Polyvalent
73	Joueur Polyvalent	Joueur Polyvalent	Joueur Polyvalent
39	Joueur Polyvalent	Joueur Assaut	Joueur Assaut
65	Joueur Type Furtif qui exploite FC5	Joueur Type Furtif qui exploite FC5	Joueur Type Furtif qui exploite FC5
77	Joueur Type Furtif qui exploite FC5	Joueur Type Furtif qui exploite FC5	Joueur Type Furtif qui exploite FC5
78	Joueur Type Furtif qui exploite FC5	Joueur Type Furtif qui exploite FC5	Joueur Type Furtif qui exploite FC5
92	Joueur Assaut qui exploite FC5	Joueur Assaut qui exploite FC5	Joueur Assaut
87	Joueur Assaut qui exploite FC5	Joueur Assaut qui exploite FC5	Joueur Type Furtif qui exploite FC5
36	Joueur Assaut qui exploite FC5	Joueur Polyvalent	Joueur Assaut
38	Joueur Assaut qui exploite FC5	Joueur Assaut qui exploite FC5	Joueur Assaut qui exploite FC5

FIGURE 4.6 – Comparaison de la performance de classification des modèles choisis

Chapitre 5

Présentation des résultats

Suite au chapitre précédent, nous avons montré que la méthode hiérarchique ascendante était plus performante dans notre cas car elle arrivait à mieux classer les individus.

5.1 Méthodologie utilisée

Comme nous l'avons expliqué dans la section 3.3 *Choix de la découpe des bases de données*, il a été décidé de traiter l'aspect temporel en séparant la base de données en 6. Ainsi, chaque nouvelle base de données produite s'intéresse à une période de jeu en particulier :

— Les 40 premières minutes de jeu,

— entre 41 et 60 minutes de jeu,

— entre 61 minutes et 2 heures de jeu,

— entre 2h01 et 3h de jeu,

— entre 3h01 et 4h de jeu,

— entre 4h01 et 5h de jeu.

En d'autres termes, la méthode de clustering a été refaite à chaque période de jeu indépendemment.

5.2 Résultats

Comme expliqué dans la revue de littérature, au delà de toutes les méthodes de choix de groupes possibles, il est plus important de trouver des groupes qui font du sens à l'analyste selon son sujet de recherche. En effet, le plus difficile est de définir les groupes et donc de les expliquer.
Avant toute démarche statistique, il a donc été important de réfléchir sur les éventuels groupes pouvant exister dans la base de données. Pour cela, une très bonne connaissance du jeu de données est primordiale et donc une bonne analyse explicative est importante afin de tirer le meilleur de la base de données. Dans le cadre de cette étude, un moyen supplémentaire s'est offert : les tests utilisateurs.
Ceux-ci ont pour but d'expliquer le comportement du joueur, ainsi il s'agit d'un excellent moyen de compréhension de données. En effet, vu que le comportement du joueur est retranscrit de manière quantitative à travers la télémétrie, l'observation du joueur permet de mieux comprendre les tendances que peuvent faire transparaître la télémétrie.

Grâce à notre classification manuelle, nous avons pu en conclure que la méthode hiérarchique ascendante prédisait mieux que la méthode hybride dans notre cas. Ainsi, en prenant en considération ce que nous venons tout juste de dire, tous les résultats présentés ci-dessous sont ceux issus de la méthode hiérarchique ascendante.

Procédure pour l'obtention des résultats

Une procédure unique a été utilisée afin de déterminer le nombre de groupes à retenir ainsi que l'analyse des groupes à chaque période de temps définie.

- Pour le choix du nombre de groupes, 2 étapes sont effectuées :
  - Dans un premier temps, nous nous basons premièrement sur un package R **NbClust** qui compare les résultats obtenus par plus de 30 indices différents concernant le nombre de groupes à retenir (parmi les indices

analysés, figure notamment les quantités SPRSQ et RSQ ou encore la règle du coude). La sortie générée montre un histogramme avec en abscisse le nombre de groupes à retenir et en ordonnée la somme des indices qui ont proposé ce nombre à retenir.

— Dans un second temps, comme nous l'avons vu dans le chapitre 2, *La Revue de Littérature*, la détermination du nombre de groupes reste la partie la plus difficile à faire en segmentation. Ainsi, la détermination de ce nombre dépend aussi de l'analyste qui peut jouer avec le nombre de groupes à retenir afin de choisir celui qui convient le mieux à l'analyse et qui permet de produire des groupes interprétables.

Ainsi, le choix du nombre à retenir est d'abord guidé par l'histogramme qui indique le nombre de groupes à retenir. Ensuite, nous « jouons » avec le nombre recommandé, généralement en testant notre modèle avec un groupe en moins et en plus que celui qui est recommandé afin de voir à quel point cela influence l'interprétation des groupes. Pour cela nous utilisons le dendrogramme formé pour voir la répartition des individus entre les groupes. Nous recherchons des tendances de jeu générales, ainsi nous portons une attention toute particulière au nombre de groupes qui distribue de manière homogènes les individus entre les groupes (c'est à dire qu'il y ait plus ou moins le même nombre d'individus par groupe).

Des explications plus précises seront données au fur et à mesure de l'avancement de l'analyse.

— Une fois le nombre de groupes retenu, nous passons à l'interprétation des groupes formés par notre modèle. Cette étape se base sur les données des joueurs de chaque groupe, où des moyennes sont formées afin de faciliter l'interprétation.

Ainsi, la présentation des résultats ci-dessous se base intégralement sur cette procédure.

Au bout de 40 minutes de jeu

Comme nous l'avons expliqué dans le chapitre 3, dans la section *Catégorisation des actions prises par le joueur*, nous avons délibérément décidé de ne pas étudier les 20 premières minutes de jeu.
Au bout de 40 minutes de jeu, le joueur se trouve normalement encore dans le tutoriel dynamique. Ainsi, il est guidé dans ses actions afin de pouvoir comprendre les mécanismes de jeu. Cependant au bout de 40 minutes, il a eu l'occasion de tester les mécanismes de combat proposés par le jeu.

Comme nous le montrent les graphiques ci-dessous, deux groupes sont conseillés. Le dendrogramme montre ainsi que seuls 4 joueurs ont une attitude bien différente de l'autre groupe qui représente la majorité des joueurs.
Deux groupes explicables peuvent être identifiés :

— **Le joueur assaut qui exploite FC5** : En d'autres termes ce groupe se distingue d'une part par sa manière de jouer qui est orientée sur les armes d'assaut et d'autre part par son utilisation active des ressources caractéristiques du jeu FC5.

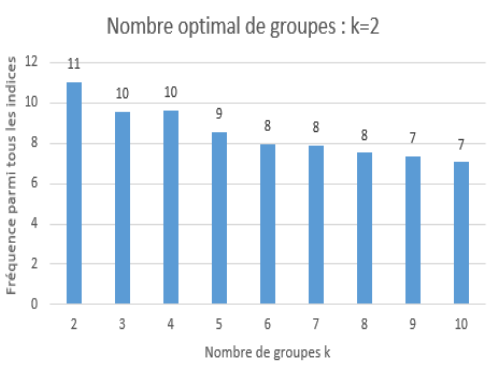
Ce joueur fait le plus de victimes à une distance inférieure à 20 mètres ainsi qu'avec des armes d'assaut. En plus de cela il s'agit du groupe qui aime le combat rapproché pour s'attaquer à ses ennemis. Par ce fait, nous pouvons en conclure qu'il s'agit d'un joueur de type assaut.

Tous les joueurs de ce groupe ont débloqué le mercenaire disponible lors du tutoriel tout en ayant profité du monde ouvert offert par Far Cry 5 (notamment en découvrant de nombreux lieux).

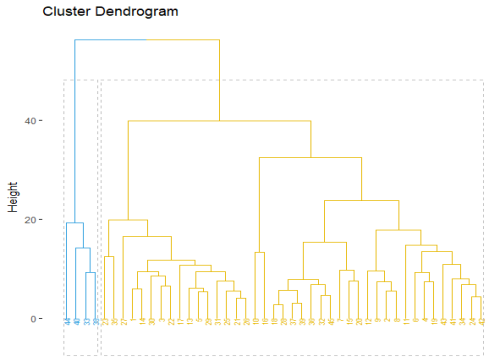
— **Le joueur type furtif** : Ce groupe représente la majorité des individus. Le joueur caractéristique de ce groupe effectue le plus de headshots ainsi qu'exploite le mieux le sniper pour tuer ses ennemis avec 1 mort pour 16 ennemis tués. Enfin, il possède le meilleur ratio d'ennemis tués avec une arme qui privi-

légie le type furtif. Aucun joueur n'a débloqué le mercenaire disponible lors du tutoriel dynamique. Enfin, ce dernier n'utilise aucune ressource caractéristique de Far Cry 5.

Il est important de noter qu'il est tout à fait normal d'avoir une grande majorité de joueurs regroupés dans un seul et même groupe à ce stade du jeu. En effet, comme expliqué précédemment le tutoriel a pour but de guider le joueur dans une série d'actions afin de pouvoir comprendre les mécanismes du jeu, que cela soit au niveau de la manette que des ressources disponibles dans le jeu. Ainsi, vu que le tutoriel accompagne le joueur, il est normal que la plupart jouent de la même manière pendant cette période de temps.



(a) 40 minutes k = 2 groupes



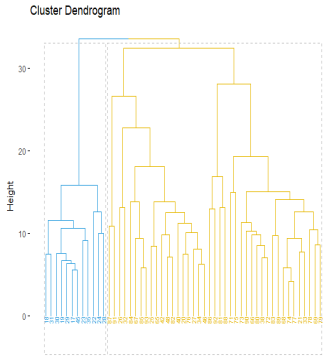
(b) 40 minutes Dendrogramme pour k= 2

FIGURE 5.1 – 40 minutes : Nombre de groupes et dendrogramme

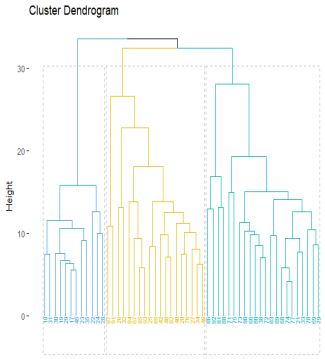
À 60 minutes de jeu

Le joueur se trouve vers la fin du tutoriel voire l'a terminé et commence donc à goûter réellement aux ressources qu'offre le Monde Ouvert de Far Cry 5. Il se retrouve donc libre de parcourir le monde, d'aller de région en région, de commencer plusieurs quêtes à la fois ou juste profiter du Monde ouvert.

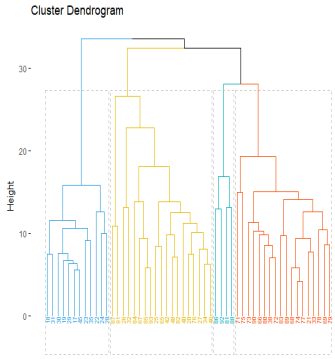
La sortie **NbClust** nous recommande 2 groupes. Cependant, comme expliqué dans la procédure, le point de vue de l'analyste compte afin de déterminer des groupes interprétables et qui ont du sens avec l'étude. Ainsi, nous avons tester notre méthode hiérarchique ascendante avec différents nombres de groupes : 2, 3 et 4. Les dendrogrammes sont présentés dans la figure 5.2 :



(a) 60 minutes Dendrogramme pour k = 2 groupes.



(b) 60 minutes Dendrogramme pour k= 3 groupes.



(c) 60 minutes Dendrogramme pour k= 4 groupes.

FIGURE 5.2 – 60 minutes : Les Dendrogrammes formés pour 2, 3 et 4 groupes

Comme nous l'avons déjà expliqué dans la procédure, nous cherchons une répartition homogène des individus entre les groupes afin de cibler les principaux groupes et travailler sur l'amélioration de ces stratégies de jeu-ci. En effet, vu la multitude de possibilités qu'a le joueur dans le jeu et les courtes échéances en entreprise, il est important de focaliser les efforts d'amélioration d'utilisation des stratégies de jeu sur les styles de jeu principaux afin de répondre à la demande de la majorité des joueurs.

Ainsi, en partant avec cette logique, nous remarquons dans notre figure 5.2, que 3 groupes serait un bon choix : les individus sont bien divisés entre les groupes contrairement à 4 groupes qui propose un groupe seulement formé de 4 individus, ce qui

ne cible pas notre objectif de groupes principaux.

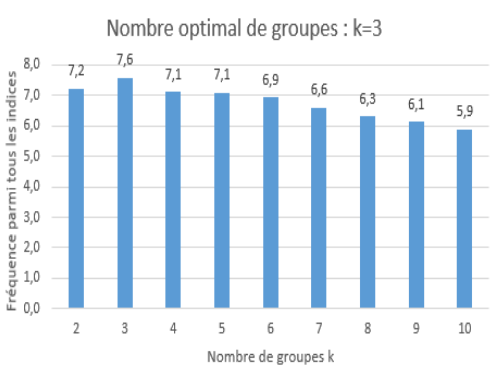
Par conséquent, trois groupes sont retenus et le dendrogramme montre ainsi une répartition équilibrée des joueurs entre ces trois groupes.

— **Joueur polyvalent** : Ce joueur maîtrise aussi bien le style furtif que l'assaut. Cependant, ce dernier ne profite pas des ressources disponibles dans le jeu : il effectue peu d'activités du monde ouvert ainsi que n'utilise pas les mercenaires mis à disposition.

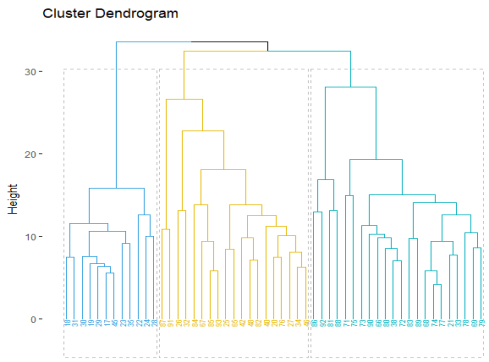
— **Le Joueur type furtif exploitant Far Cry 5** : Ce joueur utilise des armes de type furtif pour tuer ses ennemis. Ce dernier utilise le plus l'arc ainsi que les mercenaires.Il excelle dans le style FPS en affichant le meilleur ratio d'ennemis tués par mort de joueur. Aussi, il exploite le mieux les ressources de Far Cry 5 en faisant le plus d'activités du monde ouvert et en utilisant beaucoup les mercenaires dans sa stratégie de jeu.

— **Le joueur Assaut exploitant Far Cry 5** : Ce joueur exploite les ressources de Far Cry en utilisant les mercenaires, ainsi qu'en effectuant les activités disponibles à part les outposts. Comparé aux autres groupes, ce dernier meurt le plus en campagne, mais enregistre aussi le plus d'ennemis tués. Il tue essentiellement à l'aide d'armes de type assaut ainsi qu'à une distance inférieure à 20 mètres.

Les groupes trouvés entre 40 minutes et maintenant 1 heure de jeu montrent une évolution intéressante : en effet, le joueur de style furtif est le groupe qui exploite maintenant le plus les ressources disponibles du jeu, aussi bien concernant le monde ouvert que les mercenaires.



(a) 60 minutes k = 3 groupes



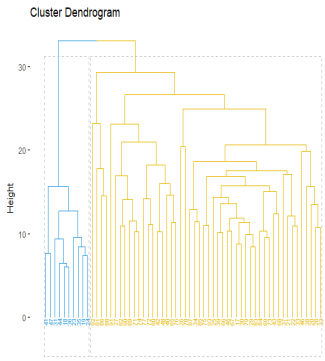
(b) 60 minutes Dendrogramme pour k= 3

FIGURE 5.3 – 60 minutes : Nombre de groupes et dendrogramme

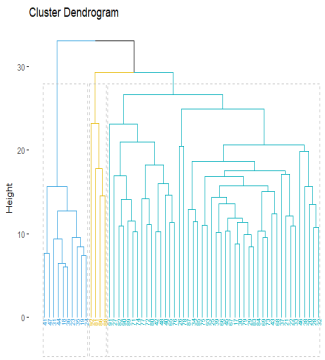
Au bout de 2 heures de jeu

Le joueur profite pleinement du monde ouvert offert et est libre d'aborder le jeu comme il lui souhaite : réaliser l'histoire proposée, interagir avec l'environnement, faire les missions annexes, etc.

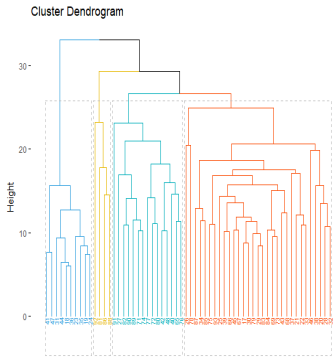
Nous remarquons sur le graphique 5.5 que 2 groupes sont recommandés. Nous avons donc tenté de voir ce que notre modèle donne avec : 2, 3 et 4 groupes. Les dendrogrammes sont présentés dans le graphique 5.4.



(a) 2 heures Dendrogramme pour k = 2 groupes.



(b) 2 heures Dendrogramme pour k= 3 groupes.



(c) 2 heures Dendrogramme pour k= 4 groupes.

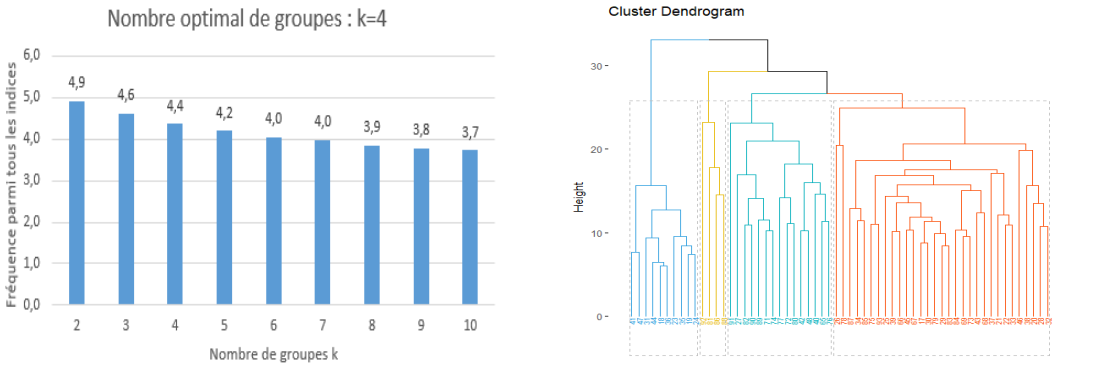
FIGURE 5.4 – 2 heures : Les Dendrogrammes formés pour 2, 3 et 4 groupes

Nous remarquons que pour 2 groupes, nous avons presque la même répartition des joueurs qu'après 40 minutes de jeu. Pour 3 groupes, le groupe supplémentaire

semble être très spécifique à 4 joueurs. Pour 4 groupes, nous remarquons que notre plus gros groupe se divise en 2 principaux groupes homogènes. Cette division intéressante nous a décidé pour baser l'interprétation sur 4 groupes :

- **Le joueur type furtif exploitant Far Cry** : Ce groupe est le même que celui présent lors de la première heure de jeu : ce joueur privilégie l'approche furtive en effectuant le plus de headshots ainsi qu'en affichant le meilleur ratio d'ennemis tués avec une arme de type furtif. Enfin, ce dernier exploite bien les mercenaires ainsi que les activités disponibles dans le monde ouvert.
- **Joueur polyvalent** : Ce dernier maîtrise aussi bien les armes furtives que d'assaut avec une préférence pour le type furtif. En effet, il possède le 2^{eme} meilleur ratio d'ennemis tués avec des armes furtives ainsi que le meilleur ratio d'ennemis tués avec des armes de type Sniper.
- **Le joueur assaut** : Ce joueur assaut est celui qui utilise le moins les mercenaires comparé aux autres groupes. De plus, ce dernier exploite très peu le monde ouvert mis à sa disposition.
- **Le joueur assaut qui exploite Far Cry 5** : Ce joueur aussi de type assaut a comme particularité d'utiliser les mercenaires pour mettre à bien sa stratégie d'assaut. Il s'agit de la majeure différence avec le groupe juste au dessus.

Entre 1 et 2 heures de jeu, nous remarquons une évolution dans les groupes. Les 3 groupes identifiés au bout d'une heure de jeu se sont divisés en 4 groupes distincts au bout de 2 heures de jeu. Sur les 4 groupes, 2 se démarquent comme des joueurs qui emploient une stratégie furtive et 2 autres qui privilégient l'assaut. Les 3 groupes trouvés au bout d'une heure de jeu sont retrouvés au bout de 2 heures de jeu. Le nouveau groupe supplémentaire qui est apparu au bout de 2 heures de jeu est un groupe assaut qui n'exploite pas du tout les ressources (activités du monde ouvert et mercenaires) de Far Cry 5 comparé aux autres groupes présents.

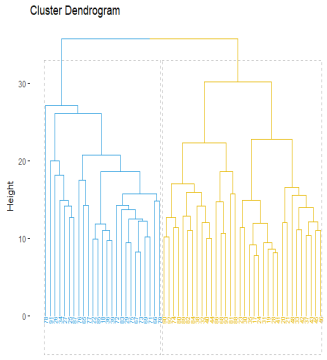


(a) 2 heures k = 4 groupes (b) 2 heures Dendrogramme pour k= 4

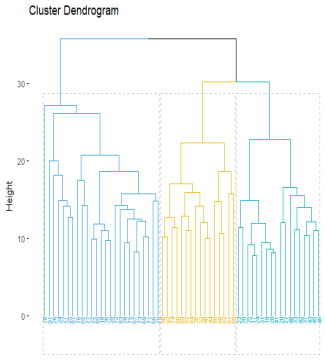
FIGURE 5.5 – 2 Heures : Nombre de groupes et dendrogramme

Au bout de 3 heures de jeu

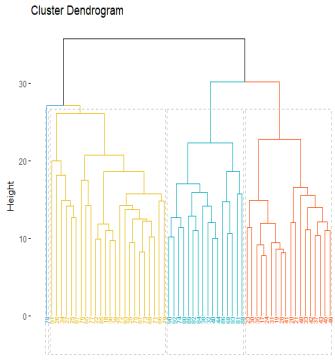
Nous remarquons que deux principaux groupes sont recommandés dans le graphique 5.6. Nous avons donc étudié notre modèle selon 2, 3 ou 4 groupes pour voir le nombre qui correspondrait le mieux à notre étude.



(a) 3 heures Dendrogramme pour k = 2 groupes.



(b) 3 heures Dendrogramme pour k= 3 groupes.



(c) 3 heures Dendrogramme pour k= 4 groupes.

FIGURE 5.6 – 3 heures : Les Dendrogrammes formés pour 2, 3 et 4 groupes

Pour la première fois depuis le début de notre étude, la division des individus en 2 groupes ne ressemble pas à celle présente au bout de 40 minutes de jeu. En effet, les individus sont mieux répartis entre ces deux groupes. La meilleur répartition des joueurs se trouve lorsque notre modèle divise les individus en 3 groupes distincts : les groupes formés montre une répartition homogène des individus. Pour 4 groupes, un individu forme à lui tout seul un groupe, ce qui n'est pas ce que nous cherchons

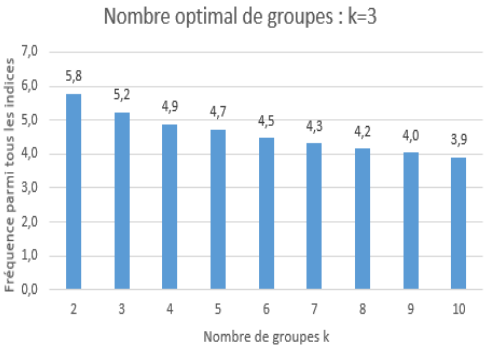
à étudier dans notre cas. Ainsi, 3 groupes ont été retenus pour la suite de l'analyse.

Tout comme lors de la première heure nous avons donc des groupes avec une répartition équitable des joueurs.

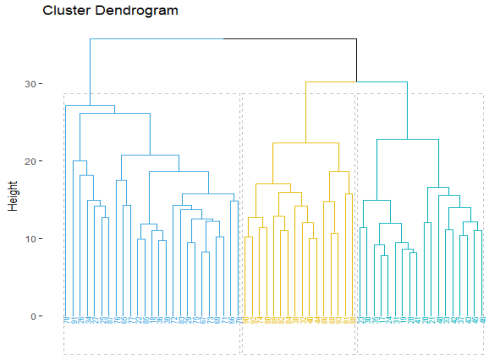
— **Le joueur Type Furtif qui exploite Far Cry 5** : Ce dernier a le meilleur ratio de morts par ennemi tué, prouvant une excellente maîtrise du jeu FPS. En plus de cela, ce dernier exploite le mieux les ressources disponibles de Far Cry 5 en effectuant le plus d'activités du monde ouvert ainsi qu'en utilisant le plus les mercenaires.

— **Le joueur assaut qui exploite Far Cry 5** : Ce joueur tue le plus d'ennemis surtout avec des armes d'assaut. Par la même occasion, ses mercenaires tuent aussi le plus d'ennemis, ce qui prouve que leur utilisation sert à la stratégie du joueur. Enfin, ce joueur apprécie le contenu de Far Cry 5 car il effectue de très nombreuses activités du monde ouvert.

— **Le joueur assaut** : Ce joueur de type assaut apprécie le combat rapproché mais n'exploite pas les ressources disponibles de Far Cry 5. En effet, il a comme particularité de débloquer un nombre très important de mercenaires, surtout de type résistant, mais ne les utilise absolument pas.



(a) 3 heures k = 3 groupes



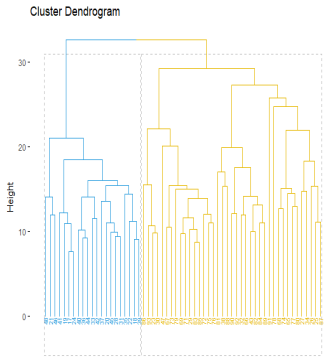
(b) 3 heures Dendrogramme pour k= 3

FIGURE 5.7 – 3 Heures : Nombre de groupes et dendrogramme

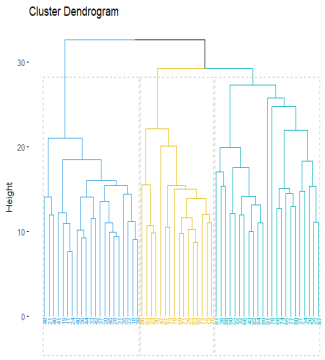
Nous remarquons que nous sommes passés de 4 à 3 groupes entre 2 et 3 heures de jeu. En effet, nous retrouvons toujours les joueurs type furtif et assaut qui exploitent les ressources de Far Cry 5 ainsi que le joueur assaut qui ne profite pas des particularités de Far Cry 5. Cependant, le joueur polyvalent s'est vu absorbé par ces trois groupes-ci. Afin de mieux comprendre l'évolution de ces joueurs, une représentation graphique plus détaillée sera donnée dans la section 5.2, *Représentation Graphiques*.

Au bout de 4 heures de jeu

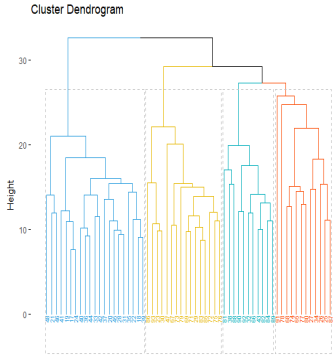
Encore une fois, d'après la figure 5.9, 2 groupes sont recommandés. Comme pour les analyses précédentes, une étude a été faite pour 2, 3 et 4 groupes pour voir si un autre nombre de groupes ne correspondrait pas mieux à notre analyse. Les résultats sont présentés ci-dessous dans la figure 5.8.



(a) 4 heures Dendrogramme pour k = 2 groupes.



(b) 4 heures Dendrogramme pour k= 3 groupes.



(c) 4 heures Dendrogramme pour k= 4 groupes.

FIGURE 5.8 – 4 heures : Les Dendrogrammes formés pour 2, 3 et 4 groupes

Nous remarquons encore une fois que 2 groupes sont recommandés d'après les indices de détermination du nombre de groupes présentés dans le graphique 5.9. Cependant, il a été décidé de jouer avec ce paramètre et regarder la répartition des joueurs pour 2, 3 et 4 groupes.

Si nous choisissons 3 groupes, nous nous retrouvons avec des groupes semblables à ceux trouvés au bout de trois heures de jeu. A l'inverse, avec 4 groupes nous nous retrouvons avec l'équivalent des groupes trouvés au bout de 2 heures de jeu. Cependant, nous remarquons qu'au bout de 4 heures, les 4 groupes formés montrent un nombre équivalent d'individus par groupe. Ainsi, nous avons décidé d'étudier ces 4 groupes distincts au lieu de 3.

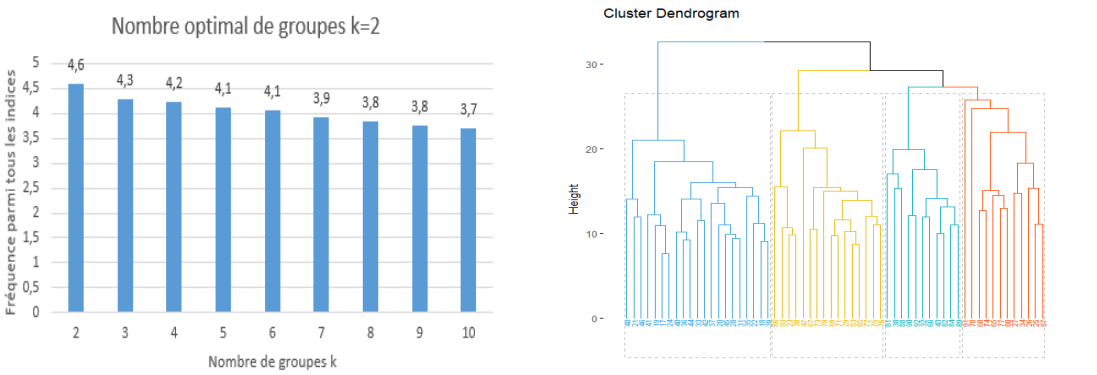
— **Le joueur assaut qui exploite Far Cry 5** : Ce joueur déjà présent au bout de trois heures de jeu est caractérisé par une bonne maîtrise des armes d'assaut ainsi qu'aime le combat rapproché. Il exploite les ressources disponibles dans le jeu en effectuant des activités du Monde Ouvert ainsi qu'en utilisant les mercenaires

— **Le joueur type furtif qui exploite Far Cry 5** : Il s'agit du même groupe que depuis le début : ce dernier fait le plus de headshots et tue le plus d'ennemis à une distance supérieure à 20 mètres. Il privilégie les armes de type furtif. Enfin, il fait partie du groupe qui exploite le mieux le Monde Ouvert de Far Cry 5 avec la meilleure utilisation des mercenaires.

— **Le joueur Assaut** : Lui aussi se retrouve dans les heures précédentes. Il privilégie les armes d'assaut tout en n'exploitant pas du tout les ressources du jeu Far Cry 5.

— **Le joueur polyvalent** : Ce joueur qui avait disparu lors de la troisième heure de jeu fait sa réapparition ici en plus grande proportion que ce que nous avons vu au bout de 2 heures de jeu. Ce groupe a comme nouvelle particularité d'être plus assaut que furtif comparé à ce que nous avions plus haut.

Au bout de 4 heures de jeu, nous nous retrouvons avec des groupes semblables à ceux présents au bout de 2 heures. Comparé aux groupes trouvés au bout de 3 heures, dans notre cas un groupe est réapparu : le groupe des joueurs polyvalents. Ce dernier a cependant une particularité comparé à ce que nous avions au bout de 2 heures de jeu : il a une plus grande tendance pour l'assaut que le style furtif.



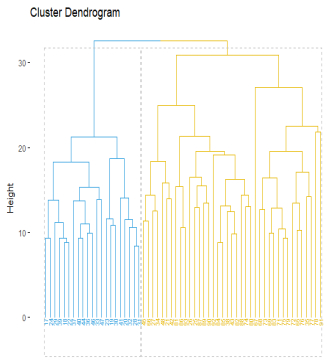
(a) 4 heures k = 4 groupes (b) 4 heures Dendrogramme pour k= 4

FIGURE 5.9 – 4 Heures : Nombre de groupes et dendrogramme

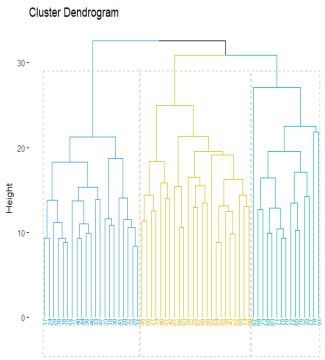
Au bout de 5 heures de jeu

Nous arrivons à la dernière heure de jeu qui peut représenter la première véritable expérience de jeu au joueur. En effet, en 5 heures de jeu, le joueur connaît les mécanismes de jeu disponibles, est rentré (ou non) dans l'histoire et peut donc se construire sa propre opinion du jeu et décider (ou non) de continuer à jouer.

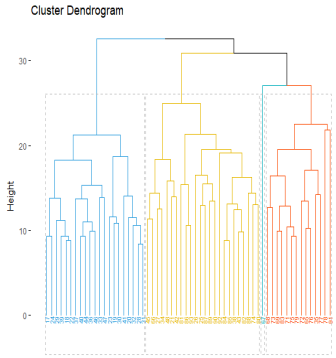
Cette fois-ci, 3 groupes sont recommandés d'après le graphique 5.11. Afin d'être certain qu'il s'agit bien du bon nombre à retenir, nous testons notre modèle avec 2, 3 et 4 groupes. Les résultats sont présentés dans l'ensemble des graphiques 5.10.



(a) 5 heures Dendrogramme pour k = 2 groupes.



(b) 5 heures Dendrogramme pour k= 3 groupes.



(c) 5 heures Dendrogramme pour k= 4 groupes.

FIGURE 5.10 – 5 heures : Les Dendrogrammes formés pour 2, 3 et 4 groupes

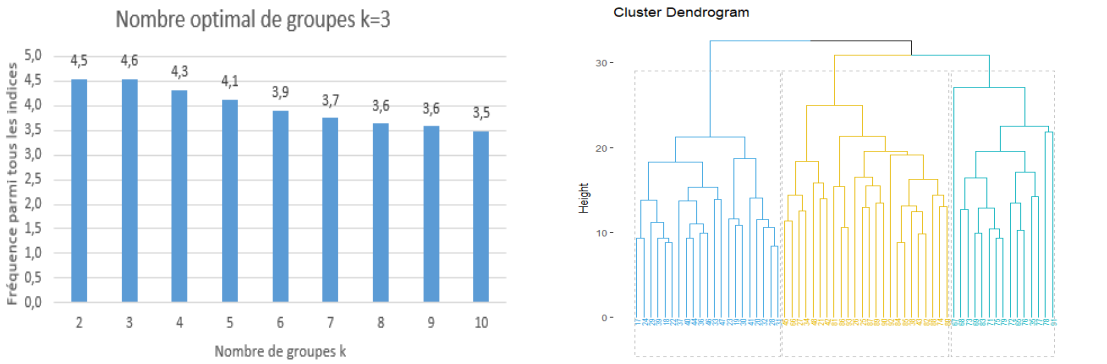
Nous remarquons que les individus sont répartis de manière homogène entre les

3 groupes proposés. Pour 4 groupes, nous avons une répartition semblable à celle trouvée au bout de 3 heures de jeu, c'est-à-dire qu'un individu forme à lui tout seul un groupe. De ce fait, nous avons décidé d'étudier les 3 groupes recommandés :

— **Le joueur type Furtif qui exploite Far Cry 5** : Ce joueur présent depuis la deuxième heure de jeu reste relativement stable en affichant toujours le plus de headshots ainsi que le plus d'ennemis tués à l'aide d'une arme furtive. De plus, ce dernier possède un excellent ratio de 34 ennemis tués pour 1 mort enregistrée. Enfin, ce dernier exploite le plus les ressources de Far Cry 5 en effectuant de nombreuses activités du Monde Ouvert ainsi qu'en utilisant le mieux ses mercenaires dans sa stratégie de jeu.

— **Le joueur Assaut qui exploite Far Cry 5** : Ce joueur présent depuis le début du jeu tue le plus d'ennemis avec des armes de type assaut ainsi qu'en combat rapproché. Il exploite très bien le Monde Ouvert en y effectuant la majorité des activités proposées. Enfin, il est utilise aussi ses mercenaires.

— **Le joueur Assaut** : Ce dernier qui lui aussi est apparut au bout de 2 heures de jeu se décrit par l'utilisation active d'armes de type assaut ainsi que du combat rapproché. Ce groupe débloque le plus de mercenaires mais ne les utilise pas du tout. Il n'exploite pas non plus les ressources du Monde Ouvert que propose Far Cry 5.



(a) 5 heures k = 3 groupes (b) 5 heures Dendrogramme pour k= 3

FIGURE 5.11 – 5 Heures : Nombre de groupes et dendrogramme

La grande différence entre la 4ème heure et 5ème heure de jeu est que nous sommes maintenant passés de 4 à 3 groupes. En effet, le groupe caractéristique du joueur polyvalent a disparu.

5.3 Représentations Graphiques

Cette section a pour but d'appuyer les résultats trouvés précédemment ainsi que de les enrichir en exploitant d'autres aspects de visualisation disponibles. De plus, le but principal de cette étude est de pouvoir présenter des informations facilement interprétables à des équipes business. Ainsi, des représentations claires, précises et utiles sont à privilégier.

Dans le cadre de notre étude nous avons décidé d'exploiter deux types de représentations : une représentation des positions géographiques des joueurs sur la carte de Far Cry 5 ainsi qu'une visualisation représentant le flux de joueurs passant d'un groupe à un autre au cours du temps. Ainsi, ces deux représentations nous permettront réellement de voir l'évolution du style de jeu des joueurs au cours du temps ainsi que les parties dans le jeu plus propices à un certain style de jeu qu'à un autre.

Représentation du flux des joueurs d'un groupe à un autre

Cette représentation graphique a comme but de mettre en évidence le flux de joueurs qui passent d'un groupe à un autre au cours du temps. Cela nous permettra d'expliquer plus facilement l'évolution des groupes entre les heures de jeu que nous avons abordé précédemment. Afin d'y arriver nous avons décidé de faire un diagramme de Sankey.

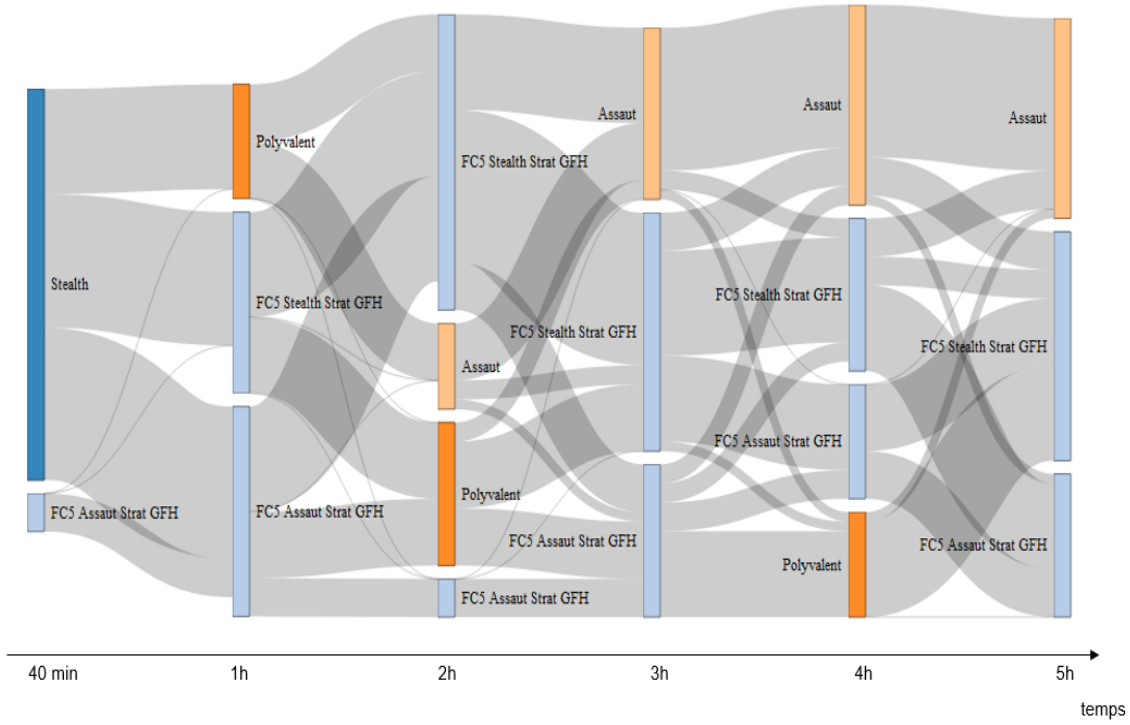


FIGURE 5.12 – Diagramme de Sankey représentant le flux des joueurs en fonction du style de jeu

Nous avons brièvement mentionné l'évolution que nous pouvions voir des groupes formés à travers le temps de jeu dans la section 5.1, *Résultats*. Ce diagramme permet d'expliquer beaucoup plus facilement comment les individus sont passés de groupe en groupe au fur et à mesure de l'avancée du jeu.

A des fins de simplification et pour ne pas encombrer le graphique, nous avons renommé les groupes trouvés dans la partie 5.1 comme-suit :

— *Stealth* équivaut au groupe *Le Joueur type furtif* ;

— *FC5 Assaut Strat GFH* équivaut au groupe *Le Joueur Assaut qui exploite Far Cry 5* ;

— *Polyvalent* ne change pas de celui que nous avions en section 5.1 ;

— *FC5 Stealth Strat GFH* équivaut au groupe *Le Joueur type Furtif qui exploite Far Cry 5* ;

— Enfin, *Assaut* équivaut au groupe *Le Joueur Assaut*.

Interprétation du diagramme de Sankey

Entre 40 et 60 minutes de jeu

Dans un premier temps, la représentation nous montre qu'au bout d'une heure de jeu les joueurs de type furtif trouvés au bout de 40 minutes de jeu se sont divisés de manière équitable pour devenir soit des joueurs au style polyvalents, soit de type furtif qui exploitent les ressources FC5 ou encore qui sont devenus des joueurs assaut qui exploitent Far Cry 5.
Comme nous pouvions l'imaginer, la majorité des joueurs de type assaut qui exploitent Far Cry 5 restent dans ce même groupe après une heure de jeu.

Entre 61 minutes et 2 heures de jeu

Comme nous l'avons vu dans la section 5.1, on passe de 3 à 4 groupes entre 1 heure et 2 heures de jeu : le groupe représentant les joueurs de type assaut apparaît. Alors que nous pourrions imaginer que le groupe polyvalent conserve les mêmes joueurs entre 1 heure et 2 heures de jeu, nous remarquons une tendance étonnante : les joueurs polyvalents au bout d'une heure de jeu se séparent de manière équitable en 2 pour soit aller rejoindre le groupe des joueurs de type furtif qui exploitent Far Cry 5 soit pour le groupe des joueurs assaut. Même si le groupe polyvalent au bout

d'une heure de jeu n'est plus du tout composé des mêmes joueurs qu'au bout de 2 heures de jeu, la séparation est logique vu que ces joueurs maîtrisent aussi bien le type furtif que l'assaut.

Une autre étonnante tendance se fait voir pour les joueurs de type assaut qui exploitent Far Cry 5 au bout d'une heure de jeu : Une grosse partie d'entre eux se retrouvent dans le groupe des joueurs de type furtif qui exploitent Far Cry 5. Cette tendance peut peut-être s'expliquer par la suite avec la deuxième représentation des résultats que nous présentons : la représentation géographique des styles de jeu. En effet, il est possible que certaines parties de la carte du jeu soient plus propices au style furtif que de l'assaut.

Si ce n'est ces tendances intéressantes pour certains des joueurs, les autres flux semblent logiques. Par exemple, les joueurs de style furtif qui exploitent Far Cry 5 et qui ne sont pas devenus polyvalents au bout de 2 heures de jeu restent dans le même groupe. Même logique pour les joueurs assaut qui exploitent Far Cry 5.

Le groupe des joueurs assaut qui exploitent Far Cry 5 rétrécit de manière considérable à cause d'un flux vers le style polyvalent et de type furtif qui exploite les ressources du jeu.

Entre 2h01 et 3 heures de jeu

Même si une grande majorité des joueurs de type furtif qui exploitent Far Cry 5 restent dans ce même groupe au bout de 3 heures de jeu, nous remarquons qu'une partie non négligeable devient des joueurs assaut ainsi que des joueurs assaut qui exploitent les ressources de Far Cry 5. 2 possibilités se font voir pour expliquer ce phénomène : soit les joueurs se trouvent dans une zone où il faut privilégier l'assaut, soit il s'agit des joueurs qui étaient de type assaut exploitant Far Cry 5 après 1 heure de jeu qui sont revenus dans cette catégorie.

Quant à eux, les joueurs polyvalents trouvés au bout de 2 heures de jeu se divisent de manière équitable entre les joueurs de style furtif qui exploitent Far Cry 5 et les joueurs de type assaut qui exploitent les ressources du jeu. Cette séparation est logique vu qu'il s'agit de joueurs qui maîtrisent les deux styles de jeu.

Concernant les joueurs assaut trouvés au bout de 2 heures de jeu, la majorité continue à rester dans ce style-là au bout de 3 heures de jeu. Ce style prend plus d'importance qu'avant grâce notamment à l'arrivée d'une grosse vague de joueurs qui étaient de style furtif qui exploitent Far Cry 5 au bout de 2 heures de jeu.

La même tendance se voit pour le style assaut qui exploite les ressources de Far Cry 5 grâce aux flux d'anciens joueurs de style furtif qui exploitent le jeu ainsi que du style polyvalent.

Entre 3h01 et 4 heures de jeu

Alors que lors du flux entre 2 et 3 heures, les joueurs de type furtif qui exploitent les ressources du jeu migraient en grande partie vers le style assaut ; entre 3h01 et 4 heures de jeu, les joueurs de type furtif qui exploitent Far Cry 5 se divisent en grande partie entre le même style de jeu ainsi que le style assaut qui utilisent les ressources du jeu.

La très grande majorité des joueurs assaut qu'on trouvait au bout de 3 heures de jeu conservent ce même groupe à la 4^{eme} heure de jeu.

Plus qu'une petite partie des joueurs de style furtif qui exploitent Far Cry 5 migrent vers le style assaut contrairement à ce que nous avions entre 2h01 et 3 heures de jeu.

Les joueurs polyvalents, auparavant formés part des joueurs de style furtif et as-

saut, sont cette fois-ci en très grande partie constitués de joueurs de type assaut qui exploitent les ressources du jeu. Cela explique plus facilement la différence que nous avions trouvé dans la partie 5.1, à savoir la tendance plus assaut qu'ont ces joueurs comparé à ce même groupe au bout de 2 heures de jeu.

Le groupe de style furtif qui exploite les ressources de Far Cry 5 s'est considérablement réduit au bout de 4 heures de jeu à cause d'une grande migration de ces joueurs vers le style assaut qui exploite le jeu.

Enfin, à cause du fait que la majorité des joueurs de style assaut qui exploitaient le jeu à 4 heures de jeu se retrouvent dans le groupe polyvalent, le nouveau groupe de joueurs assaut qui exploitent Far Cry 5 perd même un peu de terrain comparé aux heures précédentes.

Entre 4h01 et 5 heures de jeu

Encore une fois les joueurs de type assaut restent relativement fidèles à leur groupe avec juste une minorité qui part dans le groupe des joueurs furtifs qui exploitent Far Cry 5, ce qui est une tendance répétée depuis le début de l'analyse.

Il est intéressant de voir que cette fois-ci, seule une toute petite partie des joueurs furtifs qui exploitent le jeu conservent ce même groupe, la majorité migrant vers le style assaut qui exploite le jeu.

De plus, la même tendance se fait voir pour les joueurs de type assaut qui exploitent Far Cry 5 : une plus grande majorité est partie vers le style furtif qui exploite le jeu laissant l'autre minorité dans le même groupe.

Enfin, la quasi-entièreté des joueurs polyvalents ont migré vers le style furtif qui exploite le jeu. Cela montre que pendant la 4^{eme} heure de jeu, les joueurs polyva-

lents ont plutôt eu une plus grande affinité pour les armes furtives malgré la très bonne maîtrise des armes d'assaut.

Pour la dernière heure de jeu étudiée nous retrouvons ainsi 3 groupes de taille équilibrées. Les très nombreux flux présents au sein de ces 5 heures de jeu montrent de manière très claire que les joueurs sont à l'aise de naviguer d'un style à un autre d'heure en heure.

Comme le joueur est libre de faire le contenu du jeu dans l'ordre qu'il souhaite, il est très difficile d'analyser de manière générale ces flux, sauf si on étudie individu par individu afin de comprendre le contenu qu'ils ont fait.
Cependant, un moyen qui pourrait nous aiguiller serait une analyse de ces styles de jeu sur la carte du jeu Far Cry 5. En effet, comme nous l'expliquions, certains contenus de jeu peuvent privilégier un style plutôt qu'un autre, ce qui pourrait expliquer ces flux.

La représentation des styles de jeu sur la carte de Far Cry 5

Comme expliqué plus haut, le but de cette représentation est de voir si certaines parties de la carte de Far Cry 5 sont plus propices à un style de jeu qu'à un autre. De plus, comme nous l'avons expliqué dans la section 1.5, *Problématique*, les concepteurs de jeu pourront utiliser cet outil afin de voir comment les joueurs ont abordé leur contenu.
En effet, vu que ces derniers sont en charge du contenu de jeu, ils ont une connaissance accrue de la carte du jeu afin de s'assurer de bien répartir les missions, les lieux à découvrir, les missions annexes et les activités du monde ouvert dans le but de proposer un contenu diversifié.

Lors du développement du jeu, le concepteur de jeu essaye de se mettre dans la

peau du joueur pour essayer d'imaginer toutes les manières possibles de compléter son contenu. Ainsi, il n'est pas rare à la fin d'un test utilisateur, de fournir aux concepteurs de jeu les positions des joueurs sur la carte aux concepteurs de jeu afin qu'ils puissent s'assurer d'avoir bien anticipé (ou non) le comportement des joueurs lors du contenu joué.

Par exemple, pour appuyer ce point, il est possible que le concepteur de jeu imagine que son contenu (cela peut-être une mission, un lieu à découvrir, une mission annexe, etc) ne soit accessible sur la carte que par la route, la traversée d'une rivière ou par la forêt. Afin de s'assurer que le joueur ne puisse prendre que ces 3 options suivantes il met en place un terrain suffisamment raide qui fait glisser le personnage si jamais il essaye de grimper pour atteindre ce contenu.
Par la suite, il met en place toute la difficulté de son contenu autour de ces 3 portes d'accès (rivière, route ou forêt), c'est à dire qu'il concentre les ennemis à ces endroits stratégiques pour mettre plus de défi au joueur. Cependant, il est possible qu'en test utilisateur, un ou plusieurs joueurs aient réussi à trouver une route alternative, que le concepteur n'avait pas anticipé, pour atteindre le contenu.
Par conséquent, toute la stratégie que ce dernier a essayé de mettre en place dans son contenu ne fonctionne plus et le joueur peut accomplir la mission très facilement, par manque d'ennemis à cet endroit-ci ou encore utiliser un style de jeu que le concepteur n'avait pas anticipé (par exemple, il est possible que le concepteur fasse un contenu qui privilégie le style furtif en mettant un environnement où le joueur peut facilement se cacher).
A l'inverse, les positions du joueurs qui montrent le style de jeu permettraient aussi de confirmer ou non si ces derniers jouent le style que le concepteur a essayé de mettre en avant dans son contenu pour amener plus de défi.

La télémétrie permet d'enregistrer les actions des joueurs dans le jeu. L'une des actions enregistrée particulièrement intéressante pour analyser la consommation du

jeu est la position des joueurs dans le jeu. La position du joueur nous est envoyée toutes les 2 secondes de jeu, ainsi nous avons associé toutes les positions de chaque joueur avec son temps de jeu. Nous avons considéré que pendant les périodes de temps utilisées pour notre étude, le joueur conserve le même style de jeu.
Par exemple, entre 40 minutes et 1 heure de jeu, le joueur peut soit avoir un style polyvalent, furtif qui exploite le jeu ou assaut qui exploite Far Cry 5.

Ainsi, pour pouvoir représenter graphiquement l'évolution du style de jeu des joueurs sur la carte de Far Cry 5 nous avons donc décidé d'associer les positions des joueurs dans le jeu avec les groupes trouvés. En d'autres termes, toutes les positions dans le jeu entre 40 minutes et 1 heure de jeu sont soit des joueurs polyvalents, furtifs qui exploitent le jeu ou assauts qui exploitent Far Cry 5.
Ainsi, nous avons été capables de représenter facilement l'évolution des groupes sur les 5 premières heures de jeu.

Cette représentation nous permet de voir que certaines parties de la carte sont plus exploitées par un style de jeu bien défini. Les concepteurs du jeu peuvent donc savoir comment les missions ont été abordées par les joueurs.
Avant d'aller plus loin, la carte se divise en 5 parties :

— Le tutoriel : qui se trouve en plein centre, il s'agit de la plus petite des deux presque-îles présentent sur la carte.

— La région de Faith : Un des 3 personnages importants que le joueur doit tuer pour avancer dans le jeu.

— La région de John : Un des 3 personnages importants que le joueur doit tuer pour avancer dans le jeu.

— La région de Jacob : Un des 3 personnages importants que le joueur doit tuer pour avancer dans le jeu.

— La région de Joseph : Il s'agit de l'ennemi le plus important que le joueur

rencontre une fois que les 3 personnages précédents ont été tués (John, Jacob et Faith).

La figure 5.13 nous donne un point de repère visuel des 5 régions présentées :

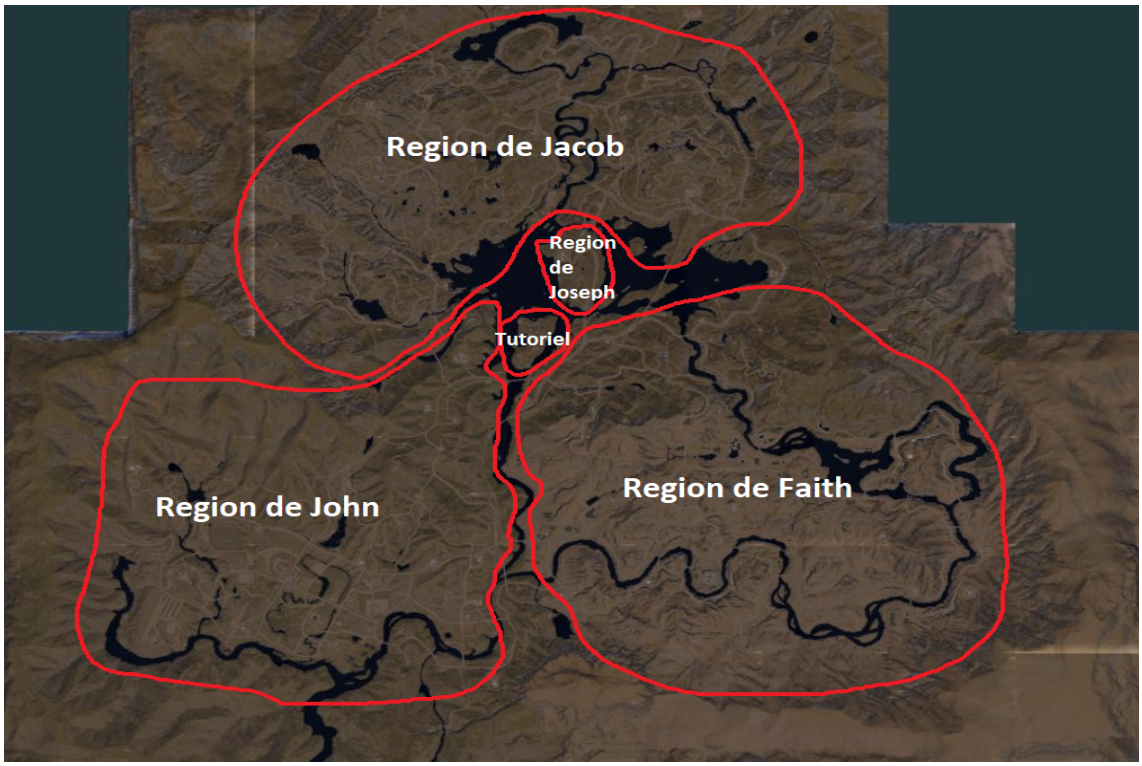


FIGURE 5.13 – Présentation des principales régions de Far Cry 5.

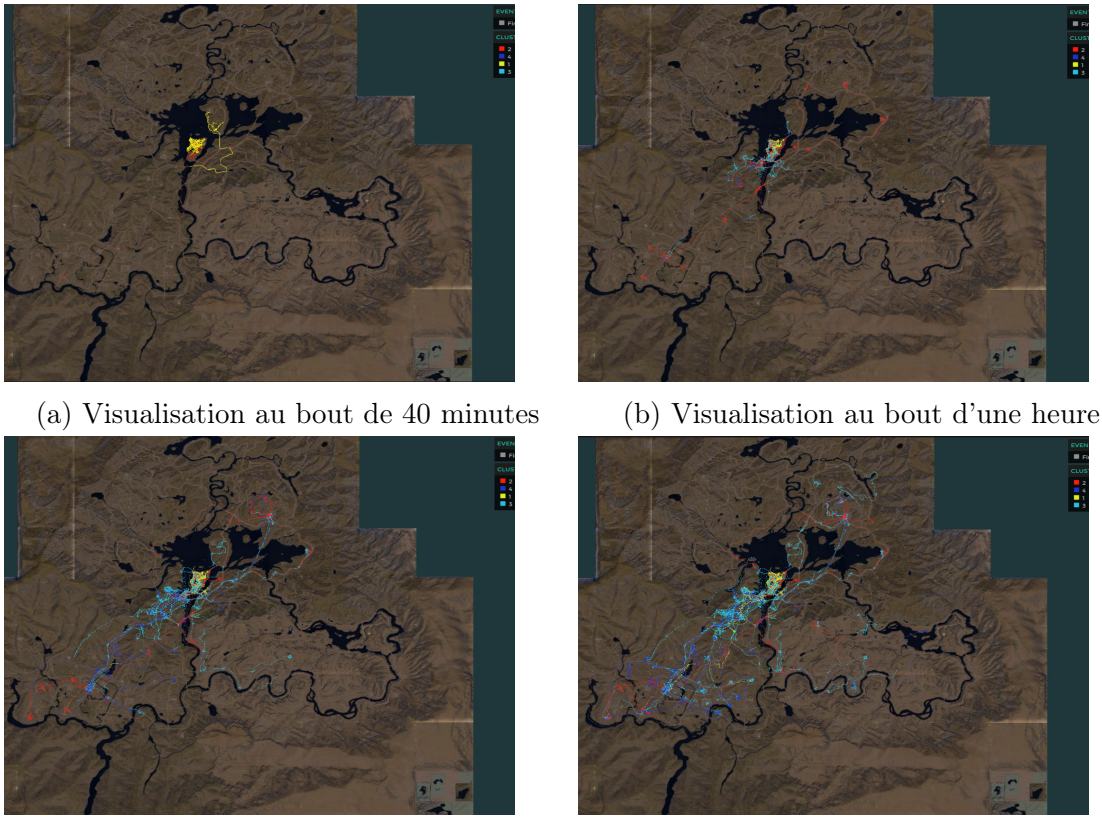
Ainsi, un découpage de la vidéo finale de l'évolution des styles de jeu sur la carte est présenté ci-dessous (voir les figures 5.14 et 5.15). Les trajectoires représentent les positions des joueurs dans le jeu et le code couleur est le suivant :

— **Jaune** : Joueur type furtif qui exploite les ressources de Far Cry 5 (afin de ne pas surcharger les images nous avons englobé le style furtif dans le style furtif qui exploite les ressources de Far Cy 5 car ce style disparaît définitivement après 40 minutes de jeu)

— **Rouge** : Joueur Assaut qui exploite les ressources de Far Cry 5

— **Bleu ciel** : Joueur Polyvalent

— **Bleu marine** : Joueur Assaut



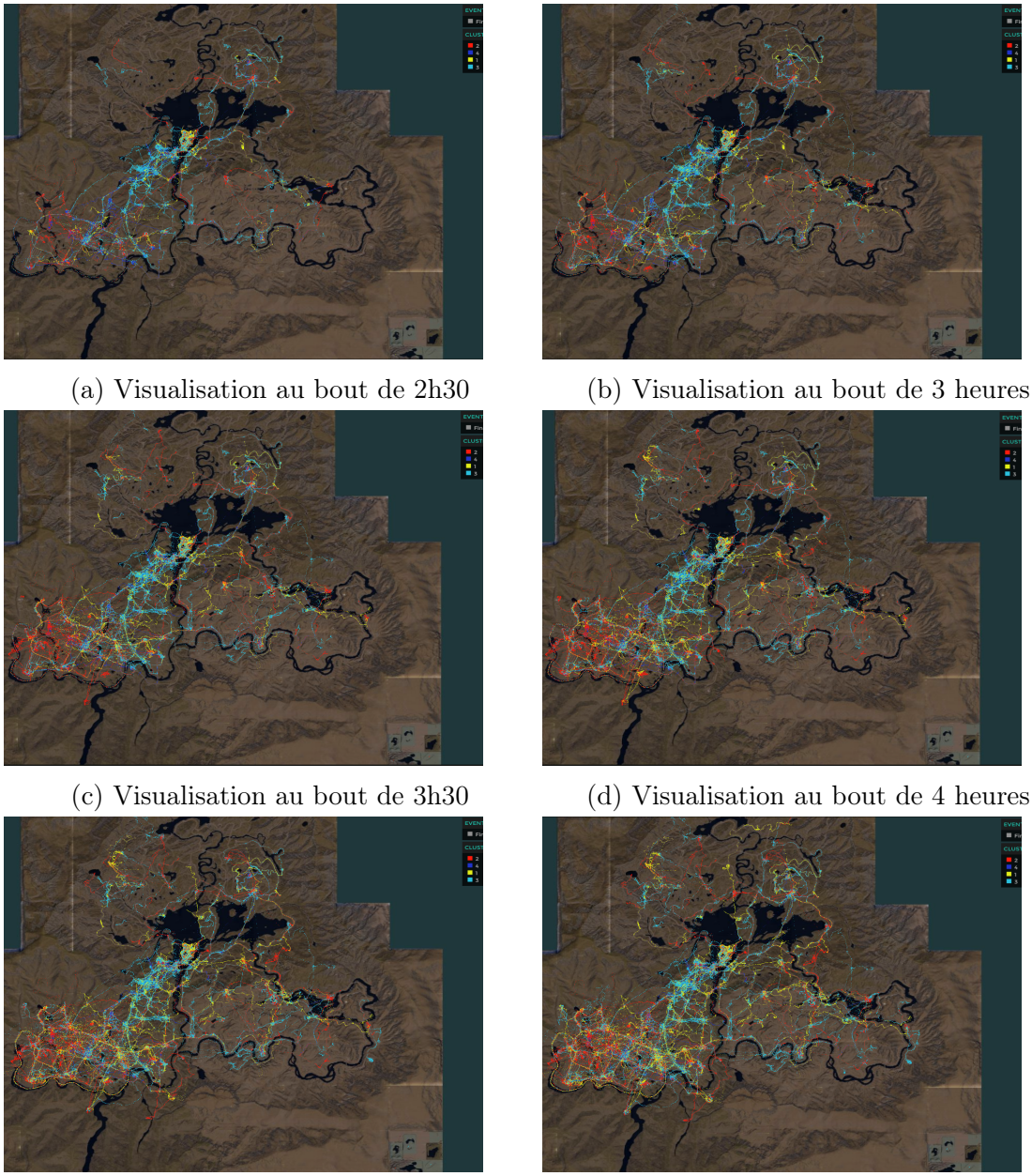
(a) Visualisation au bout de 40 minutes

(b) Visualisation au bout d'une heure

(c) Visualisation au bout de 1h30

(d) Visualisation au bout de 2 heures

FIGURE 5.14 – Représentation des styles de jeu sur la carte de Far Cry 5 entre 40 minutes et 2 heures de jeu



(a) Visualisation au bout de 2h30

(b) Visualisation au bout de 3 heures

(c) Visualisation au bout de 3h30

(d) Visualisation au bout de 4 heures

(e) Visualisation au bout de 4h30

(f) Visualisation au bout de 5 heures.

FIGURE 5.15 – Représentation des styles de jeu sur la carte de Far Cry 5 entre 2h30 et 5 heures de jeu

Au bout de 40 minutes

Le tutoriel se déroule sur une petite presque île en plein centre de la carte, la première visualisation (a) présentée dans la figure 5.14 montre bien que la majorité des contenus du tutoriel ont été fait en mode furtif. Cependant, nous remarquons que la partie sud de la presque île privilégie le type assaut. Quelques joueurs ont réussit à terminer le tutoriel en moins de 40 minutes vu les positions enregistrées en dehors de cette presque-île.

Au bout d'une heure de jeu

Le tutoriel est terminé et tous les joueurs se retrouvent en dehors de la zone du tutoriel. Nous remarquons de nombreuses positions enregistrées dans la région de John, ce qui montre que, malgré la liberté des joueurs de faire le contenu qu'ils souhaitent après le tutoriel, il y a une tendance pour se diriger dans la même zone. Cela s'explique par le fait qu'après le tutoriel, le scénario invite le joueur à se diriger dans cette région en particulier. Cependant, nous remarquons aussi une tendance à aller vers la région de Jacob.

Le style de jeu préconisé une fois sortie du tutoriel est le style polyvalent (en bleu ciel). En opposition, plus le joueur s'éloigne de la zone du tutoriel, plus le style assaut qui exploite les ressources de Far Cry 5 (en rouge) est utilisé. Ce qui montre que le joueur privilégie ce style de jeu pour explorer le monde.

Entre 1h01 et 2 heures de jeu

Au bout d'une heure et demi, nous remarquons que le style polyvalent se développe majoritairement dans la région de John ainsi qu'un peu dans les régions de Jacob et Faith ce qui montre que ce style est utilisé pour explorer le monde.
Au bout d'une heure et demie de jeu, le style assaut est utilisé (bleu foncé) lorsque le joueur se rapproche du sud de la région de John.

Il faut attendre 2 heures de jeu pour que le style furtif qui exploite Far Cry 5 apparaisse. Ce style se développe surtout dans la région de John.

Entre 2h01 et 3 heures de jeu

Au bout de deux heures et demi de jeu, le style furtif qui exploite Far Cry 5 se répand dans les 3 régions, surtout dans celle de John. Ce même style devient le style le plus répandu dans la région de Faith au bout de 3 heures de jeu. Le style assaut qui exploite Far Cry 5 s'intensifie dans le sud-ouest de la région de John, ainsi qu'un peu dans la partie sud de la région de Jacob.

Entre 3h01 et 4 heures de jeu

Au bout de trois heures et demi de jeu, la région de John continue à être visitée notamment avec le style furtif et assaut qui exploitent Far Cry 5.
Le style polyvalent est surtout utilisé dans la région de Faith.
La région de Jacob est toujours la région la moins visitée mais où la zone nord contient surtout le style assaut qui exploite Far Cry 5 alors que tout le long de la zone ouest (du nord vers le sud) est exploité par le style furtif et polyvalent.
Entre 3h30 et 4 heures de jeu, c'est le style Polyvalent qui a surtout été utilisé

Entre 4h01 et 5 heures de jeu

Au bout de 4h30 de jeu, des joueurs continuent à explorer la région de John en descendant encore plus au sud avec le style furtif et assaut qui exploitent Far Cry 5. Dans les autres régions, c'est surtout le style furtif qui exploite Far Cry 5 a été utilisé pour explorer un peu plus Faith et Jacob.
Aussi, la région de Joseph commence à être visitée avec notamment un point important qui montre l'utilisation du style furtif qui exploite Far Cry 5. Ce point peut représenter une mission en particulier à faire sur cette île.

Ainsi, au bout de 5 heures de jeu nous remarquons que les joueurs ont surtout visité la région de John et préfèrent utiliser le style polyvalent et furtif dans le nord de cette région et le style assaut qui exploite Far Cry 5 dans le sud. Nous remarquons aussi dans cette région de nombreux points d'intérêt qui représentent l'accumulation de positions des joueurs à ces endroits précis. Ainsi, dans le sud de la région, de nombreux points sont rouge et jaune, ce qui équivaut sûrement à des missions en particulier disponibles dans cette région-ci.

A l'inverse dans la région de Faith, les styles polyvalent et furtif qui exploite Far Cry 5 sont privilégiés. Le style assaut qui exploite Far Cry 5 est tout de même utilisé dans le côté est de la région.

La région de Jacob est la moins visitée et les joueurs semblent avoir plutôt avoir visité les côtés de cette zone en privilégiant les style polyvalent et furtif. Cependant, le style assaut qui exploite les ressources de Far Cry 5 est plus présent que dans la région de Faith.

Chapitre 6

Conclusion

Dans cette partie-ci, nous résumons les principaux résultats : la pertinence de l'utilisation d'une méthode hiérarchique ascendante sur un ensemble de jeux de données pour représenter l'évolution du style du joueur au cours du temps ainsi que l'utilisation de deux représentations graphiques pour accompagner les résultats de la classification. Nous mentionnerons ensuite les limites de l'étude ainsi qu'une ouverture pour les futurs travaux

6.1 Résumé des enjeux de la problématique

L'objectif de ce mémoire était de mettre en place des outils permettant de mieux représenter l'évolution du style de jeu qu'un joueur peut avoir au cours du temps. Notre étude possède plusieurs particularités :

Avant toute chose, il s'agit d'un développement de modèle applicable immédiatement en entreprise et qui peut facilement être reproduit d'un jeu à un autre. En effet, une implantation standard est requise afin de pouvoir rapidement appliquer le modèle sur n'importe quel projet. Pour cela, nous avons utilisé exclusivement des données de télémétrie que nous retrouvons dans tous les *FPS*. Effectivement, les variables comme le nombre de victimes provoquées par un certain type d'armes,

le temps passé dans les missions ou activités annexes sont des variables que nous retrouvons dans d'autres *FPS*. Ainsi, tout en prenant en considération les variables spécifiques à chaque jeu, notre étude peut être répliquée sur d'autres jeux. Cela permet ainsi de plus facilement comparer un jeu à un autre, aspect important pour une entreprise qui veut comparer ses différents produits entre eux.

Un deuxième aspect important à prendre en considération est le développement de visualisations adéquates pour répondre encore mieux à la problématique. En effet, en entreprise, il est important de bien savoir présenter ses données et ses résultats. Afin d'y arriver il est recommandé d'utiliser des outils qui mettent en évidence ce que les résultats reflètent. Ainsi, il était important de consacrer toute une section sur des méthodes de visualisation qui peuvent être utilisées pour représenter l'évolution du style de jeu du joueur au cours du temps.

6.2 Résumé des modèles utilisés

Le procédé de la segmentation

Comme nous venons tout juste de le dire, il était donc primordial de développer un modèle simple à expliquer, à mettre en place et à reproduire d'un projet à un autre. Ainsi, comme nous l'avons déjà expliqué au cours de cette étude, la segmentation semble être une bonne solution. En effet, la segmentation est une méthode très utilisée en entreprise car en plus d'être peu contraignante à mettre en place, elle est aussi simple à expliquer.
De plus, toute entreprise veut connaître ses clients, les catégoriser en groupe ou segment afin d'aider dans les stratégies marketing à prendre.
En d'autres termes, il s'agit du choix numéro 1 à prendre lorsque le but d'une étude est de comprendre le profils des individus. Comme nous l'avons vu dans la Section 2.2, *L'analyse par regroupement*, il existe un nombre exhaustif de revues de littéra-

ture traitant ce sujet-là. Enfin, il existe de très nombreuses méthodes au sein même de la segmentation, ce qui permet de trouver, pour chaque spécificité que la base de données peut offrir, une méthode adéquate.

Les méthodes retenues

Comme expliqué, au sein même de la segmentation de très nombreux modèles existent. Comme nos jeux de données ne présentaient que des données quantitatives, cela n'a pas aidé à réduire le choix des modèles. De plus, nous voulions procéder à une méthode d'apprentissage non-supervisé, nous nous sommes intéressés à deux modèles souvent vu dans la littérature : la méthode non-hiérarchique et la méthode hiérarchique.

L'une des deux méthodes est une hiérarchie ascendante sans ACP. Vu que cette méthode commence avec autant de groupes que d'individus à classer pour ne finir plus qu'avec un seul groupe regroupant tous les joueurs, elle permet de produire des groupes plus détaillés de meilleure qualité

La deuxième est une méthode hybride de plus en plus utilisée dans la revue de littérature qui combine une méthode des K-Moyennes pour réduire le temps de calcul ainsi qu'une méthode hiérarchique pour avoir des groupes de qualité tout en utilisant une ACP. Nous avons donc décidé de faire l'étude de ces deux méthodes et de voir l'impact que cela pouvait avoir sur des petits jeux de données.

Afin de définir quel modèle performait le mieux dans notre cas, nous avons décidé de classifier manuellement 20% de nos joueurs et voir comment nos modèles les classifiaient. Il en est sorti que la méthode hiérarchique ascendante sans ACP classifiait mieux les joueurs que la méthode hybride avec ACP.

Le choix des visualisations

Le but de cette étude est à la fois pour les concepteurs de jeu et les équipes business qui souhaitent voir comment le jeu est joué.
Comme cite Confucius « Une image vaut mille mots », le choix des visualisations a été un élément important à prendre en considération : en effet, la visualisation doit apporter des réponses aux résultats présentés et non plus de questions. Pour cela, nous nous sommes intéressés à deux types de représentations très différentes qui enrichissent les résultats trouvés par notre segmentation.
D'une part, les concepteurs connaissent parfaitement l'emplacement des missions dont ils sont en charge dans le jeu et doivent anticiper les entrées possibles du joueur dans une mission (par exemple ne pas oublier de mettre des ennemis sur des accès que le joueur pourrait prendre pour atteindre la mission afin de ne pas la rendre trop facile). Ainsi, une représentation visuelle du déplacement des joueurs est un moyen efficace pour fournir une information pertinente au concepteur de jeu.

Dans notre cas, il a été décidé de donner une représentation graphique de l'évolution du style des joueurs selon les positions qu'ils ont dans le jeu. Cette visualisation permet donc au concepteur de jeu de voir comment ses missions sont en règle générale jouées et peut donc aussi détecter si un style de jeu prône à certains endroits afin de pouvoir réajuster cela pour donner plus de possibilités au joueur. En effet, comme nous l'avons dit dans le chapitre 5, section 5.2.2 *La représentation des styles de jeu sur la carte de Far Cry 5*, lors du développement du jeu, le concepteur de jeu essaye de se mettre dans la peau du joueur pour essayer d'imaginer toutes les manières possibles de compléter son contenu. Ainsi, il n'est pas rare à la fin d'un test utilisateur, de fournir aux concepteurs de jeu les positions des joueurs sur la carte aux concepteurs de jeu afin qu'ils puissent s'assurer d'avoir bien anticipé (ou

non) le comportement des joueurs lors du contenu joué.

D'autre part, les équipes business veulent souvent connaître ses joueurs et la manière dont ils jouent au jeu. Cependant, ces équipes ont besoin de pouvoir rapidement comprendre les résultats. Ce défi peut être facilement résolu avec la visualisation que nous avons choisi : le diagramme de Sankey. Cette visualisation nous permet de pouvoir montrer l'évolution des styles de jeu sur les 5 premières heures de jeu en plus de montrer le flux de joueurs qui passent d'un style à un autre au cours du temps.

6.3 Résumé des résultats obtenus

Les résultats obtenus ont montré que le style de jeu des joueurs de Far Cry 5 peut se découper en 3 ou 4 groupes selon l'avancée dans le jeu. D'une heure à une autre les types de groupes restent relativement les mêmes avec au moins trois principaux groupes qu'on retrouve en tout temps : le joueur assaut qui exploite les ressources de Far Cry 5, le joueur de type furtif qui possède toujours le meilleur ratio d'ennemis tués par mort enregistrée et qui utilise au mieux les ressources disponibles dans Far Cry 5 ainsi que le joueur typiquement assaut qui n'exploite pas les ressources de Far Cry 5. Un dernier groupe a fait son apparition au bout de 2 heures et 4 heures de jeu : le joueur polyvalent. Il s'agit d'un joueur qui maîtrise aussi bien l'approche furtive que l'assaut avec une légère préférence pour l'assaut.
Même si la définition de ces groupes reste la même au cours du temps, nous avons remarqué, grâce au diagramme de Sankey, un flux important de joueurs d'un groupe à un autre.
Notamment, de nombreux joueurs qui étaient de type furtif à un moment devenaient assaut l'heure suivante ou vice-versa. Ces passages d'un style à un autre nous permet de dire que les joueurs sont à l'aise de passer d'un style à un autre. Aussi,

nous remarquons que la migration du style furtif au style assaut est plus facile que l'inverse. Cela peut notamment s'expliquer que le style furtif demande plus d'armes appropriées et une plus grande stratégie de jeu.

Comme le joueur est libre de faire le contenu du jeu dans l'ordre qu'il souhaite,il est très difficile d'analyser de manière générale ces flux, sauf si on étudie individu par individu afin de comprendre le contenu qu'ils ont fait. La représentation des styles de jeu sur la carte de *Far Cry 5* permet de mieux cerner les zones à tendances assaut ou furtive.
Ainsi, nous remarquons qu'au bout de 5 heures de jeu :

— Dans la région de John : il s'agit de la région la plus visitée. Le sud de la région se partage entre le type furtif et assaut qui exploitent les ressources de Far Cry 5. Cependant, jusqu'à 3 heures de jeu, le style polyvalent était privilégié, surtout du nord au centre de la région.

— Dans la région de Faith : même si cela semble difficile à voir à cause de la résolution de l'image, le style furtif semble légèrement dominer dans cette région-ci. Les bords de la région qui longent les rivières sont les parties les plus visitées pour après 5 heures de jeu.

— Dans la région de Jacob : cette région reste la moins visitée au profit des 2 autres régions. Les styles furtif et assaut qui exploitent les ressources de Far Cry 5 ne se mélangent pas, c'est à dire que dans certaines parties de la région, nous voyons essentiellement que de l'assaut et dans d'autres du style furtif.

6.4 Les limitations de la méthodologie utilisée et futurs travaux

En présentant nos résultats nous avons parlé des avantages des méthodes utilisées dans le cadre de notre étude. Néanmoins, il est important de noter qu'il existe des

limitations.

La télémétrie : à la fois une information véridique et trompeuse

Comme les jeux Triple A proposent de plus en plus de flexibilité quant au *gameplay*, de très nombreuses variables rentrent en jeu et peuvent donc influencer la télémétrie. Par exemple, un joueur peut très bien être détecté par un ennemi et décider de le tuer presque au corps-à-corps avec une arme de type stealth : dans le cadre de l'observation, le modérateur en aurait conclu qu'il s'agissait d'une action de type assaut alors que la télémétrie considérerait cela comme une action de type furtive.
En d'autres termes, même si les données quantitatives permettent de mettre en place des modèles statistiques qui peuvent classifier le style de jeu du joueur, les données qualitatives, comme les remarques des joueurs ou l'observation des modérateurs, permettent d'avoir une représentation sans faille du style du joueur. Il serait donc intéressant de pouvoir utiliser des données qualitatives pour mieux appréhender le style de jeu du joueur.

La méthodologie : la segmentation trop rigide

Comme expliqué dans la revue de littérature, les méthodes que nous avons étudiées dans le cadre de l'étude sont des méthodes exclusives, c'est à dire qu'un individu ne peut appartenir qu'à un seul groupe. Or, le joueur adapte souvent son style de jeu selon la difficulté et les ennemis qu'il rencontre. Ainsi, il est fort probable qu'au sein d'une même mission ce dernier mélange le style assaut et furtif. Xie and Beni (1991) parlent de méthodes non-exclusives comme la mixture de Gaussienne (plus communément appelée *fuzzy clustering*) qui tolère justement qu'un individu appartienne à plusieurs groupes. Cette méthode, plus proche du comportement réel d'un individu permettrait de donner des résultats différents de ceux obtenus avec une

méthode exclusive. Il serait intéressant à l'avenir d'essayer de faire une classification dans le temps avec une méthode non-exclusive ou encore utiliser cette méthode sur certaines missions pour donner encore plus de contenu aux concepteurs de jeu.
Aussi, il aurait été intéressant de reproduire la même étude avec un nombre fixe de groupes au cours du temps. Ainsi, une comparaison de l'évolution du style de jeu avec toujours 2 groupes, 3 groupes ou 4 groupes aurait pu nous donner des conclusions intéressantes et sans doute différentes de celles que nous avons développé ici. Par exemple, il est possible qu'en forçant le modèle a trouver plus de groupes que ceux retenus lors de notre étude, on trouverait des groupes composés de très peu de joueurs mais qui ont un style de jeu bien défini et bien différent des autres groupes. En d'autres termes, plus de groupes nous permettrait de trouver des comportements atypiques dans le style de jeu.

Robustesse de l'interprétation

Notre méthode utilisée se base exclusivement que sur des données quantitatives. Ces données récoltées ont comme principale caractéristique d'être retrouvées d'un jeu à un autre. En effet, tous les jeux récoltent les informations sur les ennemis tués, les types d'armes utilisées, les headshots, etc... Seules certaines variables sont caractéristiques du jeu en question comme les mercenaires ou le temps passé dans les différentes régions. Ainsi, outre ces variables plus spécifiques à Far Cry 5, le modèle pourrait être facilement réutilisable sur un autre jeu *FPS* en ajoutant des variables spécifiques au jeu. Cependant, l'analyse des groupes est indépendante d'un jeu à un autre : toute l'analyse serait à refaire sur un autre *FPS*.

De plus, comme nous l'avons longuement évoqué dans le chapitre 2, dans la sous-section 2.2.3 *Détermination du nombre de groupes*, le nombre de groupes reste la partie la plus délicate dans l'analyse par regroupement. En d'autres termes, il existe une certaine part de subjectivité dans l'analyse vu que l'analyste prend la décision

du nombre de groupes pour former des groupes interprétables dans le cadre de son étude. Toujours dans cet ordre d'idée, comme le style de jeu est encore un phénomène peu étudier dans le milieu universitaire, nous nous basons sur des hypothèses pour former et analyser nos groupes. Par exemple, nous prenons pour acquis qu'un joueur qui utilise une arme de type furtif a forcément joué furtif. Ou encore, ce qui peut paraître correcte comme classification d'armes de type furtif pour l'analyste peut être vu différemment par un autre analyste.

Enfin, un autre problème est le manque de validation humaine. Comme le modèle se base intégralement sur des données quantitatives prélevées de la télémétrie dans le jeu, il est possible que certaines valeurs soient erronées à cause d'éventuels bugs de jeu ou encore que le joueur adopte une stratégie de jeu que nous allons mal classifier par la suite. Pour illustrer cela, un joueur pourrait utiliser une arme de type furtif lors d'un combat assaut.

Application en entreprise

Ce modèle a été appliqué sur le jeu Far Cry 5 et compte être utilisé sur d'autres projets non annoncés. De part l'intérêt que cela provoque pour l'équipe analytique, l'équipe d'analystes en recherche utilisateur ainsi que pour les concepteurs de niveaux, ce modèle sera sans doute amélioré afin de répondre à de futures demandes.

Bibliographie

Richard Bartle. Hearts, clubs, diamonds, spades : Players who suit muds. *Journal of MUD research*, 1(1) :19, 1996.

Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6) :1554–1563, 1966.

Boyan Bontchev and Olga Georgieva. Playing style recognition through an adaptive video game. *Computers in Human Behavior*, 82 :136–147, 2018.

Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1) :1–27, 1974.

camadegames.com. Aaa games, 2014. [Online ; Retrieved 2014-01-30].

Bernard Chen, Phang C Tai, Robert Harrison, and Yi Pan. Novel hybrid hierarchical-k-means clustering method (hk-means) for microarray analysis. In *Computational Systems Bioinformatics Conference, 2005. Workshops and Poster Abstracts. IEEE*, pages 105–108. IEEE, 2005.

Benjamin Cooley. Detecting learning styles in video games. 2015.

Douglass R Cutting, David R Karger, Jan O Pedersen, and John W Tukey. Scatter/gather : A cluster-based approach to browsing large document collections. In *ACM SIGIR Forum*, volume 51, pages 148–159. ACM, 2017.

Rusel DeMaria and Johnny L Wilson. *High score ! : the illustrated history of electronic games*, volume 1. McGraw-Hill/Osborne Berkeley, CA, 2002.

Anders Drachen, Alessandro Canossa, and Janus Rau Møller Sørensen. Gameplay metrics in game user research : Examples from the trenches. In *Game analytics*, pages 285–319. Springer, 2013.

Colas Duflo. *Jouer et philosopher*. Presses universitaires de France, 1997.

Mirko Ernkvist. Down many times, but still playing the game : Creative destruction and industry crashes in the early video game industry 1971-1986. 2008.

J-M Bouroche et G. Saporta. L'analyse de données. *Pour la Science*, 1985.

Eric Yergeau et Martine Poirier. Analyse en composantes principales. `http://spss.espaceweb.usherbrooke.ca/media/images/Site20v17/acp3.jpg`.

Brian S Everitt, Graham Dunn, et al. *Applied multivariate data analysis*, volume 2. Wiley Online Library, 2001.

Edward Ez'. L'histoire du jeu vidéo - la naissance du fps. `https://www.youtube.com/watch?v=DEQpNpPbUi0t=2s`, 2015a.

Edward Ez'. La démocratisation du fps. `https://www.youtube.com/watch?v=Ph63d9-X328`, 2015b.

Peter D. Jakab James S. Tieman Maurice R. Ferre. Position Tracking and Imaging system with error detection for use in medical applications. *Divisional of U.S*, (08/527.517), 1995. doi : https ://patentimages.storage.googleapis.com/e4/21/a3/b537f4a2176486/US5676673.pdf.

Ali Ghorbani and Sara Farzai. Fraud detection in automobile insurance using a data mining based approach. 2017.

Reynaldo J Gil-Garcia, José Manuel Badia-Contelles, and Aurora Pons-Porrata. A general framework for agglomerative hierarchical clustering algorithms. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 2, pages 569–572. IEEE, 2006.

JF Hair and B Babin Black. anderson, r. e.; tatham, r. l. et al. *Multivariate data analysis*, 1998.

André Hardy. On the number of clusters. *Computational Statistics & Data Analysis*, 23(1) :83–96, 1996.

John A Hartigan. Clustering algorithms. 1975.

J.E Jackson. *A User's Guide to Principal Components*. Wiley edition, 1991.

A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data. Prentice-Hall advanced reference series*. Prentice-hall, inc. edition, 1988.

I.T. Jolliffe. *Principal Component Analysis. 2nd Edition*. Springer edition, 2002.

Robert S. Jones. Home video games are coming under a strong attack. `//news.google.com/newspapers?nid=1320dat=19821212id=L2tWAAAAIBA Jsjid=q-kDAAAAIBAJpg=1609,4274079hl=en`, 1982. [Online; accessed 12-Dec-1982].

Martin T.E. Karr, R.J. Random numbers and principal components : further searches for the unicorn. 1981.

Alboukadel Kassambara. *Practical guide to cluster analysis in R : Unsupervised machine learning*, volume 1. STHDA, 2017.

Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data : an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

Kevin M Kieffer. Orthogonal versus oblique factor rotation : A review of the literature regarding the pros and cons. 1998.

Kolb. Experiential learning theory. In *Encyclopedia of the Sciences of Learning*, pages 1215–1219. Springer, 2012.

Wojtek J Krzanowski and YT Lai. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, pages 23–34, 1988.

Jacques Lallain. Le jour où : Le premier jeu vidéo. `https://www.youtube.com/watch?v=HBO-HYAVjHo`, 2015.

Denis Larocque. *Notes de cours : Analyse multidimensionnelle appliquée (6-602-07)*. HEC Montréal, Département des sciences de la décision, 2016.

Yi Lu, Shiyong Lu, Farshad Fotouhi, Youping Deng, and Susan J Brown. Incremental genetic k-means algorithm and its application in gene expression data analysis. *BMC bioinformatics*, 5(1) :172, 2004.

Emma McDonald. The global games market will reach $108.9 billion in 2017 with mobile taking 42%. `https://newzoo.com/insights/articles/the-global-games-market-will-reach-108-9-billion-in-2017-with-mobile-taking-42/`, 2017. [Online ; Retrieved 2017-April-20].

Glenn W Milligan and Martha C Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2) :159–179, 1985.

Mquantin. Illustration du déroulement de l'algorithme des k-means. `https://upload.wikimedia.org/wikipedia/commons/f/fb/ K-means.png`, 2017.

Fionn Murtagh. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4) :354–359, 1983.

nintendo.com. Licensed and unlicensed products. `https://www.nintendo.com/consumer/licensed.jsp`, 2018. [Online ; Retrieved 2014-01-30].

Mathieu Triclot Olivier Lejade. *La Fabrique des jeux vidéo : Au cœur du gameplay.* Éditions de la Martinière, 2013. ISBN 2732456373 et 978-2732456379.

OJ Oyelade, OO Oladipupo, and IC Obagbuwa. Application of k means clustering algorithm for prediction of students academic performance. *arXiv preprint arXiv :1002.2425*, 2010.

Marjorie A Pett, Nancy R Lackey, and John J Sullivan. *Making sense of factor analysis : The use of factor analysis for instrument development in health care research.* Sage, 2003.

R.A Pimentel. Morphometrics : the multivariate analysis of biological data. 1979.

Girish Punj and David W Stewart. Cluster analysis in marketing research : Review and suggestions for application. *Journal of marketing research*, pages 134–148, 1983.

Shiwani Rana and Roopali Garg. Application of hierarchical clustering algorithm to evaluate students performance of an institute. In *Computational Intelligence & Communication Technology (CICT), 2016 Second International Conference on*, pages 692–697. IEEE, 2016.

Gustavo Ratto, Guillermo J Berri, and Ricardo Maronna. On the application of hierarchical cluster analysis for synthesizing low-level wind fields obtained with a mesoscale boundary layer model. *Meteorological Applications*, 21(3) :708–716, 2014.

Alvin C Rencher. Principal component analysis. *Methods of Multivariate Analysis, Second Edition*, pages 380–407, 2002.

C. Shalizy. Distances between clustering, hierarchical clustering 36-350, data mining, 2009.

Weifang Shi and Weihua Zeng. Application of k-means clustering to environmental risk zoning of the chemical industrial area. *Frontiers of Environmental Science & Engineering*, 8(1) :117–127, 2014.

Padhraic Smyth. Clustering sequences with hidden markov models. In *Advances in neural information processing systems*, pages 648–654, 1997.

Lúcia Sousa and João Gama. The application of hierarchical clustering algorithms for recognition using biometrics of the hand. 2014.

Garton E.O. Steinhorst R.K. Stauffer, D.F. A comparison of principal components from real and random data. 1985.

Michael Steinbach, George Karypis, Vipin Kumar, et al. A comparison of document clustering techniques. In *KDD workshop on text mining*, volume 400, pages 525–526. Boston, 2000.

Scott Matthew Steinberg. *Videogame marketing and PR*. iUniverse, 2007.

Tufféry Stéphane. *Data mining et statistique décisionnelle : l'intelligence des données*. Editions Technip, 2012.

Catherine A Sugar and Gareth M James. Finding the number of clusters in a dataset : An information-theoretic approach. *Journal of the American Statistical Association*, 98(463) :750–763, 2003.

George Taylor. Why is gaming more popular than music and film ? `https://www.huffingtonpost.co.uk/george-taylor/why-is-gaming-more`

-popular-than-music-and-film_b_10095376.html, 2016. [Online; Retrieved 2016-May-23].

Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301) :236–244, 1963.

Sarah Boslaugh Paul Andrew Watters. *Statistics in a Nutshell, A desktop Quick Reference*. O'reilly edition, 2008.

Xuanli Lisa Xie and Gerardo Beni. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (8) :841–847, 1991.