

TABLE DES MATIÈRES

	Page
INTRODUCTION	1
CHAPITRE 1 JEUX SÉRIEUX	5
1.1 Introduction.....	5
1.2 Définition d'un jeu sérieux	5
1.3 Classification des jeux sérieux.....	6
1.4 Jeux sérieux dans le contexte de l'éducation	7
1.4.1 Les jeux sérieux éducatifs et l'intelligence artificielle	7
1.4.2 Les jeux sérieux éducatifs et la motivation des apprenants	8
1.4.3 Les jeux sérieux éducatifs et l'apprentissage collaboratif : Jeu 2.0	9
1.5 Jeux sérieux pour la stimulation cognitive.....	11
1.6 Adaptation des jeux sérieux	14
1.6.1 Définition de l'adaptation	14
1.6.2 Exemples d'études sur l'adaptation des jeux sérieux	14
1.7 Jeux sérieux et évaluation	17
1.8 Jeux sérieux et forage de données.....	18
1.8.1 Définition du forage de données (<i>Data Mining</i>).....	18
1.8.2 Étapes de forage de données	18
1.8.3 Classification des techniques de forage de données	19
1.8.4 Exemples d'études sur le forage de données des jeux sérieux.....	19
1.9 Conclusion	21
CHAPITRE 2 MÉTHODES DE FORAGE DE DONNÉES : RÉGRESSION LINÉAIRE ET REGROUPEMENT	23
2.1 Introduction.....	23
2.2 Modèles de régression linéaire.....	23
2.2.1 Régression linéaire simple	25
2.2.2 Régression linéaire multiple	26
2.2.3 Indicateurs de qualité des modèles de régression linéaire	27
2.3 Regroupement (<i>Clustering</i>)	28
2.3.1 Regroupement : définition et méthodes	28
2.3.2 Méthode de partitionnement K-moyennes (<i>Kmeans</i>)	30
2.3.3 Indicateurs de qualité du regroupement.....	31
2.4 Conclusion	33
CHAPITRE 3 SCÉNARIOS D'ANALYSE.....	35
3.1 Introduction.....	35
3.2 Scénarios d'analyse.....	35
3.2.1 Scénario d'analyse basé sur le jeu	35
3.2.2 Scénario d'analyse basé sur le joueur	36
3.2.3 Choix du scénario d'analyse	37

3.3	Conclusion	39
CHAPITRE 4 EXPÉRIMENTATION		41
4.1	Introduction.....	41
4.2	Prétraitement.....	41
4.2.1	Description du jeu sérieux « Science en jeu »	41
4.2.2	Description de la base de données «Scienceenjeu»	43
4.2.3	Description des outils.....	44
4.2.4	Préparation des données.....	45
4.2.4.1	Transformation des données	46
4.2.4.2	Détection et élimination des valeurs aberrantes.....	47
4.2.4.3	Détection et élimination des valeurs extrêmes.....	48
4.3	Modélisation : modèle de régression linéaire multiple	54
4.3.1	Contexte de l'expérimentation.....	54
4.3.1.1	Choix de la méthode de régression et critères d'évaluation du modèle.....	54
4.3.1.2	Variables dépendantes et variables indépendantes	54
4.3.1.3	Échantillonnage.....	55
4.3.2	Corrélation entre les variables	56
4.3.3	Résultats de l'application du modèle de régression linéaire multiple.....	60
4.3.3.1	Résultats de la régression linéaire multiple pour le nombre de neurones	60
4.3.3.2	Résultats de la régression linéaire multiple pour le nombre de talents	61
4.3.3.3	Résultats de la régression linéaire multiple pour le nombre de cortex.....	62
4.3.3.4	Résultats de la régression linéaire multiple pour la durée totale des sessions	62
4.3.4	Analyse des résultats.....	63
4.3.5	Interprétation des résultats	65
4.4	Modélisation : Regroupement (<i>Clustering</i>)	66
4.4.1	Contexte de l'expérimentation.....	66
4.4.2	Résultats de l'application du regroupement avec K-moyennes.....	67
4.4.2.1	Résultats du regroupement pour le nombre de neurones	68
4.4.2.2	Résultats du regroupement pour le nombre de talents	70
4.4.2.3	Résultats du regroupement pour le nombre de cortex	71
4.4.2.4	Résultats du regroupement pour la durée totale des sessions	75
4.4.3	Analyse des résultats.....	77
4.4.4	Interprétation des résultats	78
4.5	Conclusion	79
CONCLUSION.....		81
RECOMMANDATIONS		85
ANNEXE I	CODIFICATION DE L'ATTRIBUT INTÉRÊT SCIENTIFIQUE	87

ANNEXE II	GRAPHIQUES DE DISTRIBUTION DES VALEURS DES ATTRIBUTS.....	89
ANNEXE III	RÉSULTATS DE L'APPLICATION DU MODÈLE DE RÉGRESSION LINÉAIRE MULTIPLE (SUITE)	97
ANNEXE IV	GRAPHIQUES DE L'APPLICATION DU MODÈLE DE RÉGRESSION LINÉAIRE MULTIPLE SUR LES SOUS-POPULATIONS.....	101
ANNEXE V	RÉSULTATS DE L'APPLICATION DU MODÈLE DE RÉGRESSION LINÉAIRE MULTIPLE SUR PLUSIEURS ÉCHANTILLONS.....	105
ANNEXE VI	RÉSULTATS DE L'APPLICATION DU MODÈLE DE REGROUPEMENT	107
ANNEXE VII	DÉSCRIPTION DU JEU SÉRIEUX « SCIENCE EN JEU »	113
	LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES.....	117

LISTE DES TABLEAUX

	Page
Tableau 2.1	Liste des paramètres utilisés dans les définitions de la régression linéaire 24
Tableau 4.1	Les attributs présélectionnés pour l'étude..... 44
Tableau 4.2	Les attributs sélectionnés pour l'étude..... 48
Tableau 4.3	Subdivision de l'ensemble de données en sous-populations 53
Tableau 4.4	Variables indépendantes 55
Tableau 4.5	Variables dépendantes 55
Tableau 4.6	Coefficients de corrélation entre les variables Langue, Âge, Sexe, Intérêt scientifique, Nb_visite, Abonnement, Nb_quêtes, Nb_avantages, Nb_victoire et Nb_defaite..... 57
Tableau 4.7	Coefficients de corrélation entre les variables dépendantes et les variables indépendantes..... 59
Tableau 4.8	Coefficients de corrélation entre les variables Nb_neurones, Nb_talents, Nb_cortex, Duree_totale et Duree_moyenne..... 59
Tableau 4.9	MRLM avec Nb_neurones pour la sous-population Vs01Nr1 60
Tableau 4.10	MRLM avec Nb_neurones pour la sous-population Vs70Nr 61
Tableau 4.11	MRLM avec Nb_talents pour la sous-population Vs01T11 61
Tableau 4.12	MRLM avec Nb_cortex pour la sous-population Vs70Cr 62
Tableau 4.13	MRLM avec Duree_totale pour la sous-population Vs70Dt 63

LISTE DES FIGURES

	Page
Figure 1.1	Architecture d'un jeu éducatif (<i>learning game</i>) 2.0.....11
Figure 1.2	Couche d'abstraction jeu-système tutoriel.....16
Figure 4.1	Ile d'entrée du jeu « Science en jeu »42
Figure 4.2	Neurones, quêtes et avantages dansVs01.....68
Figure 4.3	Neurones, langue et visites dansVs20.....69
Figure 4.4	Neurones, visites et avantages dans Vs70.....70
Figure 4.5	Talents, quêtes et avantages dansVs0171
Figure 4.6	Cortex, visites et avantages dansVs0172
Figure 4.7	Cortex, visites et quêtes dansVs20.....73
Figure 4.8	Cortex, quêtes et avantages dansVs70.....74
Figure 4.9	Durée totale, visites et quêtes dansVs01.....75
Figure 4.10	Durée totale, sexe et visites dans Vs70.....76

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

BD	Base de données
CHUM	Centre Hospitalier de l'Université de Montréal
F-stat	F-statistique
IA	Intelligence artificielle
JS	Jeux sérieux
MRLM	Modèle de régression linéaire multiple
R^2	Coefficient de détermination
RLM	Régression linéaire multiple
TDAH	Trouble du déficit de l'attention avec ou sans hyperactivité

INTRODUCTION

Les jeux vidéo connaissent actuellement un grand succès. Ils sont devenus un moyen de divertissement incontournable et très répandu parmi les jeunes. À partir du caractère divertissant et attrayant des jeux vidéo auquel s'ajoute un but de formation ou d'entraînement, une nouvelle catégorie de jeux appelée les jeux sérieux (JS) s'est développée. C'est une voie prometteuse de développement et d'innovation qui a engendré de nombreuses applications dans l'éducation, la santé, la défense, l'industrie, etc.

Plusieurs études portant sur les jeux sérieux dans le domaine de la santé mentale, plus précisément la stimulation cognitive, démontrent leur potentiel à devenir des outils efficaces pour les thérapeutes et les patients. Les résultats encourageants des études effectuées auprès de personnes souffrantes de maladies telles que : la maladie d'Alzheimer (Imbeault, 2011), l'autisme (Sehaba, 2005) ou les troubles du comportement (Coyle, 2005) (Conconi, 2008), ont inspiré l'idée d'utiliser les jeux sérieux comme outil d'aide à la stimulation cognitive avec de jeunes patients atteints de psychose (Wartena, 2013).

Dans cette même perspective, des chercheurs de l'ÉTS et du centre hospitalier de l'Université de Montréal (CHUM) ont initié un projet multidisciplinaire de développement d'outils technologiques d'aide à la stimulation cognitive pour de jeunes adultes atteints de troubles psychotiques, en collaboration avec des partenaires industriels (développement de jeux vidéo et jeux sérieux). Notre travail s'inscrit dans ce cadre de recherche et tente de relever l'un des nombreux défis dont fait face ce projet ambitieux. Ainsi, nous nous sommes intéressés spécifiquement dans la présente étude à la problématique des données issues des jeux. En effet, une grande partie de la réponse à la question de la pertinence et de l'efficacité de ces outils viendrait de l'analyse des données des sessions de jeu et ceux des joueurs. Il s'agit d'un grand volume de données pour lequel il faudra déterminer :

- Quel est l'objectif de l'analyse?
- Quelles sont les méthodes d'analyse à appliquer?
- Quelles sont les données à sélectionner pour l'étude et de quel type?

- Quelle interprétation fait-on des résultats obtenus?

Nous proposons, dans cette étude, de faire l'analyse des données des sessions du jeu sérieux « Science en jeu » qui est une plateforme de jeux destinée à l'apprentissage des sciences et qui offre aussi des activités de stimulation cognitive. Par conséquent, il peut être utilisé dans les contextes de l'éducation et de l'entraînement des fonctions cognitives et les résultats de notre étude peuvent être appliqués et exploités dans les deux contextes aussi. En plus, « Science en jeu » a été choisi pour être testé par des patients atteints de psychose à l'hôpital Notre-Dame du CHUM et fait l'objet d'une étude de convivialité dans le cadre d'une autre étude.

L'objectif de ce travail est d'exploiter les méthodes de forage de données appropriées afin: 1) d'identifier des données pertinentes du jeu et du joueur, et les méthodes de forage de données appropriées pour les étudier et 2) de déduire des caractéristiques des profils des joueurs dans la perspective de la personnalisation du jeu.

Différentes techniques de forage de données sont appliquées pour analyser les données des jeux vidéo et des JS et déterminer par exemple : les profils de joueurs, leurs styles de jeu, leur motivation ou la prédiction de leurs performances. Dans notre cas, nous avons sélectionné les données les plus influentes sur les performances des joueurs grâce au modèle de régression linéaire multiple (MRLM). Ensuite, nous avons utilisé la technique de partitionnement (méthode de classification non supervisée) sur ses données pour explorer l'existence de regroupements et déduire les caractéristiques des profils des joueurs.

Notre travail s'organise autour de quatre chapitres. Le premier chapitre définit les JS et passe en revue les études portant sur leur utilisation dans les domaines de l'éducation et de la stimulation cognitive. Il traite aussi de l'adaptation et de l'évaluation de ces jeux et donne un aperçu sur le forage de données, ses techniques et les travaux qui se sont intéressés à l'analyse des données des jeux vidéo et des JS. Le deuxième chapitre décrit les deux méthodes choisies dans notre approche, à savoir : la régression linéaire et le regroupement

(*Clustering*) et défini leurs indicateurs de qualité qui permettent d'évaluer leurs résultats. Le troisième chapitre expose les scénarios d'analyse et annonce le scénario retenu pour notre étude.

Le chapitre quatre présente les étapes de l'expérimentation et les résultats obtenus. Il décrit le jeu étudié « Science en jeu » et sa base de données (BD) et explique les phases de préparation des données. Ensuite, le contexte de l'expérimentation, les résultats, leur analyse et leur interprétation, sont détaillés pour chacune des deux méthodes (régression linéaire multiple et regroupement). Nous terminerons notre travail par une conclusion et des recommandations.

CHAPITRE 1

JEUX SÉRIEUX

1.1 Introduction

Les études portant sur les JS dans les domaines de l'éducation et de la santé témoignent de leur potentiel à devenir des outils efficaces aussi bien pour les apprenants que pour les enseignants, ainsi que pour les thérapeutes et les patients. Si leur efficacité est prouvée, ces outils permettraient de compléter les outils actuels et de développer de nouvelles approches et méthodes en éducation et en santé. Cependant, pour atteindre ces objectifs, de nombreux défis sur le plan conceptuel et technique doivent être relevés.

Dans ce chapitre, nous introduisons plusieurs définitions importantes relatives aux JS et nous présentons des études portant sur ce type de jeux et leur utilisation dans les domaines de l'éducation et de la stimulation cognitive. Ensuite, nous nous intéressons à l'adaptation des JS et au forage de leurs données, et nous terminons avec une conclusion.

1.2 Définition d'un jeu sérieux

Le terme « jeu sérieux » caractérise les jeux vidéo possédant une dimension sérieuse telle que l'enseignement, l'apprentissage, la simulation, en plus de la dimension ludique. Alvarez *et al.* ont défini le jeu sérieux comme « une application informatique, dont l'intention initiale est de combiner avec cohérence, à la fois des aspects sérieux (*serious*) tels que, [...] l'enseignement, l'apprentissage, la communication, ou encore l'information, avec des ressorts ludiques issus du jeu vidéo (*game*) » (Alvarez, 2010, p. 17). L'ancêtre des JS est le jeu Army Battlezone développé par Atari en 1980 afin d'entraîner des militaires américains (Muratet, 2010).

1.3 Classification des jeux sérieux

Parmi les premières classifications des jeux vidéo, nous retrouvons celle proposée par Crawford qui subdivise principalement les jeux en deux classes (Crawford, 1984):

- les jeux de dextérité/action tels que les jeux de combats, les jeux de labyrinthe et les jeux de courses;
- les jeux de stratégie tels que les jeux d'aventure, les jeux de rôle et les jeux éducatifs.

Une classification plus récente différencie les jeux en 42 catégories selon l'interaction du joueur avec le jeu. Nous citons à titre d'exemple les catégories suivantes (Wolf, 2001) :

- *Capturing* : jeux de capture d'objet ou de personnages,
- *Escape* : jeux de poursuite,
- *Fighting* : jeux de combat entre des personnages sans utilisation d'armes,
- *Maze* : jeux de déplacement dans un labyrinthe,
- *Strategy*: jeux se basant sur l'utilisation de stratégies.

Concernant les JS, Alvarez a proposé une classification basée sur le domaine d'application, à savoir : militaire, militant, marketing, éducatif/formatif, informatif et médical. Il a souligné aussi l'importance d'introduire le critère d'interactivité dans les classements des JS (Alvarez, 2007).

Le choix du type de jeu le mieux adapté à un domaine d'application précis devient plus difficile à cause de la complexité des jeux développés combinant parfois plusieurs types de jeux, l'existence de plusieurs classifications et les similitudes dans les définitions. Muratet a même parlé de floue et de confusion (Muratet, 2010). Il serait alors souhaitable de définir de nouveaux critères pour la distinction des jeux, et d'actualiser les classifications existantes. Cela pourrait faire l'objet de futures études de recherches qui considéreront les nouvelles approches et les récentes réalisations dans le monde des JS.

1.4 Jeux sérieux dans le contexte de l'éducation

En se basant sur une composante pédagogique, les jeux sérieux dans le contexte de l'éducation visent l'apprentissage, ou la formation. Les jeux vidéo sont des outils efficaces pour l'apprentissage selon Kebritchi et son collègue grâce à (Kebritchi, 2008):

- L'utilisation de l'action au lieu de l'explication;
- La création de la motivation et de la satisfaction;
- L'accommodement de multiples styles et aptitudes d'apprentissage;
- Le renforcement de la maîtrise des compétences;
- Le contexte interactif et de prise de décision des jeux.

Dans leur travail, Johnson *et al.* ont défini les éléments clés du contexte des JS éducatifs, à savoir : le *Gameplay* (activités et stratégies employées pour maintenir l'engagement et la motivation), la rétroaction (*Feedback*), l'interface simple et fluide, les niveaux de difficulté adaptés à l'expérience du joueur, le scénario du jeu (maintien de l'intérêt du joueur) et le divertissement (Johnson, 2005).

Les jeux éducatifs peuvent rivaliser avec les jeux vidéo de divertissement en termes d'immersion, d'engagement et de motivation, s'ils arrivent au juste équilibre entre l'apprentissage et le jeu, tout en s'appuyant sur les théories psychologiques et pédagogiques (Kickmeier-Rust, 2007). Dans ce qui suit, nous exposons des exemples de travaux de recherche qui se sont intéressés à plusieurs aspects des JS en éducation.

1.4.1 Les jeux sérieux éducatifs et l'intelligence artificielle

Dans leur étude, Johnson *et al.* ont souligné le rôle de l'intelligence artificielle (IA) dans le soutien des objectifs éducatifs et l'introduction d'enseignement adapté aux besoins sans compromettre les principes de conception du jeu. Ils ont étudié l'apport de l'IA dans le contexte d'un programme qui soutenait l'acquisition des langues étrangères et des compétences culturelles (Johnson, 2005). Il s'agissait d'un jeu 3D interactif dans lequel on

retrouvait : 1) un générateur de compétences (exercices interactifs ciblés); 2) un tuteur virtuel (évaluation du discours du joueur, aide en cas d'erreurs et encouragements) et 3) un glossaire hypertexte adaptatif (vocabulaire des leçons, et structure grammaticale des phrases apprises). À la fin de cette étude, ils ont conclu que l'IA intervenait efficacement dans le contrôle du comportement des personnages non joueurs du jeu et le support du tuteur intelligent.

Également, Mills et Dalgarno ont développé un prototype de système tutoriel intelligent basé sur un jeu 3D (apprentissage de la physique newtonienne), et implémenté à l'aide de multi agents et des technologies d'apprentissage automatique (Mills, 2007). L'étude pilote, qu'ils ont menée avec un petit groupe de participants, a révélé que ces derniers ont trouvé le jeu très intéressant, agréable avec un défi amusant. En plus, les résultats du post-test étaient très encourageants. En conclusion, ils ont recommandé une autre évaluation avec un plus grand échantillon et un groupe de contrôle pour valider les résultats obtenus.

Après l'étude des principales contributions de l'IA et de la personnalisation dans le domaine des JS, Brisson *et al.* ont conclu que ces deux aspects sont cruciaux pour le développement de ce type de jeu. Seulement, leur conception et leur implémentation efficace présentent des défis majeurs. Selon ces chercheurs, la simplification du processus de création et d'adaptation permettrait aux personnes sans connaissance en programmation de développer leurs propres jeux (Brisson, 2012). Aussi, le développement d'outils auteurs réunissant les fonctionnalités de création de jeux et le développement de leurs composants « intelligents » est déterminant pour l'essor des JS éducatifs.

1.4.2 Les jeux sérieux éducatifs et la motivation des apprenants

Maintenir l'attention et la motivation des joueurs demeure parmi les principaux objectifs des JS. Plusieurs recherches montrent que la motivation et l'apprentissage peuvent s'accroître à travers plusieurs éléments d'une activité incluant la rétroaction (*Feedback*), la fantaisie, la personnalisation, le choix et la curiosité (Jackson, 2011).

Dans cette perspective, Jequier et son collègue ont étudié le jeu en ligne Science Express comme étant un moyen de motivation. Le jeu consistait à résoudre des énigmes portant sur les sciences fondamentales (Mathématiques, physique, chimie, etc.) tout en effectuant un voyage exploratoire (Jequier, 2012). L'aide proposée au joueur était sous forme d'indices cachés donnant accès à des ressources (lecture, vidéo). Après l'expérimentation du jeu auprès des étudiants (université Bordeaux 1) pendant le semestre de remise à niveau, l'analyse de leurs réponses (discussions et questionnaire) a révélé des résultats positifs sur les plans de la motivation, la cohésion du groupe, la prise de conscience du potentiel ludique des jeux et des ressources pédagogiques offertes. Les auteurs de l'étude ont suggéré l'utilisation du jeu dès la première année. Néanmoins, des améliorations apportées au : 1) design du jeu (graphisme de bande dessinée) et son adaptation à la clientèle; et 2) système d'analyse des réponses textuelles et des délais occasionnés par le facteur humain; pourront favoriser une utilisation efficace et une meilleure acceptation par les apprenants.

Pour développer des stratégies de lecture auprès d'un groupe d'étudiants (9 étudiants), Jackson et ses coéquipiers ont expérimenté iSTART-ME (système tutoriel intelligent basé sur le jeu) pendant 8 semaines (Jackson, 2011). Ils ont constaté, d'après les résultats du questionnaire (post-test), que les participants préféraient les aspects du système basé sur le jeu (iSTART-ME) nettement plus que celui sans le jeu (iSTART). Ces résultats renforcent la pertinence de l'utilisation d'un tel système pour favoriser le divertissement et potentiellement augmenter l'acquisition de compétences.

1.4.3 Les jeux sérieux éducatifs et l'apprentissage collaboratif : Jeu 2.0

Le jeu 2.0 est défini comme étant « toute application permettant à un utilisateur de créer, d'échanger et de jouer à un contenu ludique » (Djaouti, 2011, p. 201). Ses technologies offrent des moyens intéressants pour la conception des jeux, mais il demeurerait un manque sur le plan des JS et leur potentiel éducatif selon une étude comparative réalisée par Djaouti et ses collègues (Djaouti, 2010).

D'autre part, Monterrat et son équipe ont expérimenté le *modding*¹ sur les jeux éducatifs pour en faire un moyen d'appropriation des connaissances contenues dans les jeux 2.0 (approche constructiviste de l'apprentissage) (Monterrat, 2012). Ils ont alors proposé un modèle d'activité éducative et une architecture générique du système informatique supportant cette activité (Figure 1.1) et se basant principalement sur la triple relation entre la structure du jeu modifié, la carte des connaissances (ontologie) et les discussions. Les résultats du test du prototype réalisé avec 16 participants ont montré une tendance à l'adoption de l'apprentissage avec les outils collaboratifs grâce à plusieurs facteurs déterminants, à savoir : la motivation (aspect ludique du jeu), l'interaction entre les apprenants et l'apprentissage collaboratif. Cependant, certaines améliorations ont été proposées par l'équipe des chercheurs tels que l'expérimentation à plus grande échelle et sur une plus longue durée, la consolidation des interactions entre participants (outils de discussion) et la création d'outils pour les enseignants (élaboration de carte de connaissance et évaluation) ainsi que l'intégration des composants de l'activité sur une plateforme unique. Aussi, le *modding* a été expérimenté dans d'autres études pour l'apprentissage de la programmation, où des outils comme Crytek, Warcraft III, Web Driver et Unreal Engine 2.5 ont permis à des étudiants d'apprendre et de mettre en application les bases de l'algorithmique, des langages de programmation et le processus de développement logiciel (McAtamney, 2005) (El-Nasr, 2006).

¹ *Modding* : le *modding* (modification) dans le domaine des jeux vidéo consiste à créer un jeu vidéo à partir d'un autre, ou en modifiant un jeu original (Technologysource, 2003).

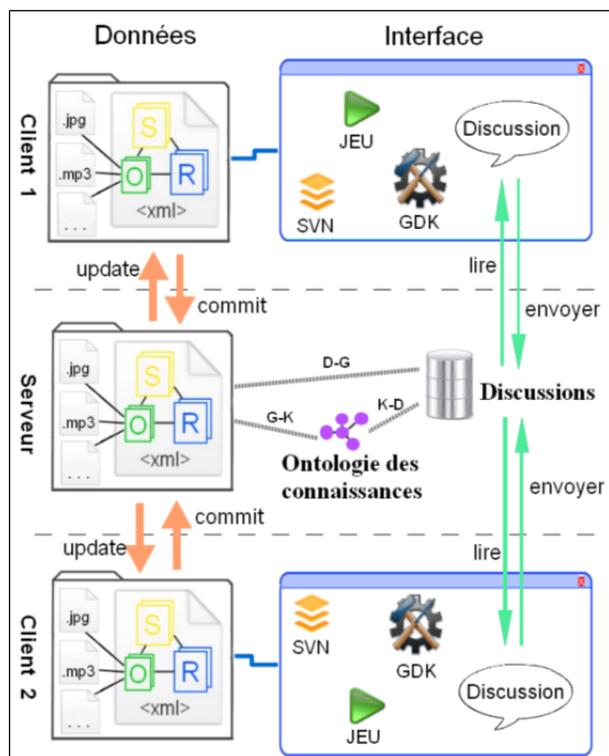


Figure 1.1 Architecture d'un jeu éducatif (*learning game*) 2.0
Tirée de Monterrat (2012, p. 103)

Cette approche est prometteuse grâce à l'intérêt qu'elle susciterait auprès des apprenants familiers avec les notions du Web 2.0 et du jeu 2.0. Aussi, la dimension collaborative et/ou compétitive pourrait maintenir la motivation et renforcer le processus d'apprentissage. Toutefois, il reste à résoudre les problématiques liées à la disponibilité des outils intégrant toutes les fonctionnalités de développement, la complexité de l'évaluation de l'apprentissage dans un contexte collaboratif et les manières d'intégration de ces jeux éducatifs dans les contextes actuels d'enseignement.

1.5 Jeux sérieux pour la stimulation cognitive

Les résultats de plusieurs études sur la stimulation cognitive à l'aide de jeux vidéo ont montré une amélioration ou un renforcement de certaines fonctions cognitives telles que : l'attention visuelle (Green, 2003), les fonctions exécutives chez les personnes âgées (Basak, 2008) et la mémoire spatiale et verbale chez un patient atteint de lésion cérébrale causée par

un accident (Caglio, 2009). Ces résultats encourageants ont ouvert la voie au développement de nouveaux outils pour la stimulation cognitive basés sur les jeux vidéo. Il s'agit des JS qui ont été expérimentés dans différents contextes. Nous présentons dans cette section quelques exemples pour documenter cette tendance.

Bellack et ses collègues ont réalisé une étude pilote sur la remédiation cognitive chez les patients atteints de schizophrénie en utilisant des exercices sur ordinateur pour travailler la mémoire, la verbalisation, la résolution de problème, la planification des tâches, etc. Les chercheurs ont remarqué que les sessions d'entraînement étaient bien reçues par les patients et que ces derniers étaient plus motivés pour s'exercer (Bellack, 2005). Une autre recherche a expérimenté un jeu vidéo employant un récit d'enquête policière afin de renforcer l'engagement des adolescents ayant des problèmes de santé mentale (dépression, anxiété, trouble du comportement) aux psychothérapies (Coyle, 2005). Avec l'aide d'un thérapeute, le patient apprenait de nouvelles stratégies pour résoudre ses propres problèmes tout en jouant au détective. Les résultats de l'étude pilote et l'étude de cas de cette recherche ont révélé les bénéfices d'une telle approche dans l'amélioration de l'engagement et la motivation des participants ainsi qu'un développement plus rapide des relations avec les thérapeutes.

Également, Wartena et son coéquipier ont conçu et implémenté un JS comme outil d'aide dans le traitement de l'autostigmatisation² chez des patients en rémission d'un premier épisode psychotique (Wartena, 2013). Leur approche est caractérisée par deux aspects importants : 1) l'introduction d'un outil technologique (JS) en psychothérapie et 2) l'introduction d'une activité thérapeutique dans la conception de jeux. Selon les auteurs de l'étude, cette approche permettra un apprentissage autorégulé par le patient et l'augmentation de sa motivation et de son engagement dans les thérapies, elle fournira aussi un moyen intégré et efficace de suivi et d'évaluation de la progression des patients. Ils ont proposé la

² Autostigmatisation : elle «se produit lorsque les personnes atteintes de maladies mentales et leurs familles intériorisent les attitudes négatives de la société à leur égard, ce qui les amène à se blâmer et à avoir une faible estime de soi.» (TermiumPlus, 2013).

validation de leur travail à travers des tests contrôlés et l'exploration de leur JS pour d'autres troubles de santé mentale.

En étudiant les jeux sérieux pour la réhabilitation, Rego *et al.* ont réalisé une taxonomie de ces jeux en se basant sur les critères suivants : le champ d'application, la technologie d'interaction, l'interface, le nombre de joueurs, le type de jeu, l'adaptabilité, la rétroaction sur la performance du joueur, le contrôle de suivi et la portabilité du jeu. Aussi, ils ont effectué l'étude de RehaCom comme système de référence (Rego, 2010). C'est un système modulaire composé de procédures pour entraîner l'attention, la mémoire, les fonctions exécutives, le champ de vision et visuomoteur³ et nécessitant un thérapeute expérimenté. Les auteurs de la recherche ont conclu que RehaCom offrait de nombreux avantages. Mais, certains aspects restaient à améliorer tels que son interface et la dimension coopérative et compétitive en permettant l'interaction entre plusieurs joueurs appartenant à une même équipe ou à des équipes adverses. Parmi leurs propositions, nous notons : 1) l'identification et l'évaluation de l'impact des aspects pouvant améliorer la pertinence et l'efficacité des jeux pour la réhabilitation; 2) l'incorporation de la dimension sociale (collaboration et compétition) et son lien avec l'accroissement de l'efficacité de ces jeux; et 3) le test des JS sur un plus grand nombre de patients.

Principalement, ces études montrent que le côté divertissant et ludique des JS est un facteur important pour accroître la motivation des patients et maintenir leur engagement et leur persévérance pendant les séances d'entraînement qui étaient perçues auparavant comme une corvée pénible et ennuyante. Elles nous renseignent aussi sur le besoin des thérapeutes de nouveaux outils d'aide attrayants, efficaces et adaptés à l'âge et aux domaines d'intérêts des patients.

³Visuomoteur : relatif à la vue et à l'ensemble des fonctions sensorimotrices (TermiumPlus, 2014).

1.6 Adaptation des jeux sérieux

Le succès des JS dépend de deux facteurs essentiels à savoir la convivialité et l'acceptation. Plusieurs approches sont proposées pour garantir ces deux facteurs et améliorer l'efficacité de la formation/entraînement et la satisfaction des utilisateurs. Parmi ces approches, nous retrouvons l'utilisation des principes généraux de conception du jeu (*Game Design*) et des interfaces homme-machine appropriées et attractives ainsi que l'adaptation dynamique (Hocine, 2011). Dans cette section, nous définissons l'adaptation des jeux, et nous discutons des résultats de certaines recherches (à titre d'exemple) qui se sont intéressées à ce sujet.

1.6.1 Définition de l'adaptation

Selon Hocine *et al.*, l'adaptation est « [...] un processus qui va interagir avec le système du jeu en modifiant structurellement sa présentation, ses modèles de données ou sa logique de contrôle.» (Hocine, 2011, p. 257).

Ismailović *et al.* ont proposé la répartition du processus d'adaptabilité en quatre étapes (Ismailović, 2012) :

- L'observation des actions du joueur pendant les sessions de jeu et la collecte de ses résultats;
- La caractérisation du joueur et la détermination de son profil;
- La génération de l'évaluation du niveau du joueur basée sur son profil et les observations effectuées;
- L'intervention adaptative par l'ajustement automatique du contenu d'apprentissage et des éléments du jeu selon des théories spécifiques d'apprentissage.

1.6.2 Exemples d'études sur l'adaptation des jeux sérieux

Hocine *et al.* ont analysé plusieurs études traitant de l'adaptation des JS et ludiques, en utilisant un cadre d'évaluation qui se composait des critères suivants (Hocine, 2011) :

- Le périmètre de l'adaptation (fonctionnalités concernées par l'adaptation : contrôle, présentation ou contenu),
- Les paramètres de l'adaptation,
- Le modèle d'adaptation (représentation implicite ou explicite des règles d'adaptation),
- La nature du système (monojoueur ou multijoueur).

Suite à cette analyse, ils ont trouvé, à propos des JS étudiés, que :

- Le périmètre de l'adaptation englobait le contrôle ainsi que les aspects de présentation et du contenu tels que l'adaptation des aides (Conati, 2009), l'adaptation des interfaces dans les jeux thérapeutiques et éducatifs, l'adaptation des scénarios narratifs (Delmas, 2007) et des rétroactions visuelles ou sonores (Chen, 2006);
- les paramètres de l'adaptation étaient principalement issus de l'utilisateur. Selon le type du jeu, le modèle utilisateur peut contenir des informations sur l'état attentionnel de l'utilisateur, son état émotionnel, ses préférences, sa localisation physique, ses taux de réussite, ses capacités, les rétroactions, etc.;
- Dans la plupart des travaux, le modèle d'adaptation n'était pas identifié explicitement et donc faisait partie du code source de l'application. Cependant, quelques travaux utilisaient un moteur d'inférence (Peirce, 2008) et un module d'exécution indépendant de jeu;
- la nature du système était essentiellement monojoueur.

Pour devenir des outils de formation adaptés, les JS doivent permettre, selon les chercheurs Sottolare et Gilbert: 1) l'ajout des mesures physiologiques et comportementales pour soutenir la modélisation cognitive et 2) l'intégration des jeux sérieux et des systèmes tutoriels. Pour le deuxième défi, ils ont proposé l'introduction d'interfaces entre le moteur du jeu et celui du système tutoriel. En évaluant les données du jeu, les mesures physiologiques et les données de l'historique, le système tutoriel fournit les stratégies pédagogiques au moteur de jeu qui ajuste les rétroactions et les scénarios du jeu en conséquence (Figure 1.2) (Sottolare, 2011).

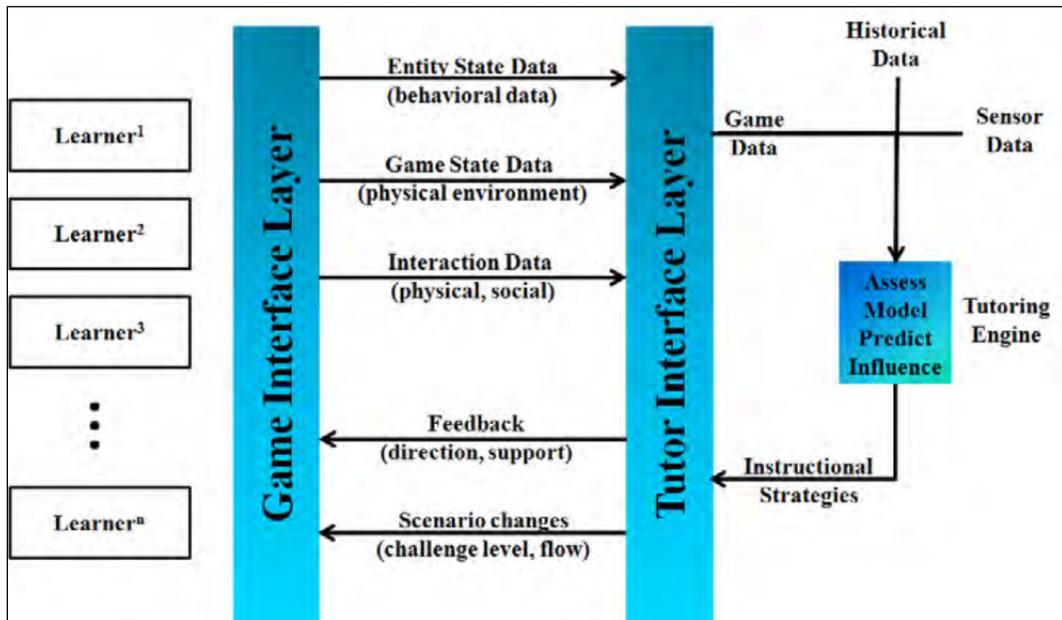


Figure 1.2 Couche d'abstraction jeu-système tutoriel
Tirée de Sottolare (2011, p.15)

Pour déterminer le profil du joueur, certaines approches utilisent l'historique de ses activités et de ses interactions ainsi que ses informations générales, ses préférences et son domaine de connaissances (Sehaba, 2005). D'autres approches utilisent les techniques de forage de données et la technique de la similarité des événements (Salmeron, 2011).

L'adaptation des JS fait face à plusieurs défis concernant la conceptualisation de l'apprenant ou du patient (capacité d'apprentissage et progression), la détection de son engagement en temps réel et le maintien de l'équilibre entre le contenu ludique et pédagogique du jeu ainsi que la prise en considération du profil du joueur, des objectifs d'apprentissage, et des personnages non-joueurs de l'environnement lors de cette adaptation (Brisson, 2012). En plus, l'avènement des jeux 2.0 dans le domaine des JS pose des défis supplémentaires à cause de la complexité de l'adaptation d'un jeu collaboratif.

1.7 Jeux sérieux et évaluation

Afin de faciliter le choix des JS et leur adaptation pour diverses situations d'apprentissage, il est important d'avoir des moyens pratiques et fiables pour leur évaluation. Or, selon Boughzala et ses collègues, peu de travaux se sont intéressés à ce sujet. Pour y remédier, ils ont proposé une grille de caractérisation et d'évaluation (*G-CE-JS: Grid for Characterizing and Evaluation Serious Game*) (Boughzala, 2013). Il s'agit d'une liste de critères subdivisés en trois sections : identification, description et évaluation. Les auteurs de l'étude ont testé la grille *G-CE-JS* avec 24 étudiants familiers avec les jeux vidéo qui ont proposé la clarification de certains termes, la revue du système de notation et l'ajout de plus de choix de réponse. Parmi ses limitations, cette grille ne peut être utilisée pour examiner la corrélation entre la qualité du JS et l'objectif éducatif.

Pour pallier la problématique reliée à la création complexe et coûteuse en moyen et en temps, Marfisi-Schottman *et al.* ont proposé un outil d'évaluation des JS en cours de conception. Cet outil est composé d'indicateurs de qualité structurés en 6 facettes : objectifs d'apprentissage, problèmes et progression, conditions d'utilisation, interactions, décorum (aspect extérieur) et coût prévisionnel. Sa validation a été réalisée par 6 experts (évaluation de 24 jeux). Les chercheurs ont prévu d'améliorer leur proposition en remplaçant certains indicateurs par des éléments chiffrables issus des traces du jeu pour aider les experts et réduire le temps d'évaluation (Marfisi-Schottman, 2012).

L'évolution des JS dans différents domaines d'application exige une évaluation fiable. Celle-ci peut se faire avec l'aide des experts en développement de jeux vidéo et de jeux sérieux (grilles d'indicateurs de qualité), et en recueillant les avis des utilisateurs sur leur degré de satisfaction (questionnaires d'appréciation) ainsi qu'en examinant les résultats des utilisateurs à des tests spécifiques visant à vérifier l'acquisition des connaissances/compétences (examens, tests neuropsychologiques, etc.). L'analyse de ces données permet d'avoir une vue plus globale du jeu, de son acceptabilité et de sa pertinence dans un contexte précis en diminuant le plus possible le biais causé par les questionnaires

d'appréciation (réponses subjectives). Une telle analyse pourrait être réalisée grâce aux techniques de forage de données.

1.8 Jeux sérieux et forage de données

Le forage de données est actuellement utilisé dans de nombreux domaines. Il combine les techniques de statistique, d'analyse de données et d'intelligence artificielle et offre ainsi un moyen puissant d'exploration et d'analyse de grande base de données afin d'en extraire de nouvelles connaissances.

Nous présentons dans cette section la définition du forage de données, les étapes de son processus et la classification de ses techniques. Nous effectuons également un survol des études portant sur le forage de données des jeux vidéo et des JS.

1.8.1 Définition du forage de données (*Data Mining*)

Tufféry définit le Forage de données (ou fouille de données) comme étant :

l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de (souvent grandes) bases de données informatiques, de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données. (Tufféry, 2010, p. 4)

Autrement dit, le forage de données est l'ensemble de techniques permettant la découverte et la description de patrons (*pattern*) dans un grand volume de données. Il aide à trouver des explications pour ces données et à faire des prédictions (Witten, 2011).

1.8.2 Étapes de forage de données

Le processus de forage de données se déroule en trois étapes principales (Ratté-1, 2013):

- Le prétraitement : Pendant cette étape, le travail se concentre sur la compréhension du domaine, la compréhension des données et leur préparation;
- La modélisation : Elle consiste à sélectionner les modèles appropriés et les tester sur les données. Par la suite, une évaluation des résultats est réalisée;
- Le post-traitement : C'est l'étape d'analyse des résultats et de leur visualisation.

1.8.3 Classification des techniques de forage de données

Les techniques de forage de données se répartissent en deux grandes classes (Tufféry, 2010) :

- Les techniques descriptives : Elles permettent de mettre en évidence les informations cachées dans les données. Nous retrouvons dans cette classe les méthodes de partitionnement, l'analyse en composantes principales, la recherche d'association, etc.;
- Les techniques prédictives : Elles permettent de faire l'extrapolation de nouvelles informations à partir des informations présentes. Parmi ces techniques, il y a les arbres de décision, la régression linéaire, l'analyse discriminante, les réseaux de neurones, etc.

1.8.4 Exemples d'études sur le forage de données des jeux sérieux

Les jeux vidéo et les JS font l'objet de plus en plus d'études en vue d'en faire de véritables outils d'aide à l'apprentissage, la stimulation cognitive, la réhabilitation et même au diagnostic. L'évaluation de l'efficacité et la pertinence de tels outils exigent l'analyse des résultats des différentes expériences (données du jeu, appréciation du joueur et son comportement). Cela constitue un volume important de données que les techniques de forage de données peuvent traiter efficacement. Dans cette section, nous décrivons quelques études qui ont utilisé le forage de données pour le traitement des données des jeux.

Afin d'évaluer la capacité prédictive d'un jeu vidéo (*Supermarket Game*) à détecter les cas de trouble du déficit de l'attention avec ou sans hyperactivité (TDAH), Santos et son équipe ont analysé les données du jeu (âge, sexe, score, durée et classification-TDAH) d'un groupe de 80 enfants (âgés de 10 à 17 ans) à l'aide de deux techniques de forage de données, à savoir :

Naive Bayes et les arbres de décision. Les résultats obtenus ont permis de conclure que le jeu était capable de classer des enfants comme ayant le TDAH ou non, à travers les deux techniques employées avec de bonnes mesures de performance (spécificité, sensibilité autour de 0,7). Seulement, il n'était pas possible de préciser plus la classification avec les types du TDAH à cause du peu de cas dans le corpus d'apprentissage (Santos, 2011). De cette étude, nous retenons quatre éléments relatifs à notre approche : 1) le jeu vidéo comme outil d'aide à la prise de décision ou le diagnostic; 2) le forage des données du joueur (âge et sexe), son score et la durée du jeu et 3) la constitution du sous-ensemble d'attributs pour l'étude en sélectionnant les plus corrélés d'entre eux à la classification.

L'effet d'un jeu sérieux (*Food Force*) sur la motivation et l'apprentissage chez le joueur a été examiné dans l'étude de Derbali et Frasson. Dans le but de prédire l'état motivationnel du joueur, les chercheurs ont utilisé le modèle de régression linéaire multiple sur des mesures électrophysiologiques (électroencéphalographie EEG, conduction cutanée et fréquence cardiaque) ainsi que des évaluations prétest et post-test (questionnaires) sur la motivation et les connaissances apprises (Derbali, 2010). L'analyse des résultats a permis de : 1) déterminer les variables explicatives significatives (exemple : conduction cutanée) pour la motivation du joueur; et 2) de constater l'accroissement de la motivation (attention et confiance) et de l'apprentissage.

Également, Shim *et al.* ont construit des modèles afin de prédire la performance future du joueur dans un contexte de jeu vidéo en ligne massivement multijoueurs (Shim, 2011). En premier lieu, ils ont exploré les corrélations entre les données, ensuite ils ont appliqué le modèle de régression linéaire simple et multiple sur les données corrélées. Leurs résultats indiquaient que les modèles utilisant la performance passée et la diversité sociale (activités de mentorat entre apprenti joueur et mentor) comme variables explicatives donnent de meilleurs résultats, et la prévision des mentors est généralement meilleure que celle des apprentis.

Une autre approche effectuée par Anagnostou et son collègue consiste à proposer la modélisation des joueurs en se basant sur le regroupement de leurs données (sexe, joueur familial avec les jeux vidéo ou pas, durée, précision des tirs, armes utilisées, nombre d'ennemis détruits, résultat final, etc.) collectées durant un jeu vidéo d'action (Anagnostou, 2009). Ils ont défini deux types de joueurs : le joueur d'action et le joueur tactique et ils ont utilisé l'algorithme CURE (*Clustering Using REpresentatives*) pour déterminer les regroupements de joueurs et par la suite, ils les ont comparés avec la classification faite manuellement selon le style de jeu. Les résultats obtenus étaient satisfaisants. Ainsi, la connaissance du type du joueur pourrait être utilisée pour l'ajustement des niveaux de difficulté du jeu et offrir une meilleure expérience lors des sessions de jeu.

1.9 Conclusion

Dans ce chapitre, nous avons défini les principales notions relatives à notre étude et nous avons présenté des études de recherche qui se sont intéressées aux JS dans le domaine de l'éducation et de la stimulation cognitive. Ces études nous révèlent la pertinence et la faisabilité de cette voie de recherche grâce aux résultats encourageants obtenus. Elles ont démontré des effets positifs sur la motivation et l'apprentissage du joueur et proposent des améliorations à travers par exemple : la personnalisation des JS et l'introduction des aspects collaboratifs et compétitifs. Aussi, une évaluation efficace et fiable est l'un des principaux paramètres dont dépend l'amélioration des JS. Celle-ci requiert des outils d'évaluation fiables et proposant des critères mesurables et quantifiables.

L'important volume de données issues des JS, leurs complexités et leurs étendues représentent un réel défi pour l'analyse et l'interprétation et surtout l'extraction de nouvelles connaissances. Il est alors primordial de sélectionner les données les plus pertinentes et les techniques de forage les plus appropriées. Ainsi, l'analyse des caractéristiques du joueur (âge, sexe, etc.) et des données de son parcours dans le jeu (score, durée, mesures électrophysiologiques, etc.) a permis dans des études précédentes de définir les profils des

joueurs et réaliser leur classification. De tels résultats devront être renforcés par plus d'expérimentations et à plus long terme pour établir des modèles prédictifs fiables et exploitables pour l'adaptation des JS dans les contextes d'éducation et de stimulation cognitive. Cette adaptation repose sur la caractérisation du joueur (apprenant/patient) et son profil, et elle vise essentiellement la satisfaction de ses besoins en compétences et en divertissement.

Selon Anagnostou et son collègue, les techniques de regroupement appliquées aux données des jeux vidéo pourraient contribuer à l'identification des paramètres qui concernent le comportement du joueur et les aspects de sa modélisation (Anagnostou, 2009). Mais au début de notre étude, nous ne savons pas quelles sont les données les plus importantes (influentes) pour appliquer une méthode de partitionnement et déduire des regroupements de joueurs. Il s'agit donc, d'abord, d'une approche exploratoire qui nous amènera à rechercher les données influentes à l'aide de modèles à base de fonction mathématiques (corrélation, régression linéaire) et les exploiter ensuite pour le regroupement (méthode descriptive). Par conséquent, nous détaillons la régression linéaire et le regroupement dans le chapitre suivant avant de les appliquer sur nos données.

CHAPITRE 2

MÉTHODES DE FORAGE DE DONNÉES : RÉGRESSION LINÉAIRE ET REGROUPEMENT

2.1 Introduction

Conformément aux objectifs de notre étude, à savoir : 1) l'identification des données pertinentes du jeu et du joueur, et des méthodes de forage de données appropriées pour les étudier et 2) la déduction des caractéristiques des profils des joueurs, ainsi qu'à notre corpus de données (volume, dimension et types de données), nous avons choisi la régression linéaire multiple (RLM) et le regroupement comme techniques pour effectuer le forage des données du jeu. La première technique permet de déterminer éventuellement l'existence de relations entre les données du joueur. Tandis que la deuxième technique permet d'explorer l'existence de regroupements selon des caractéristiques et des performances particulières des joueurs.

Dans ce chapitre, nous définissons la régression linéaire. Ensuite, nous nous intéressons à la méthode de regroupement que nous décrivons plus en détail. Nous décrivons aussi les indicateurs de qualité pour chacune des méthodes. Enfin, nous terminons avec une conclusion.

2.2 Modèles de régression linéaire

Les modèles de régression linéaire ont pour but d'expliquer ou prédire la variance d'une variable dépendante (à expliquer) à l'aide d'une ou plusieurs variables indépendantes (explicatives). Suivant le contexte et la finalité du modèle ainsi que les données traitées, nous retrouvons globalement deux types de régression : 1) la régression linéaire (simple et multiple) qui traitent les variables dépendantes continues ou discrètes et les variables indépendantes continues ou catégorielles; et 2) la régression logistique qui permet de traiter les variables dépendantes de type dichotomique (0 ou 1) et les variables indépendantes continues ou catégorielles.

Dans cette section, nous présentons les techniques de régression linéaire simple et multiple ainsi que leurs indicateurs de qualité. Les paramètres utilisés dans les définitions de ces techniques sont identifiés et expliqués dans le tableau ci-dessous (Tableau 2.1)

Tableau 2.1 Liste des paramètres utilisés dans les définitions de la régression linéaire

Variable	Définition	Formule de calcul
n	Nombre d'observations de l'échantillon	-
m	Nombre de variables indépendantes X_j	-
$E(Y X=x)$	Espérance conditionnelle de Y sachant X	-
ε_i	Erreur pour l' $i^{\text{ème}}$ observation	-
\bar{x}	Moyenne de X	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
\bar{y}	Moyenne de Y	$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$
\hat{Y}	Valeur estimée de Y	
$Cov(X, Y)$	Covariance	$Cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$
$N(0, s^2)$	Densité normale de moyenne nulle et d'écart type s	-
s_x^2	Variance de X	$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
s^2	Variance de l'erreur	-
X^t	Matrice transposée de X	-
X^{-1}	Matrice inverse de X	-
X_j	$J^{\text{ème}}$ variable indépendante	-
$E(X)$	Espérance de X	-

2.2.1 Régression linéaire simple

La régression linéaire simple permet de « [...] mettre en relation une variable continue Y à expliquer avec une variable explicative X continue » (Tufféry, 2010, p.352). On suppose que la connaissance de la valeur de X ($X=x$) permet de connaître l'espérance conditionnelle de Y sachant X : $E(Y|X=x)$. Cette relation est exprimée par (Tufféry, 2010):

Pour tout $i = 1, \dots, n$

$$E(Y_i) = \alpha + \beta X_i \quad (2.1)$$

C'est équivalent à :

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad (2.2)$$

avec l'espérance de l'erreur ε_i nulle ($E(\varepsilon_i) = 0$) pour tout $i = 1, \dots, n$

On suppose aussi que :

- Les variables X et Y ne sont pas indépendantes;
- La variance des erreurs est la même pour toutes les valeurs de X (Variance de $\varepsilon_i = s^2$);
- Les erreurs sont linéairement indépendantes ($Cov(\varepsilon_i, \varepsilon_j) = 0$ pour tous $i \neq j$);
- Les erreurs sont normalement distribuées ($\varepsilon_i \sim N(0, s^2)$).

On cherche a et b , les estimateurs de α et β respectivement, à partir d'un échantillon en utilisant plusieurs méthodes. Parmi celles-ci, la méthode de minimisation du critère des moindres carrés ordinaires qui consiste à trouver les estimateurs a et b qui minimisent :

$$\sum_{i=1}^n (y_i - a - bx_i)^2 \quad (2.3)$$

Les moindres carrés sont minimisés par :

$$b = \frac{Cov(X,Y)}{s_x^2} \quad (2.4)$$

et

$$a = \bar{y} - b\bar{x} \quad (2.5)$$

On obtiendra pour tout X_i , la valeur prédite ou estimée de Y_i égale à $a + b X_i$. Les erreurs ε_i sont estimées par les résidus égaux à $Y_i - (a + b X_i)$

2.2.2 Régression linéaire multiple

La régression linéaire multiple est la généralisation de la régression linéaire simple. Elle met en relation une variable continue Y à expliquer avec plusieurs variables explicatives X_j . Elle est définie par l'expression suivante (Tufféry, 2010) :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m + \varepsilon \quad (2.6)$$

avec l'espérance de l'erreur ε égale à zéro : $E(\varepsilon) = 0$.

On suppose l'indépendance linéaire des variables explicatives X_j pour tout $j = 1, \dots, m$.

Pour n observations, nous obtenons $Y = X\beta + \varepsilon$ exprimé sous la forme matricielle suivante :

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1m} \\ \vdots & X_{21} & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nm} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (2.7)$$

Le traitement de la régression linéaire multiple consiste à rechercher l'estimateur $b = \begin{pmatrix} b_0 \\ \vdots \\ b_m \end{pmatrix}$

du vecteur β en utilisant par exemple la méthode des moindres carrés ordinaires. En appliquant cette méthode, l'expression à minimiser s'écrit :

$$\sum_{i=1}^n \left[y_i - b_0 - \sum_{j=1}^m x_{ij} b_j \right]^2 \quad (2.8)$$

En utilisant la notation matricielle avec (2.8), nous obtenons :

$$\|Y - Xb\|^2 = (Y - Xb)^t (Y - Xb) = Y^t Y - 2b^t X^t Y + b^t X^t X b \quad (2.9)$$

La dérivation matricielle de (2.9) sur b nous donne :

$$X^t X b - X^t Y = 0$$

Avec l'hypothèse que la matrice $(X^t X)$ est inversible, nous obtenons l'estimateur b :

$$b = (X^t X)^{-1} X^t Y \quad (2.10)$$

et les valeurs estimées (ajustées) de Y :

$$\hat{Y} = Xb = X (X^t X)^{-1} X^t Y \quad (2.11)$$

Le vecteur des erreurs (résidus) permet de connaître l'écart entre les valeurs observées (réelles) de Y et celles qui sont estimées \hat{Y} , il nous renseigne donc sur la précision du modèle de régression obtenu. Il est calculé avec :

$$\varepsilon = Y - \hat{Y} = Y - Xb \quad (2.12)$$

2.2.3 Indicateurs de qualité des modèles de régression linéaire

Plusieurs indicateurs sur la qualité des modèles de régression linéaire sont utilisés afin d'évaluer leurs résultats et choisir les plus performants pour la prédiction ou la classification. Nous commençons par définir les sommes des carrés, puis nous définissons ces indicateurs.

- Sommes des carrés (Tufféry, 2010) :

- Sommes des carrés de la régression (SCR) :

$$SCR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (2.13)$$

- Somme des carrés résiduels (erreurs) (SCE) :

$$SCE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.14)$$

- Somme des carrés totale (SCT) : $SCT = SCR + SCE$

$$SCT = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (2.15)$$

Sachant que y_i est la valeur de Y pour la $i^{\text{ème}}$ observation et \hat{y}_i est la valeur estimée de y_i par le modèle de régression.

- P-valeur (p -value) ou degré de signification : Elle est définie par « la probabilité d'obtenir une statistique de test aussi extrême (\geq ou \leq) que la valeur mesurée sur l'échantillon si H_0 est vrai » (Tufféry, 2010, p. 653). Pour rejeter l'hypothèse nulle H_0 ($H_0 : \beta_1 = \beta_2 = \dots = \beta_m = 0$) et considérer le résultat du test significatif, la p -valeur doit être inférieure à la valeur prédéterminée du seuil α (généralement $\alpha = 0,05$ dans le domaine médical).
- F-statistique (F -stat) (Tufféry, 2010) : F-statistique donne la significativité globale du modèle, elle suit une loi de Fisher à $(m, n-m-1)$ degrés de liberté. L'hypothèse nulle H_0 ($H_0 : \beta_1 = \beta_2 = \dots = \beta_m = 0$) est rejetée si la p -valeur associé à F -stat est inférieure à α ($\alpha = 0,05$). F -stat est calculée par :

$$F\text{-stat} = \frac{\frac{SCR}{m}}{\frac{SCE}{n-m-1}} \quad (2.16)$$

- Coefficient de détermination (R^2) : Il est défini par (Tufféry, 2010) :

$$R^2 = \frac{SCR}{SCT} \quad (2.17)$$

L'ajustement du modèle est d'autant meilleur que R^2 est proche de 1. Autrement dit, si SCR est proche de SCT ($\frac{SCR}{SCT} \sim 1$), la somme des carrés résiduels (erreurs) est petite. Seulement R^2 « [...] croît « mécaniquement » avec le nombre de variables, dans une proportion plus grande que l'ajustement réel du modèle linéaire [...] au nuage de points » (Tufféry, 2010, p.364) et par conséquent cet indicateur de qualité devra être interprété avec précaution et en comparant des modèles avec le même nombre de variables explicatives. De plus, la quantité R est appelée le coefficient de corrélation multiple entre Y et les variables explicatives X_i , et la quantité $R^2 \cdot 100$ représente le pourcentage de la variabilité de Y expliqué par le modèle.

- Tests sur les résidus consistent à vérifier les hypothèses de :
 - Normalité de la distribution des résidus : ($\varepsilon_i \sim N(0, s^2)$),
 - Égalité de la variance ou homoscélasticité (Variance de $\varepsilon_i = s^2$),
 - Absence d'autocorrélation : ($Cov(\varepsilon_i, \varepsilon_j) = 0$ pour tous $i \neq j$).

2.3 Regroupement (*Clustering*)

Dans cette section, nous définissons le regroupement des données (*Clustering*) et nous détaillons la méthode de partitionnement K-moyennes (Kmeans). Ensuite, nous terminons avec la présentation des indicateurs de qualité.

2.3.1 Regroupement : définition et méthodes

Le regroupement (*Clustering*) est une méthode de classification non supervisée. Elle consiste à rassembler des données en un nombre limité de groupes (*clusters*) naturels. Son objectif est

l'obtention d'une représentation concise du comportement d'un système à partir d'un grand ensemble de données (MathWorks-1, 2014).

Selon Tufféry, le regroupement est une méthode descriptive et non prédictive. Il consiste à regrouper des objets en un nombre limité de groupes ou classes, qui ont deux propriétés: 1) les groupes (ou *clusters*) ne sont pas prédéfinis par l'analyse, mais découverts au cours du traitement; 2) les groupes rassemblent les individus de la population ayant des caractéristiques similaires (homogénéité interne) et séparent ceux ayant des caractéristiques différentes (hétérogénéité externe) (Tufféry, 2010). Afin de former des groupes homogènes, des critères de proximité ou de similarité sont définis en introduisant des mesures de distance.

Dépendamment des types des attributs de l'ensemble des données (*dataset*) et de la structure des groupes à obtenir, le choix des algorithmes de regroupement est porté principalement sur (Ratté-2, 2013):

- des méthodes de partitionnement basées sur la construction de diverses partitions sur les données et dont le nombre de groupes est défini à priori (Exemples : K-moyennes, K-médoïdes);
- des méthodes hiérarchiques (ascendantes agglomérantes ou descendantes fractionnelles) basées sur la décomposition hiérarchique des données (Exemples : Farthest First, AGNES);
- des méthodes basées sur la notion de densité : les groupes sont composés de régions denses de points et séparés entre eux par des régions de faible densité (Exemples : DBSCAN, DENCLUE);
- des méthodes basées sur la notion de modèle : un modèle (hypothèse) correspond à chaque groupe (Exemples : COBWEB, CLASSIT);
- des méthodes basées sur l'analyse floue dans laquelle chaque individu de l'ensemble de données appartient à un groupe à un certain degré qui est spécifié par une classe d'appartenance.

En plus des mesures de qualité propres à chaque méthode, un bon regroupement permet la détection des structures présentes dans les données et l'obtention d'un nombre optimal de groupes bien différenciés et stables (Tufféry, 2010).

2.3.2 Méthode de partitionnement K-moyennes (*Kmeans*)

K-moyennes est une méthode regroupement basée sur le partitionnement. Elle se déroule selon les étapes suivantes (Tufféry, 2010):

- Choix des k individus comme centres initiaux des groupes;
- Calcul des distances entre chaque individu et chaque centre et affectation de l'individu au centre le plus proche (k groupes formés);
- Remplacement des k centres par les centroïdes⁴ des k groupes déterminés à l'étape 2;
- Arrêt du traitement si les centroïdes restent suffisamment stables ou le nombre fixé d'itérations est atteint. Autrement, reprise du traitement depuis l'étape 2.

L'algorithme de K-moyennes recalcule le centroïde de chaque groupe à chaque fois qu'un nouvel individu est ajouté dans le groupe. Il permet donc une convergence plus rapide et une amélioration continue des groupes obtenus. Sa complexité linéaire (temps d'exécution proportionnel au nombre d'individus) le dispose à être utilisé pour un grand volume de données. Mais, il faut noter certains inconvénients propres à ce type de regroupement tels que (Tufféry, 2010):

- Les résultats dépendent du choix des centres initiaux et de l'ordre d'ajout des individus pendant le déroulement;
- Le nombre de groupes (k) fixé au départ du traitement peut nuire aux résultats s'il ne correspond pas à la véritable configuration des données.

⁴ Centroïde : Un centroïde (centre du groupe) est représenté par la moyenne des valeurs des attributs des individus du cluster (Ratté-2, 2013).

Afin de remédier à ces inconvénients, l'algorithme de K-moyennes pourrait être appliqué sur plusieurs échantillons tirés aléatoirement pour déduire les centres du meilleur regroupement trouvé et les appliquer, par la suite, comme centres initiaux à la totalité de la population. Également, plusieurs valeurs de k (nombre de groupe) pourraient être testées afin de trouver la meilleure partition possible (Tufféry, 2010).

2.3.3 Indicateurs de qualité du regroupement

L'évaluation de la validité des résultats des méthodes de regroupement se fait grâce à deux types d'indices de validité : les indices de validité interne et les indices de validité externe.

- Les indices de validité interne : Ces indices permettent d'évaluer les méthodes de regroupement en se basant sur les données partitionnées. Parmi ces indices, nous décrivons :
 - Critère de Calinski–Harabasz (*VRC : Variance Ratio Criterion*): Il est utilisé pour évaluer le nombre optimal de groupes. Ce dernier est déduit de la solution avec la valeur la plus élevée de ce critère. Le critère de Calinski–Harabasz est mieux adapté pour la méthode K-moyennes du fait de l'utilisation de la distance euclidienne au carré. Il est défini par (MathWorks-2, 2014) :

$$VRC_k = \frac{SS_B}{SS_W} * \frac{N-k}{k-1} \quad (2.18)$$

avec SS_B : variance globale entre les *groupes*

$$SS_B = \sum_{i=1}^k n_i ||m_i - m||^2 \quad (2.19)$$

et SS_W : variance globale dans les *groupes*.

$$SS_W = \sum_{i=1}^k \sum_{x \in C_i} ||x - m_i||^2 \quad (2.20)$$

Sachant que : N est le nombre d'observations, k est le nombre de groupes, m_i est le centroïde du groupe i , m est la moyenne globale des données de l'échantillon, x est un point du groupe, c_i le $i^{\text{ème}}$ groupe, $||x - m_i||^2$ et $||m_i - m||^2$ représentent la distance entre les deux vecteurs.

- Silhouette (S_i) : À chaque individu de l'ensemble de données, est affectée une valeur de la Silhouette qui mesure la similarité de cet individu avec les individus de son

propre groupe comparativement aux individus des autres groupes. La valeur du critère Silhouette varie entre -1 et 1. Une valeur élevée de celle-ci indique que l'individu est bien associé à son propre groupe. Si la plupart des individus possèdent des valeurs élevées de silhouette, le résultat de partitionnement est approprié. Par contre, si la plupart des individus possèdent des valeurs faibles ou négatives de Silhouette, le résultat de partitionnement est considéré comme inapproprié à cause de trop ou trop peu de groupes. Elle est calculée comme suit (MathWorks-3, 2014):

$$S_i = \frac{\hat{d}_i - d_i}{\max(d_i, \hat{d}_i)} \quad (2.21)$$

Sachant que d_i est la distance moyenne entre l' $i^{\text{ème}}$ individu et les individus de son groupe et \hat{d}_i est la distance moyenne minimale entre l' $i^{\text{ème}}$ individu et les individus des autres groupes.

- Les indices de validité externe : Ces indices permettent d'évaluer les méthodes en se basant sur des données non partitionnées. Il s'agit de données de référence créées et préclassées par des experts (humains). Parmi ces indices, nous définissons :
 - Mesure de Rand (RI) : Elle calcule la similitude entre les groupes produits par la méthode de regroupement utilisée et la classification de référence. Plus RI est proche de 1, plus la répartition des éléments réalisée par la méthode est proche de la classification réelle (classification de référence). La mesure de Rand est définie par (Manning, 2008) :

$$RI = \frac{VP + VN}{VP + FP + FN + VN} \quad (2.22)$$

Où VP est le nombre des vrais positifs, VN est le nombre des vrais négatifs, FP est le nombre des faux positifs et FN est le nombre des faux négatifs.

- F-mesure (F_β): Elle prend en considération le rappel (R) et la précision (P). Un paramètre β permet de pondérer R et P . Plus la valeur de F-mesure est proche de la valeur 1, meilleur est la performance de la méthode de regroupement. Elle est définie par (Roche, 2008) :

$$F_{\beta} = \frac{(\beta^2 + 1) * P * R}{\beta^2 * P + R} \quad (2.23)$$

Où la précision P est définie par le nombre de résultats corrects (VP) divisé par le nombre de tous les résultats retournés ($VP+FP$);
 et le rappel R défini par le nombre de résultats corrects (VP) divisé par le nombre des résultats qui auraient dû être retournés ($VP+FN$).

2.4 Conclusion

Nous avons présenté dans ce chapitre les deux méthodes qui seront employées pour l'étude des données du jeu. La régression linéaire multiple est utilisée pour trouver les attributs les plus influents sur les performances des joueurs, alors que le modèle de regroupement sert à trouver des groupes de joueurs en prenant en considération les attributs influents. Les indicateurs de qualité des deux modèles nous permettent d'évaluer nos résultats et éventuellement valider nos interprétations et nos conclusions.

Dans le chapitre suivant, nous présentons notre scénario d'analyse qui guidera la partie expérimentale de notre étude.

CHAPITRE 3

SCÉNARIOS D'ANALYSE

3.1 Introduction

Afin de mener à bien notre étude, nous sommes amenés à définir des scénarios d'analyse et de discuter de leur pertinence et de leur faisabilité. Dans notre cas, nous pouvons distinguer deux principaux scénarios :

- Le scénario s'intéressant principalement au jeu et à la progression des joueurs en vue de déceler les tendances, les forces et les lacunes du jeu pour pouvoir l'améliorer;
- Le scénario s'intéressant particulièrement aux joueurs, leurs caractéristiques, leurs performances, leurs profils dans le but d'explorer les effets du jeu sur les habilités des joueurs, leurs apprentissages et éventuellement adapté le jeu à leurs profils et à leurs objectifs.

En premier lieu, nous détaillons les deux scénarios tous en discutant de leur faisabilité. En deuxième lieu, nous expliquons le scénario retenu et les questions de recherche que nous allons étudier. Nous terminons ce chapitre par une conclusion.

3.2 Scénarios d'analyse

3.2.1 Scénario d'analyse basé sur le jeu

Du point de vue de la compagnie créatrice du jeu, l'intérêt de l'analyse des données est porté sur les possibilités de pouvoir trouver les quêtes⁵ les plus populaires, les activités (ou exercices) et les actions les plus fréquentes dans les sessions de jeu. Cela permettra à la compagnie d'améliorer les quêtes les moins attrayantes et d'adapter leurs niveaux de

⁵ Quête : une quête consiste à résoudre des énigmes, mener des combats, relever des défis et obtenir des récompenses.

difficulté ainsi que le développement de nouveaux jeux pour offrir une expérience divertissante et enrichissante optimale aux joueurs et par conséquent une meilleure fidélisation de sa clientèle. Seulement, n'ayant pas les ressources et l'expertise nécessaires, la compagnie ne pouvait pas accomplir l'analyse et le forage d'un aussi grand volume de données. Notre contribution consisterait à apporter une solution à cette problématique en réalisant cette analyse.

Dans ce cadre, l'étude se focalisera sur les quêtes, les visites, et les réalisations des joueurs. L'analyse des données des joueurs (caractéristiques et performances) permettra d'extraire de nouvelles connaissances relatives au jeu et à l'expérience du joueur dans celui-ci.

3.2.2 Scénario d'analyse basé sur le joueur

L'intérêt à étudier les caractéristiques, les performances et les profils des joueurs peut être exprimé par les intervenants dans les domaines de l'éducation et de la santé tels que les enseignants et les médecins. Dans ces domaines, le jeu est utilisé comme un outil d'aide à l'apprentissage ou à la stimulation cognitive. Les objectifs d'une telle analyse peuvent être la détermination des profils des joueurs (élèves, patients), le suivi de l'évolution des joueurs et l'évaluation de l'impact du jeu sur les résultats scolaires pour des élèves ou sur les fonctions cognitives chez des patients souffrant de troubles cognitifs.

Pour cela, les caractéristiques du joueur, ses actions et ses résultats dans le jeu devront être analysés et confrontés à ses résultats dans des tests scolaires (élève) ou des batteries de tests neurocognitifs (patient) sur une période de temps significative pour pouvoir tirer des conclusions sur la pertinence du jeu et sa contribution aux résultats scolaires ou aux fonctions cognitives. La réalisation de cette étude est complexe, voire impossible, compte tenu des nombreuses exigences et de la durée relativement courte du mémoire de maîtrise. Nous dénombrons certains obstacles à cette étude dans la liste suivante :

- L'élaboration de protocole de recherche et l'obtention des accords requis pour la partie expérimentale dans les institutions de santé (ou enseignement);

- La participation de volontaires sur une période de temps plus longue (comparativement à la première approche);
- Le recrutement de jeunes patients souffrants de troubles psychotiques (population cible) plus long, faible en effectif et incertain pour toute la durée de l'étude;
- L'accès aux données du jeu, données liées à l'état de santé du patient obéit à des règles d'éthiques très strictes;
- L'intervention d'expertise en neurosciences (ou en éducation) pour l'élaboration des tests neurocognitifs (scolaires), l'interprétation des résultats de ces tests ainsi que la détermination des profils des joueurs;
- La difficulté d'obtenir un financement pour une étude se consacrant à 1% « seulement » de la population (pourcentage des personnes atteintes de troubles psychotiques au Canada).

3.2.3 Choix du scénario d'analyse

Vu la disponibilité des données du jeu, la contrainte du temps (durée pour un mémoire de maîtrise), la non disponibilité de l'expertise et la difficulté à réunir toutes les conditions requises pour procéder à l'expérimentation auprès des patients atteints de troubles psychotiques, nous avons opté pour le choix du scénario d'analyse basé sur le jeu auquel nous ajoutons la recherche de regroupements de joueurs provenant du scénario d'analyse basé sur le joueur et éventuellement la détermination des caractéristiques de ces regroupements. Il s'agit donc d'une approche « hybride » s'intéressant d'une part au jeu et son évolution et d'autre part aux joueurs et à leurs caractéristiques. Cela se justifie par : 1) la disponibilité des données du jeu issues de la compagnie créatrice (données des joueurs en ligne); 2) l'absence d'exigences en matière d'accès aux données (autorisation des participants) car les données sont anonymisées.

L'approche que nous avons choisi permettra d'explorer le jeu et ses données plus en profondeur, de déduire ses éléments à renforcer ou à améliorer, de rechercher des regroupements de joueurs et de définir leurs caractéristiques. Les résultats de cette approche,

entre autre le regroupement, pourront éventuellement être exploités par le scénario basé sur le joueur dans des études futures une fois toutes les conditions requises réunies.

Notre étude cherche à explorer les relations qui peuvent exister entre les caractéristiques des joueurs et leurs performances. Ces relations, si elles existent, pourraient aider les concepteurs de la plateforme de jeux « Science en jeu » à mieux orienter le développement de nouveaux jeux et à améliorer leur produit avec une meilleure adaptation du contenu et de la forme des jeux à leurs clients. Également, nous explorons l'existence, éventuelle, de regroupements de joueurs et les caractéristiques qui identifieront ces regroupements. Des résultats concluants dans ce sens pourraient aider dans la détermination des profils des joueurs.

En premier lieu, nous nous intéressons aux relations existantes entre les caractéristiques du joueur et ses performances dans le jeu. Nos questions de recherche dans ce cas peuvent être énoncées comme suit :

- Quelle est l'influence des caractéristiques du joueur (âge, langue, sexe et intérêt scientifique) et des caractéristiques de ses accès au jeu (nombre de visites, nombre de quêtes, nombre d'avantages, abonnement, nombre de victoires et nombre de défaites) sur les scores (nombre de neurones, nombre de talents et nombre de cortex) obtenus?
- Quelle est l'influence des caractéristiques du joueur (âge, langue, sexe et intérêt scientifique) et des caractéristiques de ses accès au jeu (nombre de visites, nombre de quêtes, nombre d'avantages, abonnement, nombre de victoires et nombre de défaites) sur les durées totales des sessions réalisées?

En deuxième lieu, nous explorons l'existence de regroupements de joueurs autour de caractéristiques spécifiques. Ainsi, nos questions de recherche sont énoncées comme suit :

- Quelles sont les caractéristiques communes (ou proches) entre les joueurs pouvant amener à des regroupements?
- Les regroupements trouvés peuvent-ils être identifiés à des profils de joueurs?

3.3 Conclusion

Nous avons défini dans ce chapitre le scénario d'analyse de notre étude qui s'intéresse principalement au jeu et son évolution ainsi qu'aux joueurs et à leurs caractéristiques. Concernant le premier volet de l'étude, les questions de recherche se rapportent aux influences des caractéristiques du joueur et de ses accès au jeu sur ses performances (scores et durée de jeu). Tandis que les questions de recherche du deuxième volet s'intéressent aux regroupements de joueurs et à l'identification de leurs profils.

Le scénario d'analyse basé sur les données du jeu nous amène à considérer les choix suivants lors de l'étape de modélisation expliquée dans le chapitre 4 :

- Modèle de régression linéaire multiple qui à défaut de donner des modèles prédictifs précis, peut nous renseigner sur l'influence de chaque variable explicative (caractéristiques du joueur) sur les variables à expliquer (performances du joueur);
- Méthode de regroupement (partitionnement) qui nous permet d'explorer (méthode non supervisée) l'existence de regroupements de joueurs et de déceler leurs caractéristiques communes. En particulier, c'est la méthode K-moyennes qui sera utilisée.

CHAPITRE 4

EXPÉRIMENTATION

4.1 Introduction

Nous présentons dans ce chapitre, les étapes détaillées de notre analyse des données du jeu sérieux « Science en jeu ». Nous avons commencé la première étape par la description du jeu et de la base de données qui lui est associée. La compréhension de cette BD nous a permis de cerner les données à retenir pour notre étude conformément au scénario d'analyse choisi. Par la suite, nous avons procédé à la préparation de ces données en éliminant les valeurs aberrantes et les valeurs extrêmes. Lors des étapes suivantes, nous avons testé deux méthodes d'analyses (régression linéaire multiple et regroupement) sur nos données et nous avons analysé et interprété les résultats obtenus. Une conclusion a permis de clore ce chapitre et de souligner les principaux résultats.

4.2 Prétraitement

4.2.1 Description du jeu sérieux « Science en jeu »

La plateforme de jeux sérieux en ligne « Science en jeu » a été créée par la compagnie CREO⁶, elle est sous la forme d'un monde virtuel dédié aux sciences, aux technologies et à la réussite scolaire (*Voir* ANNEXE VII, page 113). Les joueurs explorent le monde à l'aide de leurs avatars⁷ et découvrent divers contenus tels que: des jeux, des informations, des photos, des vidéos, ainsi que des objets d'apprentissage interactifs à travers les îles virtuelles thématiques (Figure 4.1).

⁶ CREO : CREO Inc. est une compagnie qui produit et diffuse des outils de vulgarisation scientifique multiplateformes (<http://www.creo.ca/>).

⁷ Avatar : C'est un personnage numérique, réaliste ou fantaisiste, qui représente le joueur dans un monde virtuel (Ex. : jeux vidéo, logiciels de simulation) (OQLF, 2009).

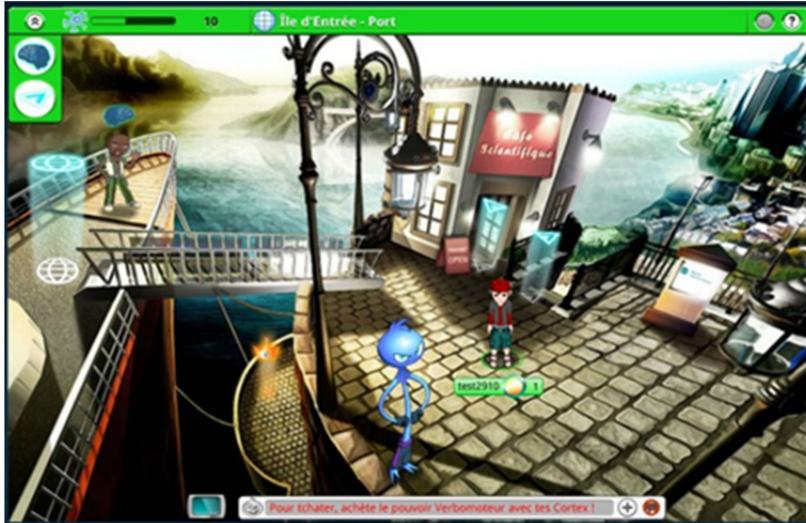


Figure 4.1 Ile d'entrée du jeu « Science en jeu »
Tirée de CREO (2014)

En progressant dans les activités du jeu, le joueur gagne trois types de points :

- des neurones qui représentent les points gagnés en accomplissant des missions dans le jeu (progression dans le jeu);
- des talents qui représentent la monnaie du monde virtuel. Ils sont gagnés en même temps que les neurones et aussi en réalisant certaines actions dans le jeu (répétition de tâches, mini-jeux, etc.) et servent à se procurer différents items;
- des cortex qui représentent le nombre de niveaux franchis par le joueur. Ils servent à acquérir des pouvoirs (interaction avec les autres joueurs, plus d'options dans le jeu, etc.).

Une BD, nommée « Scienceenjeu », est associée au jeu. Elle stocke, d'une part, les contenus et les descriptions des composants du jeu (jeu-questionnaire, concours, quêtes, évaluation, messages, items, etc.) et d'autre part, les données sur les joueurs et leurs interactions avec le jeu (accès au jeu : janvier 2009 à octobre 2013). Cette BD est une importante source d'information pour l'amélioration de la plateforme et le développement de nouveaux jeux qui répondent au mieux à la demande de la clientèle.

4.2.2 Description de la base de données «Scienceenjeu»

La BD MySQL «Scienceenjeu», de taille 655 MiB (près de 9 132 574 enregistrements), est composée de 124 tables. Avant d'être remises pour l'étude, les données des joueurs (avatars) ont été anonymisées par la compagnie propriétaire (suppression des données personnelles telles que l'adresse et le courriel du joueur). Les tables de cette BD peuvent être subdivisées en trois parties : 1) les tables pour le fonctionnement et le contenu du jeu tels que les tables pour les données sur les quêtes, les abonnements, les concours, les questionnaires, les personnages du jeu, etc.; 2) les tables pour les données des joueurs (avatars), leurs caractéristiques et leurs réalisations et 3) les tables statistiques pour recenser toutes les sessions de jeu, toutes les actions des joueurs, etc.

Conformément à l'objectif de la présente étude, nous avons sélectionné uniquement les tables pour les données des joueurs (caractéristiques et réalisations) et les tables statistiques. L'étude de ces tables nous a permis de déterminer les éléments suivants :

- Les données des joueurs :
 - Identificateur, pseudonyme, date de naissance, sexe, langue, pays, intérêt scientifique, nombre de neurones, nombre de talents, nombre de cortex, date de première visite, date de dernière visite, date de suppression, etc.;
 - Liens entre les joueurs (avatar_friend);
 - Caractéristiques des avatars associés aux joueurs (couleur de la peau, cheveux et vêtements);
 - Items utilisés, ainsi que les abonnements.
- Les données sur les réalisations des joueurs telles que les salles et les zones visitées, les votes, les réponses aux questionnaires, les requêtes soumises, les duels, les avantages obtenus, les scores et les dates de début et de fin correspondantes, complétion des étapes des quêtes, etc.;
- Le nombre total des attributs : 212 (171 attributs pour les tables des données des joueurs et 41 pour les tables statistiques);
- Les types des attributs : numérique (139), nominal (22), texte (13), date et heure (38);

- Les tables vides qui sont à éliminer de l'ensemble des tables à considérer pour la présente étude (Exemple : table des identifiants Facebook des joueurs, table des courriels des parents);
- Les attributs pertinents présélectionnés pour notre étude (Tableau 4.1) et les attributs redondants éliminés.

Tableau 4.1 Les attributs présélectionnés pour l'étude

	Attributs	Nombre
Caractéristiques du joueur	identificateur, date de naissance, sexe, langue, pays, intérêt scientifique.	6
Parcours du joueur	identificateur de session, date de début et de fin de session, nombre de talents, nombre de neurones, nombre de cortex, nombre de visites au jeu, date de suppression du joueur, abonnement, identificateur de la quête, identificateur de l'avantages, nombre de victoires, nombre de défaites.	13
	Total	19

4.2.3 Description des outils

Le choix des outils, qui permettent l'analyse de nos données, a été guidé d'une part par la nature de la BD (BD MySQL) pour l'outil phpMyAdmin et le type des fichiers de données extraites (CSV et XLSX) pour Microsoft Excel. D'autre part, le volume de données, leurs types et les méthodes envisagées pour l'étape de modélisation nous ont orientés vers l'utilisation de Matlab. Dans ce qui suit, nous présentons ces outils avec plus de détails :

- phpMyAdmin (version 3.4.10.1) sur WampServer⁸ : C'est une interface pour l'administration Web de base de données MySQL. Elle a été utilisée pour : 1) étudier la structure et le contenu de la BD; 2) déterminer la taille de la BD et de ses tables; 3) exécuter des requêtes SQL pour extraire les données pertinentes et 4) importer la BD et exporter les résultats des requêtes sous format CSV ou XLSX.
- Microsoft Excel 2010 : C'est un tableur de la suite bureautique Microsoft Office. Il a été utilisé afin de: 1) lire les fichiers CSV et XLSX contenant les données extraites de la BD, 2) faire une première analyse des valeurs des attributs.
- Matlab (version R2013a) : C'est un langage de haut niveau et un environnement interactif pour le calcul numérique, la visualisation et la programmation. Il permet aux utilisateurs l'analyse des données, le développement d'algorithmes et la création de modèles et d'applications dans différents domaines tels que : le traitement du signal et des communications, le traitement d'image et de vidéo, les systèmes de contrôle, de test et de mesure, etc. (Matlab, 2014). Nous l'avons utilisé essentiellement pour : 1) explorer et décrire nos données (statistiques descriptives, représentation graphiques,...); 2) transformer les données; 3) détecter et éliminer les données aberrantes ou extrêmes; 4) appliquer et tester des modèles sur nos données et 5) visualiser les résultats.

4.2.4 Préparation des données

Les données du jeu sérieux « Science en jeu » sont stockées dans une BD volumineuse dont les données sont variées et non uniformes; en plus leurs types, leurs sens et leur nombre changent dépendamment du type de quête. Également, les attributs sont nombreux et ne sont pas tous pertinents pour notre étude. Nous avons donc commencé par présélectionner les données pertinentes et relatives au scénario d'analyse choisie (Tableau 4.1). Ces données sont ensuite examinées en vue de connaître leur distribution, de les transformer

⁸WampServer : WampServer est une plateforme de développement Web sous Windows. Il permet de créer des applications web dynamiques à l'aide d'Apache, du langage de scripts PHP et d'une base de données MySQL (<http://www.wampserver.com/>).

éventuellement pour faciliter leur manipulation, et de détecter les valeurs rares ou manquantes et celles qui sont aberrantes ou extrêmes. Détaillée dans cette section, cette étape de préparation de données est primordiale car elle nous permet d'approfondir notre compréhension des données, de les fiabiliser et ainsi construire notre corpus de données à analyser.

4.2.4.1 Transformation des données

Nous avons effectué plusieurs transformations sur les données, car d'une part nous voulons obtenir des attributs quantifiables et faciliter leur manipulation, et d'autre part nous envisageons l'utilisation du modèle de régression linéaire multiple qui exige des variables de type numérique continues ou catégorielles. Les transformations effectuées sont comme suit :

- L'attribut date de naissance est remplacé par l'âge du joueur calculé à partir de la date d'inscription et de la date de naissance (date d'inscription - date de naissance);
- Les attributs date de début de session et date de fin de session sont remplacés par la durée de session et les durées totales et moyennes des sessions sont calculées pour chaque joueur;
- L'attribut sexe, qui contenait les valeurs « f » pour fille et « g » pour garçon, est transformé respectivement avec les valeurs « 1 » et « 2 »;
- L'attribut langue, qui contenait les valeurs « FR » pour français et « EN » pour anglais, est transformé respectivement avec les valeurs « 1 » et « 2 »;
- L'attribut intérêt scientifique contenait un grand nombre de valeurs différentes entrées par les joueurs. Nous avons donc rassemblé ces valeurs sous plusieurs catégories et nous avons attribué à chacune d'elles un numéro de catégorie. Par exemple, les valeurs « aéronautique » et « avions et fusées » sont affectées à la catégorie aéronautique et elles sont remplacées par la valeur « 1 » (*Voir ANNEXE I, Tableau-A I-1*);
- Les attributs nombre de quêtes et nombre d'avantages sont calculés pour chaque joueur à partir des tables de la BD.

4.2.4.2 Détection et élimination des valeurs aberrantes

Afin d'obtenir un corpus de données fiable sur lequel se basera notre expérimentation et notre analyse, nous devons éliminer les valeurs aberrantes. Celles-ci peuvent être soit des valeurs erronées, manquantes, incomplètes ou n'appartenant pas au cadre et à l'objectif de la présente étude. En effet, nous nous intéressons dans ce cas aux différents parcours des joueurs dont les caractéristiques sont sauvegardées et dont les performances ont été réalisées dans un contexte de jeu et non pas de test et de développement de ce jeu. Pour cela, nous avons éliminé de notre corpus les deux types de joueurs suivants:

- Les joueurs externes (ou touristes) : Ceux sont les joueurs qui participent au jeu sans s'inscrire. Ils ont des choix restreints dans le jeu et le système ne possède pas les données les caractérisant (seulement un identificateur temporaire de valeur négative attribué par le système). Il n'est pas alors pertinent de les garder, car leur expérience de jeu n'est pas comparable à celle des joueurs inscrits (accès restreint aux options de la plateforme du jeu);
- Les joueurs (avatars) « système » : Ceux sont les développeurs du jeu qui y accèdent pour tester le jeu. Une fois identifiées, les données de ces joueurs sont supprimées, car ils possèdent une expérience particulière, voire optimale, du jeu en comparaison avec le reste des joueurs. Ceci peut induire en erreur lors du traitement et l'analyse des résultats.

Également, l'étude des valeurs des attributs présélectionnés nous a révélé des problèmes de fiabilité dans certains attributs comme la présence de valeurs aberrantes ou des attributs contenant la même information avec des noms différents ou encore des attributs vides dans la majorité des enregistrements. Dans ce qui suit, nous citons des exemples de ces cas:

- Plusieurs valeurs erronées dans l'attribut date de naissance (du joueur) ont été détectées telles que: 0000-00-00, 1900-07-15, et 2988-07-02. Le calcul du pourcentage moyen d'occurrences de ces données erronées (près 1,66 %) étant faible, nous avons opté pour garder cet attribut, calculer l'âge du joueur et remplacer les valeurs de l'âge erronées (inférieures à 6 ou supérieures à 90) par la valeur 100. Ceci nous permet de repérer cette valeur et de l'écarter du traitement par la suite;

- L'attribut intérêt scientifique contenait certaines valeurs aberrantes telles que : « dddddd » et « blabla ». Les individus du corpus contenant ces données aberrantes sont supprimés;
- Les attributs nombre de points et nombre de neurones sont employés dans différentes tables (Avatar et Session de jeu) et contiennent la même information. Seulement, l'attribut nombre de neurones est retenu pour éviter les redondances;
- Après la suppression des individus du corpus avec des valeurs aberrantes et/ou manquantes, nous avons constaté que l'attribut pays possédait une seule valeur (Canada) pour tous les individus de la population étudiée. Étant devenue constante, la variable pays n'a plus d'influence sur les modèles étudiés. Elle est donc écartée de l'ensemble des attributs considérés pour cette étude.

À la fin de cette section, nous établissons la liste des attributs retenus pour l'étude présentée ci-dessous (Tableau 4.2).

Tableau 4.2 Les attributs sélectionnés pour l'étude

	Attributs	Nombre
Caractéristiques du joueur	identificateur du joueur, âge, sexe, langue, intérêt scientifique.	5
Parcours du joueur	nombre de visites, abonnement, nombre de quêtes, nombre des avantages, nombre de victoires, nombre de défaites, nombre de neurones, nombre de talents, nombre de cortex, durée totale des sessions, durée moyenne des sessions.	11
	Total	16

4.2.4.3 Détection et élimination des valeurs extrêmes

Bien qu'elles ne soient pas forcément aberrantes, les valeurs extrêmes des variables affectent plusieurs méthodes d'analyses (méthodes basées sur des calculs de variances, régression

logistique, analyse discriminante, etc.) (Tufféry, 2010). Il est alors important de détecter ces valeurs extrêmes et de les écarter de notre ensemble de données. Pour cela, nous avons utilisé les histogrammes, les graphiques « boîte à moustache » et les intervalles de confiance pour les repérer et un script Matlab pour les supprimer. Nous expliquons dans ce qui suit l'utilisation de ces outils:

- L'observation des histogrammes donne des informations à propos de la dispersion (variété des valeurs), la distribution des valeurs et la présence de valeurs extrêmes ou aberrantes;
- Les graphiques « boîtes à moustache » représentent plusieurs paramètres de la distribution d'une variable (médiane, intervalle interquartile, valeur maximale et minimale de la distribution). Ils permettent aussi de visualiser les valeurs extrêmes qui sont situées à plus de 3 longueurs de boîte à partir de la bordure inférieure ou supérieure de la boîte (SPSS, 2014);
- L'intervalle de confiance est la zone d'acceptation pour nos valeurs. Cette zone se situe autour de la moyenne où se trouve 95% des valeurs (zone de rejet avec α égale à 0,05) en supposant que nos données suivent une loi normale (Tufféry, 2010). L'intervalle de valeurs caractérisant cette zone d'acceptation est défini par (Tufféry, 2010, p. 656):

$$[\text{Moyenne} - 1,96 (\text{Écart-type}), \text{Moyenne} + 1,96 (\text{Écart-type})] \quad (4.1)$$

Nous notons que les attributs langue, sexe et intérêt scientifique sont respectivement de type numérique dichotomique (1,2) et numérique ordinal (1 à 23) et par conséquent ils ne contiennent pas de valeurs extrêmes (*Voir* ANNEXE II : Figure-A II-1, Figure-A II-3 et Figure-A II-4). Concernant le reste des attributs, nous avons identifié les valeurs extrêmes (histogrammes, graphiques « boîtes à moustaches » et intervalles de confiance) et nous les avons supprimées. À l'exception des cas où leur nombre dépassait 2% de l'ensemble de données, nous avons subdivisé ce dernier en sous-population afin d'éviter de perdre un volume important et significatif de données, et nous avons repris la recherche des points extrêmes pour chaque sous-population. En effet, nous avons sélectionné en premier lieu l'attribut nombre de visites. Ses valeurs ont été subdivisées, et ont engendré trois sous-populations. Ensuite, nous avons effectué une deuxième subdivision sur la base des valeurs

des attributs nombre de neurones, nombre de talents, nombre de cortex et durée totale de sessions (Tableau 4.3). À chaque étape, nous nous assurons de supprimer les points extrêmes (sans dépasser 2% de la population étudiée). Nous décrivons dans ce qui suit la démarche effectuée pour supprimer leurs valeurs extrêmes :

- **Attribut âge** : Dans cette étude, nous nous sommes intéressés exclusivement aux données des joueurs âgés de 20 ans et moins (*Voir* ANNEXE II, Figure-A II-2) car d'une part le contenu et la forme de la plateforme de jeux « Science en jeu » est destiné aux enfants et adolescents (niveau scolaire : primaire et secondaire) et d'autre part la compagnie créatrice des jeux (CREO) cible la clientèle âgée de 9 à 14 ans pour cette plateforme. Également, la participation des personnes plus âgées (plus de 20 ans) reste exceptionnelle et marginale et pourrait comporter un risque plus élevé de saisie de valeurs erronées pour les caractéristiques du joueur (âge, langue, sexe, pays et intérêt scientifique);
- **Attribut nombre de visites** : Tout d'abord, nous avons éliminé les individus ayant une valeur nulle pour le nombre de visites car cela veut dire que le joueur inscrit a fait juste une connexion et n'a pas participé réellement aux quêtes du jeu. Ses données ne sont pas pertinentes pour notre étude, car elles ne reflètent pas une expérience dans le jeu. Ensuite, nous avons supprimé les valeurs extrêmes (nombre de visites ≥ 400) (*Voir* ANNEXE II : Figure-A II-5 et Figure-A II-6) et nous avons subdivisé les données en sous-populations de telle manière que le nombre des valeurs considérées comme extrêmes soient inférieur à 2% de l'ensemble de données (*Voir* ANNEXE II, Figure-A II-19) :
 - La sous-population 1 (Vs01) ayant le nombre de visites inférieur à 20;
 - La sous-population 2 (Vs20) ayant le nombre de visites compris entre 20 et 70 ($20 \leq \text{nombre de visites} < 70$);
 - La sous-population 3 (Vs70) ayant le nombre de visites compris entre 70 et 400 ($70 \leq \text{nombre de visites} < 400$).
- **Attribut nombre de quêtes** : La distribution de l'attribut nombre de quêtes dans notre ensemble de données initial (*Voir* ANNEXE II : Figure-A II-7 et Figure-A II-8) a révélé un grand nombre de valeurs nulles. Cela a engendré un grand nombre de valeurs considérées comme extrêmes, car elles sont supérieures à zéro. Toutefois, la subdivision en sous-populations selon le nombre de visites a permis de prendre en compte ces valeurs

- (Voir ANNEXE II, Figure-A II-20). Il reste à noter que lorsque le nombre de visites est inférieur à 20 (sous-population Vs01), le nombre de quêtes complétées est très petit pour la majorité des individus (inférieur à 2). Il s'agit probablement de nouveaux joueurs.
- Attribut nombre d'avantages : en comparant les distributions des valeurs de l'attribut nombre d'avantages avant (Voir ANNEXE II : Figure-A II-9 et Figure-A II-10) et après la subdivision en sous-populations (Vs01, Vs20 et Vs70), nous avons remarqué l'élimination des valeurs extrêmes et la constitution de sous-populations homogènes (Voir ANNEXE II, Figure-A II-21).
 - Attribut nombre de victoires et nombres de défaites : Ces deux attributs nous renseignent sur le nombre de victoires et de défaites du joueur lors de ses duels avec d'autres joueurs. Or, les valeurs de ces attributs sont nulles pour la majorité des individus de la population (plus de 90%). Il n'est donc pas pertinent de les intégrer à notre étude, car ils n'apportent pas plus d'information.
 - Attribut Abonnement : La majorité des joueurs ne possèdent pas d'abonnement.
 - Attribut nombre de neurones : Pour chaque sous-population déterminée avec l'attribut nombre de visites, nous avons recherché les valeurs extrêmes de l'attribut nombre de neurones (Voir ANNEXE II : Figure-A II-11 et Figure-A II-12) et en cas de besoin, nous avons encore subdivisé en sous-populations selon les valeurs du nombre de neurones. Nous avons alors trouvé pour (Voir ANNEXE II : Figure-A II-22 et Figure-A II-23) :
 - La sous-population 1 (Vs01) : les valeurs extrêmes (nombre de neurones > 7000) ont été supprimées et la sous-population a été subdivisée en deux groupes:
 - Le groupe 1 (Vs01Nr1) ayant le nombre de neurones inférieur ou égal à 1000;
 - Le groupe 2 (Vs01Nr2) ayant le nombre de neurones compris entre 1000 et 7000 ($1000 < \text{nombre de neurones} \leq 7000$).
 - La sous-population 2 (Vs20) : Le groupe (Vs20Nr) est formé des individus dont le nombre de neurones est compris entre 10000 et 10802 ($10000 \leq \text{nombre de neurones} < 10802$);
 - La sous-population 3 (Vs70) : Le groupe (Vs70Nr) est formé des individus ayant plus de 10802 neurones ($10802 \leq \text{nombre de neurones}$).

- Attribut nombre de talents : Nous avons recherché les valeurs extrêmes de l'attribut nombre de talents pour chaque sous-population et nous avons effectué une deuxième subdivision selon la valeur de cet attribut (*Voir* ANNEXE II : Figure-A II-13 et Figure-A II-14). Nous avons alors obtenu pour (*Voir* ANNEXE II : Figure-A II-24 et Figure-A II-25) :
 - La sous-population 1(Vs01) : les valeurs extrêmes (nombre de talents > 4000) ont été supprimées et la sous-population a été subdivisée en deux groupes:
 - Le groupe 1 (Vs01T11) ayant le nombre de talents inférieur ou égal à 500;
 - Le groupe 2 (Vs01T12) ayant le nombre de talents compris entre 500 et 4000 ($500 < \text{nombre de talents} \leq 4000$).
 - La sous-population 2 (Vs20) : Le groupe (Vs20T1) est constitué des individus dont le nombre de talents est compris entre 4000 et 6000 ($4000 \leq \text{nombre de talents} < 6000$);
 - La sous-population 3 (Vs70) : Le groupe (Vs70T1) est formé des individus ayant plus de 6000 talents ($6000 \leq \text{nombre de talents}$).
- Attribut nombre de cortex : Nous avons exploré l'existence de valeurs extrêmes de l'attribut nombre de cortex dans chaque sous-population (*Voir* ANNEXE II : Figure-A II-15 et Figure-A II-16). Nous avons alors eu pour (*Voir* ANNEXE II, Figure-A II-26) :
 - La sous-population 1(Vs01) : les valeurs extrêmes (nombre de cortex ≥ 7) ont été supprimées et le groupe (Vs01Cr) a été obtenu;
 - La sous-population 2 (Vs20) : Le groupe (Vs20Cr) est obtenu sans valeurs extrêmes;
 - La sous-population 3 (Vs70) : Le groupe (Vs70Cr) est obtenu sans valeurs extrêmes.
- Attribut durée totale de sessions : En premier lieu, nous avons supprimé tous les individus dont la durée totale des sessions était nulle car cela signifie que soit le joueur s'est connecté au site Web du jeu sans participation aux jeux, soit les données relatives aux sessions de cet individu n'ont pas été sauvegardées dans la BD de notre étude. Ensuite, nous avons recherché les valeurs extrêmes dans les sous-populations (*Voir* ANNEXE II : Figure-A II-17 et Figure-A II-18) et nous avons trouvé pour (*Voir* ANNEXE II : Figure-A II-27 et Figure-A II-28) :

- La sous-population 1 (Vs01) : les valeurs extrêmes (durée totale > 9 h 30 min) ont été supprimées et la sous-population a été subdivisée en deux groupes:
 - Le groupe 1 (Vs01Dt1) ayant la durée totale inférieure ou égale à 2 h 30 min;
 - Le groupe 2 (Vs01Dt2) ayant la durée totale comprise entre 2 h 30 min et 9 h 30 min (2 h 30 min < durée totale <= 9 h 30 min).
- La sous-population 2 (Vs20) : Après la suppression des valeurs extrêmes dont les durées totales étaient supérieures à 24 h, nous avons obtenu le groupe (Vs20Dt);
- La sous-population 3 (Vs70) : Après la suppression des valeurs extrêmes dont les durées totales étaient supérieures à 64 h, nous avons obtenu le groupe (Vs70Dt).

Tableau 4.3 Subdivision de l'ensemble de données en sous-populations

Attribut	Valeurs extrêmes	Code de la sous-population	Intervalle des valeurs	Effectif
Nombre de visites	Visites >= 400	Vs01	Visites < 20	3867
		Vs20	20 <= Visites < 70	432
		Vs70	70 <= Visites < 400	68
Nombre de neurones	Neurones > 7000	Vs01Nr1	Neurones <= 1000	2697
		Vs01Nr2	1000 < Neurones <= 7000	1101
		Vs20Nr	10000 <= Neurones < 10802	413
		Vs70Nr		68
Nombre de talents	Talents > 4000	Vs01TI1	Talents <= 500	2694
		Vs01TI2	500 < Talents <= 4000	1127
		Vs20TI	4000 <= Talents < 6000	419
		Vs70TI	6000 <= Talents	65
Nombre de cortex	Cortex >= 7	Vs01Cr	–	3760
		Vs20Cr	–	432
		Vs70Cr	–	68
Durée totale des sessions	Durée totale > 9 h 30 min	Vs01Dt1	0 < Durée totale <= 2 h 30 min	813
		Vs01Dt2	2 h 30 min < Durée totale <= 9 h 30 min	193
	Durée totale > 24 h	Vs20Dt	–	422
	Durée totale > 64 h	Vs70Dt	–	65

4.3 Modélisation : modèle de régression linéaire multiple

4.3.1 Contexte de l'expérimentation

4.3.1.1 Choix de la méthode de régression et critères d'évaluation du modèle

La régression linéaire que nous avons appliquée est la régression linéaire multiple. Le choix est porté sur une modélisation globale. C'est à dire que la combinaison de toutes les variables indépendantes est évaluée globalement. Ces variables seront insérées par entrée forcée (toutes les variables introduites en même temps). Nous avons effectué cette application avec l'outil Matlab. Ce dernier fournit à chaque modèle généré, plusieurs paramètres pour le décrire et l'évaluer. Par souci de clarté et de lisibilité, nous avons gardé uniquement les coefficients estimés des variables explicatives et leurs *p-values* respectives, le coefficient de détermination R^2 , *F-statistique* et la *p-value* globale pour représenter les résultats de chaque modèle. Aussi, Les graphiques des modèles de régression linéaire multiple obtenus sont représentés dans l'ANNEXE IV (p. 101).

4.3.1.2 Variables dépendantes et variables indépendantes

Conformément au scénario d'analyse choisi, nous avons sélectionné les attributs sur les caractéristiques du joueur et de son parcours comme variables indépendantes et ses résultats comme variables dépendantes. La description de ces variables est introduite dans les tableaux suivants : Tableau 4.4 et Tableau 4.5.

Tableau 4.4 Variables indépendantes

Variables indépendantes	
Âge	Âge du joueur
Langue	Langue dans laquelle le joueur a participé au jeu
Sexe	Sexe du joueur
Intérêt scientifique	Domaine d'intérêt scientifique choisi par le joueur
Nb_visite	Nombre de visite du joueur au jeu
Abonnement	Présence ou absence d'un abonnement au jeu (0/1)
Nb_quêtes	Nombre de quêtes effectuées par le joueur
Nb_avantages	Nombre des avantages utilisés par le joueur
Nb_victoires	Nombre de victoires pendant les duels
Nb_défaites	Nombre de défaites pendant les duels

Tableau 4.5 Variables dépendantes

Variables dépendantes	
Nb_neurones	Nombre de neurones gagnés par le joueur
Nb_talents	Nombre de talents gagnés par le joueur (monnaie du monde virtuel)
Nb-cortex	Nombre de cortex ou nombre de niveaux franchis par le joueur
Duree_totale	Durée totale des sessions de jeu réalisées
Duree_moyenne	Durée moyenne des sessions de jeu

4.3.1.3 Échantillonnage

En raison de la subdivision de notre ensemble de données en plusieurs sous-populations, nous avons obtenus des sous-ensembles de données (*datasets*) de différentes tailles de l'ordre de plusieurs milliers, plusieurs centaines et même une soixantaine d'individus. À cause de cela, nous avons sélectionné les données à tester de deux manières :

- Extraction aléatoire des données quand la sous-population étudiée est de grande taille (plusieurs milliers). Ce type d'échantillonnage est justifié pour les raisons suivantes :
 - l'ensemble de données est de grande taille;
 - La sélection aléatoire évite le biais dans les résultats;
 - La possibilité de vérifier la stabilité des modèles obtenus sur plusieurs échantillons extraits aléatoirement de l'ensemble de données global.
- Prise en compte de la totalité de la sous-population quand celle-ci est de petite taille.

4.3.2 Corrélation entre les variables

Afin d'éviter d'introduire dans nos modèles des variables qui soient fortement corrélées, et par conséquent, qui mesurent la même entité, nous avons effectué le calcul des coefficients de corrélation entre les variables indépendantes et dépendantes deux à deux. Selon Tufféry, la corrélation entre deux variables indépendantes est considérée comme « [...] inacceptable quand le coefficient de corrélation dépasse 0,9, très dangereuse quand le coefficient dépasse 0,8 et à surveiller quand il dépasse 0,7 » (Tufféry, 2010, p. 94) dans le cas de la régression linéaire. En effet, plus la valeur calculée du coefficient de corrélation est proche de 1 (en valeur absolue), plus la relation entre les deux variables correspondantes est forte et par conséquent il serait alors important de ne garder que l'une des deux variables pour la suite de l'étude.

Après l'analyse des coefficients de corrélation obtenus entre les variables indépendantes (Tableau 4.6), nous avons noté les remarques suivantes :

- Il n'y a pas de corrélation entre la variable Langue (respectivement Âge, Sexe, Intérêt scientifique) et les autres variables indépendantes (coefficients de corrélation inférieurs en valeur absolue à 0,2). Par conséquent, ces variables peuvent être considérées ensemble dans un modèle de régression linéaire multiple;
- La présence d'une corrélation importante entre les variables Nb_visite et Nb_victoire (0,7471) et entre Nb_visite et Nb_defaite (0,7152) ainsi que la corrélation entre Nb_victoire et Nb_defaite (0,7186) nous amènent à considérer l'élimination de ces deux

dernières variables (Nb_victoire et Nb_defaite) de l'ensemble des variables indépendantes. Également, le pourcentage d'individus ayant des valeurs nulles pour les variables Nb_victoire et Nb_defaite est très élevé (93,34%). Cela peut s'expliquer par le fait que la majorité des joueurs ne participent pas à la partie duel du jeu. Il n'est donc pas pertinent de garder ces deux variables;

- Vu la valeur du coefficient de corrélation (0,6793) entre la variable Nb_quêtes et Nb_avantages, il faudra prendre des précautions lors de l'application des modèles de régression avec ces deux variables et leur interprétation;
- La majorité des individus de la population n'ont pas d'abonnement pour le jeu (99.90), il n'est donc pas pertinent de garder cette variable.

Tableau 4.6 Coefficients de corrélation entre les variables Langue, Âge, Sexe, Intérêt scientifique, Nb_visite, Abonnement, Nb_quêtes, Nb_avantages, Nb_victoire et Nb_defaite

	Langue	Âge	Sexe	Intérêt	Nb_visite	Abonne-ment	Nb_quetes	Nb_avantag	Nb_victoire	Nb_defaite
Langue	1	-0,0585	-0,0160	-0,0528	-0,0430	0,0379	-0,0540	-0,0951	-0,0236	-0,0425
Âge	-0,0585	1	0,0226	0,1943	-0,0336	-0,0277	-0,0753	-0,1689	-0,0103	-0,0504
Sexe	-0,0160	0,0226	1	-0,1355	0,0332	0,0290	0,0067	0,0250	0,0274	0,0264
Intérêt	-0,0528	0,1943	-0,1355	1	-0,0249	-0,0265	-0,0708	-0,1459	-0,0442	-0,0437
Nb_visite	-0,0430	-0,0336	0,0332	-0,0249	1	0,2641	0,4750	0,4481	0,7471	0,7152
Abonne-ment	0,0379	-0,0277	0,0290	-0,0265	0,2641	1	0,1325	0,1744	0,1756	0,2609
Nb_quetes	-0,0540	-0,0753	0,0067	-0,0708	0,4750	0,1325	1	0,6793	0,3748	0,5568
Nb_avantag	-0,0951	-0,1689	0,0250	-0,1459	0,4481	0,1744	0,6793	1	0,2782	0,4336
Nb_victoire	-0,0236	-0,0103	0,0274	-0,0442	0,7471	0,1756	0,3748	0,2782	1	0,7186
Nb_defaite	-0,0425	-0,0504	0,0264	-0,0437	0,7152	0,2609	0,5568	0,4336	0,7186	1

Aussi, nous avons calculé les coefficients de corrélation (Tableau 4.7) entre les variables dépendantes (Nb_neurones, Nb_talents, Nb_cortex, Duree_totale et Duree_moyenne) et les variables indépendantes (Âge, Sexe et Intérêt scientifique, Nb_visite, Abonnement, Nb_quêtes, Nb_avantages, Nb_victoire et Nb_defaite). Ces résultats nous donnent une première idée sur les effets que peuvent avoir les X_i (variables indépendantes) sur les Y (variables dépendantes) qui seront confirmés ou non lors de l'application du modèle de régression linéaire multiple. Ainsi, nous avons noté :

- une corrélation importante entre Nb_neurones et les deux variables Nb_quêtes (0,7091) et Nb_avantages (0,755);
- une corrélation à considérer entre Duree_totale et les variables Nb_visite (0,7117) et Nb_defaite (0,7748);
- Les coefficients de corrélation entre la variable Duree_moyenne et les variables indépendantes sont faibles de manière générale en comparant avec ceux obtenus avec la variable Duree_totale. De plus, ces deux variables représentent des durées de jeu calculées différemment. Il est donc préférable de ne garder que l'une d'elles. Nous gardons dans ce cas la variable Duree_totale pour la suite de l'étude;
- Les coefficients de corrélation entre les variables indépendantes Langue, Âge, Sexe et Intérêt scientifique et les variables dépendantes sont tous inférieurs à 0,2 (en valeur absolue). Cela refléterait possiblement une faible influence de ces variables sur nos variables dépendantes.

Tableau 4.7 Coefficients de corrélation entre les variables dépendantes et les variables indépendantes

	Langue	Âge	Sexe	Intérêt	Nb_ visite	Abonne -ment	Nb_ quete	Nb_ avantag	Nb_ victoire	Nb_ defaite
Nb_ neuron	-0,1168	-0,1155	0,0057	-0,1286	0,6005	0,1061	0,7091	0,7550	0,4125	0,5216
Nb_ talents	-0,0254	-0,0296	0,0209	-0,0436	0,5455	0,5335	0,2753	0,3040	0,4896	0,5575
Nb_ cortex	-0,0779	-0,1211	-0,0431	-0,1240	0,2034	0,0404	0,4121	0,3993	0,0360	0,1218
Duree_ totale	-0,0177	-0,0513	0,0126	-0,0391	0,7117	0,4004	0,4960	0,4292	0,6311	0,7748
Duree_ moyenn	0,0593	-0,1350	-0,0435	-0,0826	0,0861	0,0095	0,2980	0,3647	0,0383	0,1040

Concernant les corrélations entre les variables dépendantes (Tableau 4.8), il apparaît l'existence d'une corrélation entre la variable Nb_neurones et les variables Nb_cortex (0,7285) et Duree_totale (0,5233). De plus, une autre corrélation est à souligner entre Nb_talents et Duree_totale (0,7833). Ces corrélations peuvent être expliquées par le fait qu'un joueur qui participe au jeu plus longtemps, gagne plus de neurones et de talents et atteint plus de niveaux (cortex). À l'exception de Duree_moyenne, nous garderons pour la suite de l'étude l'ensemble des variables dépendantes, car chacune de ses variables pourrait nous renseigner sur le cheminement du joueur à travers le jeu en termes de score et de durée.

Tableau 4.8 Coefficients de corrélation entre les variables Nb_neurones, Nb_talents, Nb_cortex, Duree_totale et Duree_moyenne

	Nb_neurone	Nb_talents	Nb_cortex	Duree_totale	Duree_moyenne
Nb_neurone	1	0,3792	0,7285	0,5233	0,3196
Nb_talents	0,3792	1	0,1735	0,7833	0,0994
Nb_cortex	0,7285	0,1735	1	0,1511	0,2854
Duree_totale	0,5233	0,7833	0,1511	1	0,2811
Duree_moyenne	0,3196	0,0994	0,2854	0,2811	1

4.3.3 Résultats de l'application du modèle de régression linéaire multiple

Dans cette section, nous présentons les résultats de l'application du modèle de régression linéaire multiple pour chacune des variables dépendantes en spécifiant les variables dépendantes et la sous-population choisie. Les résultats les plus performants en termes de significativité et d'ajustement du modèle (*p-value* et R^2) sont détaillés dans ce qui suit alors que les reste des résultats sont présentés dans l'ANNEXE III (p. 97).

4.3.3.1 Résultats de la régression linéaire multiple pour le nombre de neurones

En appliquant la RLM sur les sous-populations, nous avons obtenu les meilleurs résultats en termes de coefficients de détermination avec les deux sous-populations Vs01Nr1 (0,559) et Vs70Nr (0,744). D'après ces résultats, les variables explicatives Nb_quêtes, Nb_avantages et Intérêt scientifique sont les plus influentes sur le nombre de neurones dans la première sous-population (Tableau 4.9). Tandis que les variables explicatives Nb_quêtes, Nb_avantages et Nb_visite sont influentes dans la deuxième sous-population (Tableau 4.10). Les résultats les moins performants sont ceux obtenus pour les sous-populations Vs01Nr2 et Vs20Nr (*Voir* ANNEXE III : Tableau-A III-1 et Tableau-A III-2). Néanmoins, il est à noter l'influence de la variable Langue dans la sous-population Vs20Nr.

Tableau 4.9 MRLM avec Nb_neurones pour la sous-population Vs01Nr1
(Nombre d'observations de l'échantillon : 270)

Coefficients estimés du modèle								
Variables	Constante	Langue	Age	Sexe	Intérêt	visite	quête	avantage
Coefficients	274	-0.510	-4.171	-26.78	-5.401	4.3618	-584.67	104.06
<i>p-value</i>	< 0,000	0.9868	0,2302	0,2014	0,0008	0,1232	0,001	< 0,000
Paramètres du modèle								
R^2			<i>F-stat</i>			<i>p-value</i>		
0,559			47,4			< 0,000		

Tableau 4.10 MRLM avec Nb_neurones pour la sous-population Vs70Nr
(Nombre d'observations: 68)

Coefficients estimés du modèle								
Variables	Constante	Langue	Age	Sexe	Intérêt	visite	quête	avantage
Coefficients	3521,3	-1279,6	-28,589	9,0158	76,044	37,48	547,38	-224,11
p-value	0,1734	0,4196	0,7959	0,99	0,1617	<0,000	< 0,000	0,0297
Paramètres du modèle								
R²			F-stat			p-value		
0,744			24,9			< 0,000		

4.3.3.2 Résultats de la régression linéaire multiple pour le nombre de talents

Concernant les modèles Nb_talents obtenus, les valeurs des coefficients de détermination sont demeurées inférieures à 0,25 pour toutes les sous-populations. C'est une assez faible valeur représentant seulement un ajustement de moins de 25% des individus au modèle. Toutefois, nous avons relevé l'effet des variables Langue, Nb_quêtes et Intérêt scientifique sur le nombre de talents gagnés dans la sous-population Vs01T11 (Tableau 4.11). Il apparaît aussi les contributions de Nb_visite et Nb_avantages (*Voir* ANNEXE III : Tableau-A III-3, Tableau-A III-4 et Tableau-A III-5).

Tableau 4.11 MRLM avec Nb_talents pour la sous-population Vs01T11
(Nombre d'observations de l'échantillon : 270)

Coefficients estimés du modèle								
Variables	Constante	Langue	Age	Sexe	Intérêt	visite	quête	avantage
Coefficients	183,98	-31,45	1,8213	-0,445	-2,711	-0,9877	-11,92	17,228
p-value	< 0,000	0,0266	0,2591	0,9631	0,0002	0,5084	<0,000	0,3321
Paramètres du modèle								
R²			F-stat			p-value		
0,242			12			< 0,000		

4.3.3.3 Résultats de la régression linéaire multiple pour le nombre de cortex

Le meilleur coefficient de détermination a été réalisé pour la variable Nb_cortex dans la sous-population Vs70Cr. Dans ce cas, nous avons noté les influences significatives des variables Langue, Nb_avantages et Nb_quêtes (Tableau 4.12) sur le nombre de cortex. Tandis que dans les autres cas (*Voir* ANNEXE III : Tableau-A III-6 et Tableau-A III-7), nous avons relevé les contributions des variables Intérêt scientifique et Nb_visite dans la sous-population Vs20Cr; et aussi des contributions des variables Âge et Sexe dans la sous-population Vs01Cr.

Tableau 4.12 MRLM avec Nb_cortex pour la sous-population Vs70Cr
(Nombre d'observations: 68)

Coefficients estimés du modèle								
Variables	Constante	Langue	Age	Sexe	Intérêt	visite	quête	avantage
Coefficients	8,6098	4,471	-0,179	-0,273	-0,0263	-0,0002	0,3702	-0,7032
<i>p-value</i>	0,0026	0,0104	0,1342	0,7244	0,649	0,9732	0,0003	< 0,000
Paramètres du modèle								
<i>R</i> ²			<i>F-stat</i>			<i>p-value</i>		
0,517			9,17			< 0,000		

4.3.3.4 Résultats de la régression linéaire multiple pour la durée totale des sessions

D'après les coefficients de détermination des modèles pour la variable représentant la durée totale des sessions dans les différentes sous-populations, nous avons constaté que celui calculé pour la sous-population Vs70Dt était le meilleur (Tableau 4.13). Les coefficients et leurs *p-values* nous ont renseigné sur l'importance de l'influence des variables Sexe, Nb_visite et Nb_avantages sur la durée totale passée dans le jeu. De plus, nous avons remarqué que le nombre de quêtes contribuait aussi dans les sous-populations Vs01Dt2 et Vs20Dt (*Voir* ANNEXE III : Tableau-A III-9 et Tableau-A III-10).

Tableau 4.13 MRLM avec Duree_totale pour la sous-population Vs70Dt
(Nombre d'observations: 65)

Coefficients estimés du modèle								
Variables	Constante	Langue	Age	Sexe	Intérêt	visite	quête	avantage
Coefficients	-0,0847	-0,237	0,0378	-0,382	-0,0056	0,0039	0,0244	0,0391
p-value	0,8537	0,3971	0,0621	0,0048	0,5662	0,0021	0,1297	0,0331
Paramètres du modèle								
R²			F-stat			p-value		
0,584			11,4			< 0,000		

4.3.4 Analyse des résultats

Globalement, nous avons constaté que les valeurs de *p-value* des modèles obtenus étaient toutes significatives ($< 0,05$). Aussi, les coefficients de détermination (R^2) calculés variaient selon la variable dépendante et la sous-population. En effet, nous avons obtenu un R^2 égal à 0,744 dans le modèle de la variable Nb_neurones pour la sous-population Vs70Nr (Tableau 4.10) reflétant un bon ajustement du nuage de points au modèle, contrairement à celui de Nb_talents pour la sous-population Vs20T1 égal à 0,052 (Voir ANNEXE III, Tableau-A III-4) suggérant l'importance de l'erreur. Également, l'application de ces modèles sur plusieurs échantillons extraits aléatoirement des sous-populations Vs01Nr1, Vs01Nr2, Vs01T11, Vs01T12, Vs01Cr et Vs01Dt1 a montré une certaine stabilité pour les valeurs des R^2 et des *p-value* (Voir ANNEXE V, page 105). Cela nous permet de consolider l'interprétation de nos résultats et nos conclusions.

Une analyse plus détaillée des résultats, nous a permis aussi de remarquer que les variables Nb_quêtes, Nb_avantages et Nb_visite ont un effet sur les scores des joueurs (neurones, talents et cortex) dans presque la totalité des sous-populations. À ces effets viennent s'ajouter les contributions des variables Langue et Intérêt scientifique ainsi que les variables Sexe et Âge, mais avec une moindre importance.

En effet, la comparaison des coefficients (β) des modèles obtenus relatifs aux scores montre que les joueurs, qui ont visité le jeu moins de 20 fois (Tableau 4.9, Tableau 4.11 et *voir ANNEXE III* : Tableau-A III-1, Tableau-A III-3 et Tableau-A III-6) ont joué à moins de quêtes ($\beta < 0$) et ont utilisé plus les avantages du jeu ($\beta > 0$). Leurs talents et cortex augmentent avec le nombre de visites qu'ils ont effectuées. Il est à noter aussi que quand la variable Langue diminuait (passant de 2 pour anglais à 1 pour français), le nombre de talents augmentait. Également, les filles et les plus jeunes joueurs ($\beta < 0$) ont atteint plus de niveaux (cortex) par rapport aux autres. Il apparaît aussi que les joueurs ayant participé au jeu entre 20 et 70 fois (*Voir ANNEXE III* : Tableau-A III-2, Tableau-A III-4 et Tableau-A III-7) ont accumulé plus de points en accédant à plus de quêtes ($\beta > 0$) et en utilisant moins d'avantages ($\beta < 0$) pour obtenir plus de cortex. De plus, ces joueurs ont gagné plus de neurones et de cortex en jouant en français ($\beta < 0$). D'autre part, les joueurs qui ont joué très souvent ($Nb_visite > 70$), ont participé à plus de quêtes et à moins d'avantages. Parmi eux, ceux qui ont joué en anglais ont obtenu plus de cortex (Tableau 4.10, Tableau 4.12 et *voir ANNEXE III* : Tableau-A III-5).

L'analyse des modèles correspondants à la variable *Durée_totale* nous apprend que les joueurs restaient plus longtemps dans l'environnement du jeu en participant aux quêtes (*Voir ANNEXE III*, Tableau-A III-9), en utilisant plus d'avantages (*Voir ANNEXE III*, Tableau-A III-10) et en le visitant plus souvent. De plus, il apparaît que les filles avaient tendance à jouer plus longtemps que les garçons en se fiant au coefficient de la variable *Sexe* ($\beta < 0$) dans le cas des joueurs visitant très souvent le jeu ($Nb_visite > 70$) (Tableau 4.13).

Les valeurs des coefficients de détermination (R^2) obtenus dans les modèles *Vs01Nr1* (0,559), *Vs70Nr* (0,744), *Vs70Cr* (0,517) et *Vs70Dt* (0,584) nous permettent de consolider l'analyse des résultats de ces modèles et leur interprétation.

4.3.5 Interprétation des résultats

Notre interprétation de la contribution des variables explicatives (indépendantes) aux modèles résultants de la régression linéaire multiple (présenté dans la section 4.4.3) est portée essentiellement sur les caractéristiques du joueur et son expérience du jeu qui pourraient expliquer de telles influences et de tels résultats en prenant en considération les spécificités de chaque sous-population, les critères de qualité des modèles obtenus et leurs coefficients estimés (valeur et signe).

En phase d'apprentissage et de découverte des possibilités offertes par le jeu, les nouveaux joueurs ou joueurs occasionnels (moins de 20 visites) participent à peu de quêtes (ou les mêmes quêtes à chaque visite) et utilisent les avantages gagnés. Il apparaît aussi que les joueurs ayant joué entre 20 et 70 fois et qui ont gagné le plus de neurones ont essentiellement joué en français.

En jouant fréquemment au jeu ($Nb_visite > 70$), nous remarquons que les joueurs obtiennent une expérience plus complète du jeu (plus de visites et de quêtes) et augmentent leurs scores et leurs niveaux. Dans ce cas, les filles persévèrent plus longtemps dans les sessions tandis que les sessions des garçons durent moins longtemps. De plus, ceux qui ont joué avec la version anglaise du jeu se démarquent avec le nombre de niveaux (cortex) atteints.

L'analyse des résultats a permis aussi de noter que l'âge et le sexe ne contribuaient pas aux scores (neurones, talents et cortex) pour les joueurs habitués (plus de 20 visites). Ceci nous porte à croire qu'une fois habitués au jeu, les filles comme les garçons, ayant le même âge ou non, performaient de manière semblable. Cela est peut-être expliqué par le fait que le contenu et la forme du jeu est adapté à l'âge et au sexe des joueurs, car : 1) le jeu est offert dans les deux langues; 2) donne plusieurs choix de quêtes selon différents thèmes scientifiques pouvant intéresser les filles comme les garçons à différents âges (entre 5 et 20 ans); 3) le choix de l'avatar (fille, garçon) et de son aspect extérieur (vêtements, accessoires) est fait par le joueur et 4) probablement la majorité des participants au jeu sont réellement dans la

tranche d'âge ciblé par le jeu même si ces participants remplissent la rubrique Date de naissance avec des valeurs différentes de la réalité. Également, nous avons constaté que pour les joueurs avec moins de 20 visites au jeu, l'âge et le sexe n'influençaient pas leur gain de neurones et de talents contrairement au gain de cortex où il apparaît que les filles et les plus jeunes joueurs obtenaient plus de cortex. Possiblement qu'aux premiers contacts avec l'environnement du jeu et ses différentes quêtes, les filles et les plus jeunes joueurs persévéraient plus et atteignaient davantage de niveaux (cortex) en comparaison avec les autres.

4.4 Modélisation : Regroupement (*Clustering*)

4.4.1 Contexte de l'expérimentation

Nous avons appliqué la méthode K-moyennes (Kmeans) pour rechercher d'éventuels regroupements d'individus de notre population. Afin de choisir les variables avec lesquelles K-moyennes s'exécutera, nous nous sommes basés sur les résultats obtenus des modèles de régression linéaire multiple. En effet, nous avons sélectionné uniquement les variables indépendantes ayant obtenu des valeurs de *p-value* inférieures à 0,05 et ayant donc un effet sur nos variables dépendantes. Également, nous avons utilisé les sous-populations réparties selon le nombre de visites expliquées plus haut (Vs01, Vs20 et Vs70).

La fonction K-moyennes de Matlab offre plusieurs options aux utilisateurs pour paramétrer l'application de l'algorithme K-moyennes (MathWorks-4, 2014). Dans notre expérimentation, nous avons choisi les valeurs suivantes pour nos paramètres :

- distance : distance euclidienne au carré (valeur par défaut),
- nombre de groupes: 3,
- groupe vide: signalement d'une erreur en cas de groupe vide,
- nombre de répétitions de l'algorithme de regroupement en faisant varier les centres initiaux : 5,
- choix des k individus comme centres initiaux : aléatoire.

En spécifiant la distance euclidienne au carré comme distance permettant de comparer les individus entre eux, Matlab établit le centroïde comme étant la moyenne des valeurs des attributs. Aussi, nous avons choisi parmi les indices de validité interne l'indice Silhouette pour évaluer les résultats de K-moyennes (regroupements obtenus). Tandis que les indices de validité externe ne peuvent pas être utilisés dans ce cas, car ces indices se basent sur des classifications de référence qui ne sont pas disponibles d'une part. D'autre part, nous n'avons pas l'expertise de réaliser de telles classifications.

4.4.2 Résultats de l'application du regroupement avec K-moyennes

Nous avons appliqué la méthode K-moyennes sur les sous-populations Vs01, Vs20 et Vs70 en prenant en considération une variable dépendante et deux variables indépendantes influentes sur celle-ci à chaque expérimentation. Ensuite, nous avons étudié les regroupements formés et les valeurs de Silhouette correspondantes. Dans cette section, nous présentons les résultats les plus interprétables et les plus performants en termes de valeurs de Silhouette alors que le reste des résultats est présenté dans ANNEXE VI (p. 107). Chaque figure présentée dans cette section comporte le graphique de partitionnement résultant de K-moyennes ainsi que le graphique représentant les valeurs du critère Silhouette.

4.4.2.1 Résultats du regroupement pour le nombre de neurones

Après l'application de K-moyennes sur la variable Nb_neurones dans la sous-population Vs01, nous avons remarqué la formation d'un grand groupe englobant presque la totalité des individus (*Cluster 2*). Ce groupe représente les joueurs avec peu de quêtes complétées, peu d'avantages utilisés (Figure 4.2), ayant divers intérêts scientifiques (*Voir ANNEXE VI: Figure-A VI-1 et Figure-A VI-2*) et n'ayant pas gagné un grand nombre de neurones.

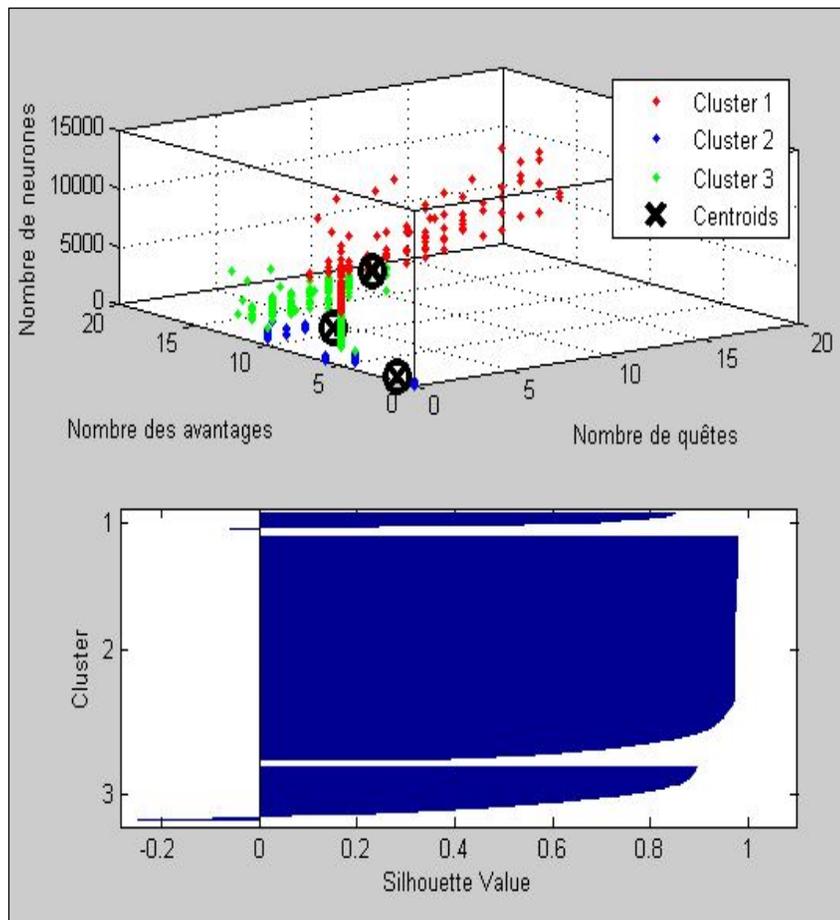


Figure 4.2 Neurones, quêtes et avantages dans Vs01

De manière générale, nous avons obtenu deux principaux regroupements (*Clusters* 1 et 3) (Figure 4.3 et voir ANNEXE VI : Figure-A VI-3 et Figure-A VI-4) qui reflètent des individus ayant divers intérêts scientifiques, et ayant participé au jeu entre 20 et 40 fois. Ces regroupements se différencient par le nombre de neurones gagnés qui est possiblement dû au nombre de quêtes complétées (moins de 5 quêtes pour l'un et plus de 5 quêtes pour l'autre).

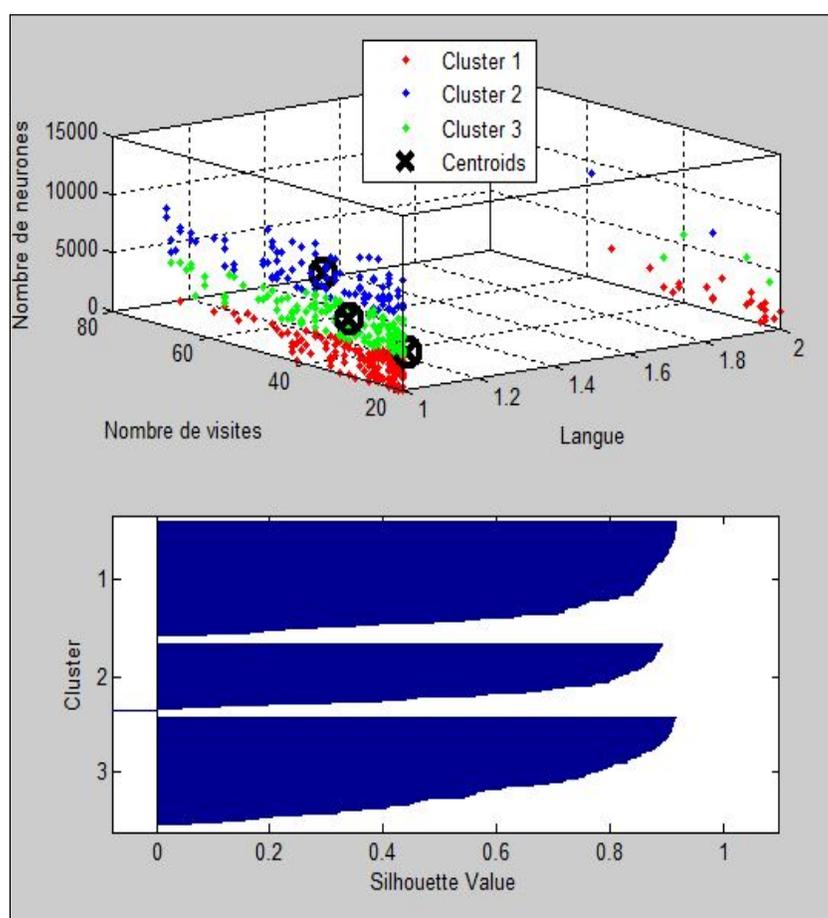


Figure 4.3 Neurones, langue et visites dans Vs20

Les résultats obtenus (Figure 4.4 et voir ANNEXE VI Figure-A VI-6) pour la sous-population de joueurs visitant le plus souvent le jeu ou ayant participé sur une plus longue période ($Nb_visite > 70$), montrent essentiellement deux groupes (*Clusters* 2 et 3). Ces groupes se distinguent entre eux par le nombre de visites, le nombre de quêtes et le nombre d'avantages et par conséquent le nombre de neurones gagnés.

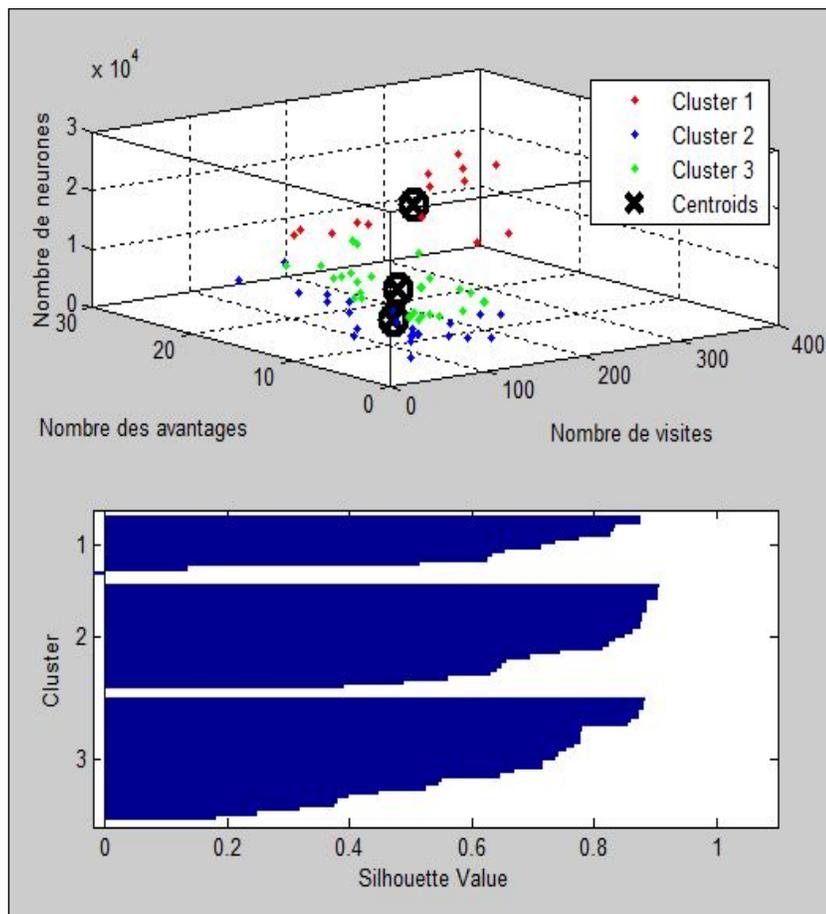


Figure 4.4 Neurones, visites et avantages dans Vs70

4.4.2.2 Résultats du regroupement pour le nombre de talents

L'application de l'algorithme de regroupement sur la sous-population Vs01 ($Nb_visite < 20$) a donné un grand groupe englobant presque la totalité des individus (*Cluster* 1). Ce groupe

représente les joueurs ayant divers intérêts scientifiques (*Voir* ANNEXE VI, Figure-A VI-8), jouant majoritairement en langue française (*Voir* ANNEXE VI, Figure-A VI-7), ayant peu de talents ($Nb_talents < 5000$), et dont le nombre de quêtes et le nombre des avantages sont globalement inférieurs à 10 (Figure 4.5).

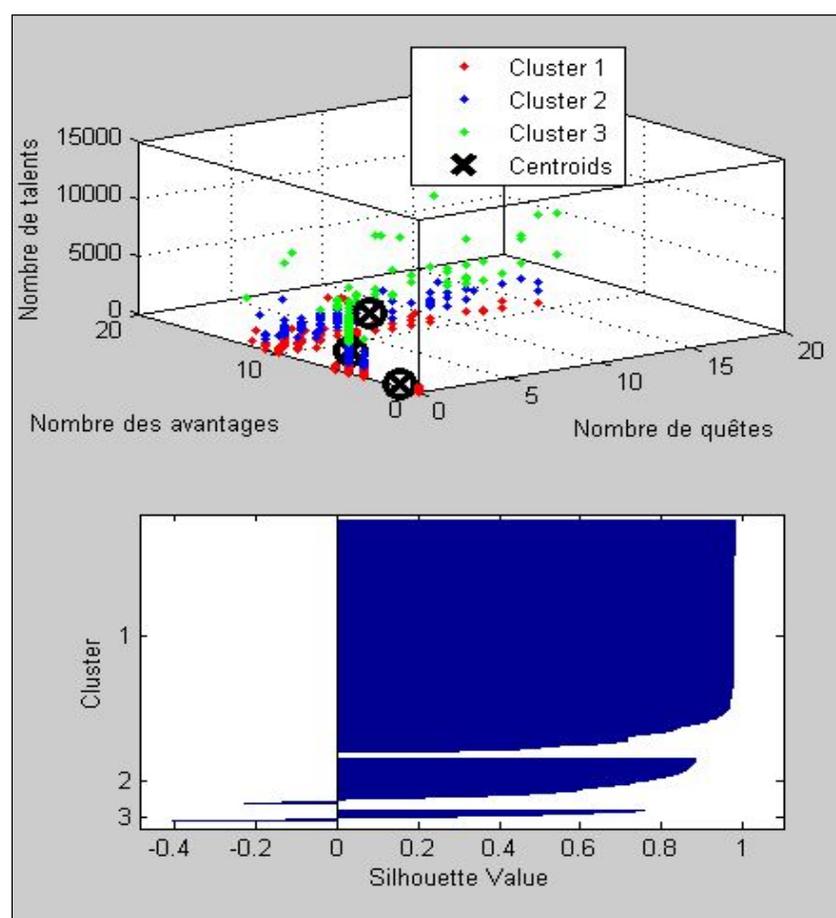


Figure 4.5 Talents, quêtes et avantages dansVs01

4.4.2.3 Résultats du regroupement pour le nombre de cortex

En appliquant le regroupement avec les variables Nb_visite et $Nb_avantages$ (Figure 4.6) nous avons obtenu un groupe principal (*Cluster* 3) caractérisant les joueurs (possiblement nouveaux) avec peu de cortex (moins de 5 cortex), peu d'avantages (moins de 5 avantages) et

moins de 10 visites. Ensuite, son application avec les variables Sexe, Nb_visite (Voir ANNEXE VI, Figure-A VI-13), a donné un groupe regroupant la majorité des individus de la sous-population (*Cluster 3*) et caractérisant les joueurs (fille et garçon), ayant moins de 6 visites et au plus 9 cortex.

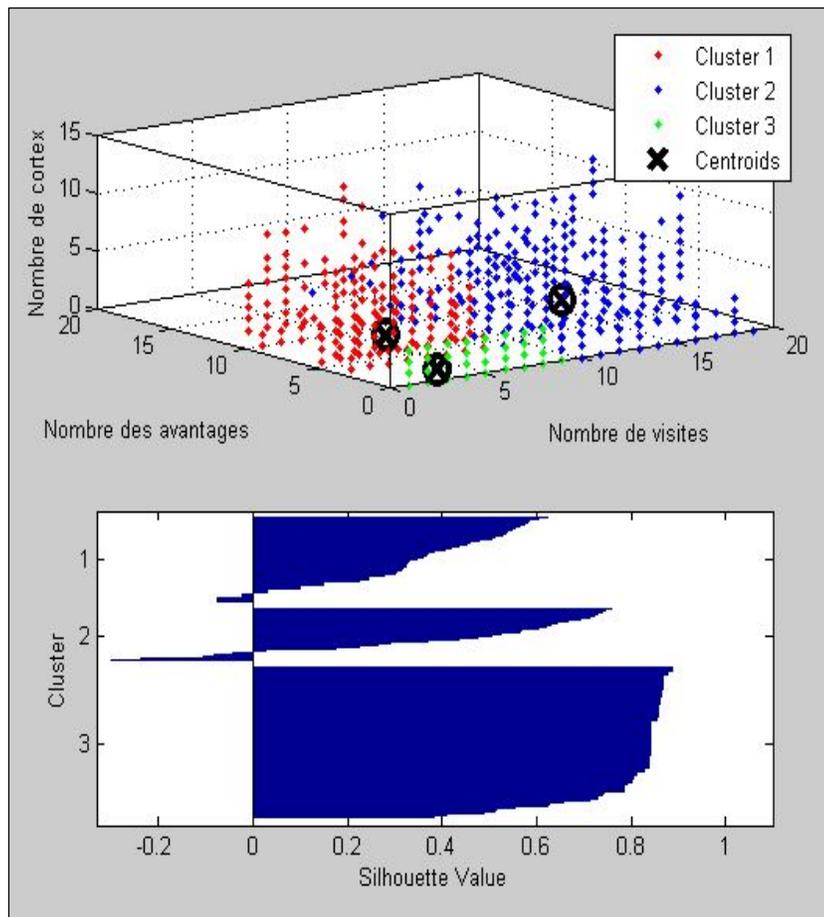


Figure 4.6 Cortex, visites et avantages dansVs01

En testant le regroupement avec les variables Intérêt scientifique et Langue (respectivement Intérêt scientifique et Nb_visite, Intérêt scientifique et Nb_quêtes) dans la sous-population Vs20, nous avons obtenu une répartition des joueurs selon leurs intérêts scientifiques (*Voir* ANNEXE VI : Figure-A VI-14, Figure-A VI-16 et Figure-A VI-17). Toutefois, nous notons que la majorité des individus ont joué en français, ont visité plus de 30 fois le jeu et ont obtenu au plus 10 cortex (*Voir* ANNEXE VI, Figure-A VI-15). Également, le regroupement obtenu en testant avec les variables Nb_visite et Nb_quêtes (respectivement Nb_visite et Nb_avantages) est caractérisé par moins de 10 quêtes complétées, une quarantaine de visites (au plus 20 avantages utilisés) et au plus 5 cortex gagnés (Figure 4.7) (respectivement *voir* ANNEXE VI, Figure-A VI-18).

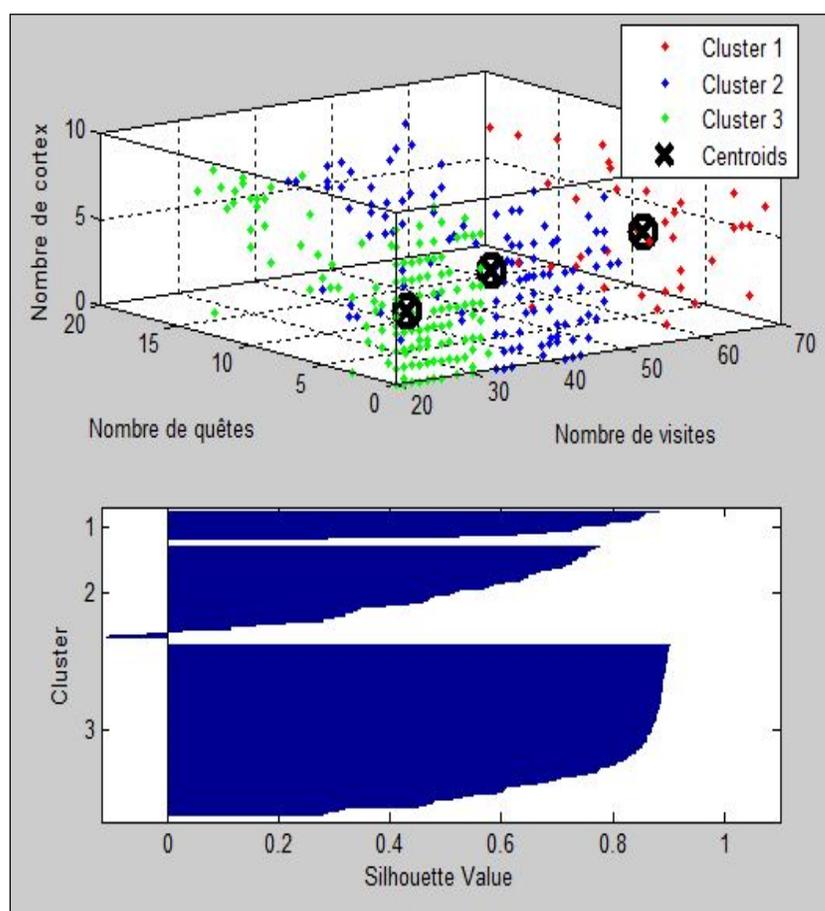


Figure 4.7 Cortex, visites et quêtes dans Vs20

Dans le cas des joueurs qui ont participé au jeu plus de 70 fois, le regroupement avec les variables Nb_quêtes et Nb_avantages a donné deux regroupements principaux (Figure 4.8). Le premier regroupement (*Cluster 1*) représente les joueurs ayant joué sur au plus 5 quêtes et ayant obtenu moins de 10 avantages et au plus 10 cortex. Le second regroupement (*Cluster 2*) représente les joueurs ayant moins de 5 cortex, jouant sur au plus 20 quêtes et utilisant plus de 10 avantages.

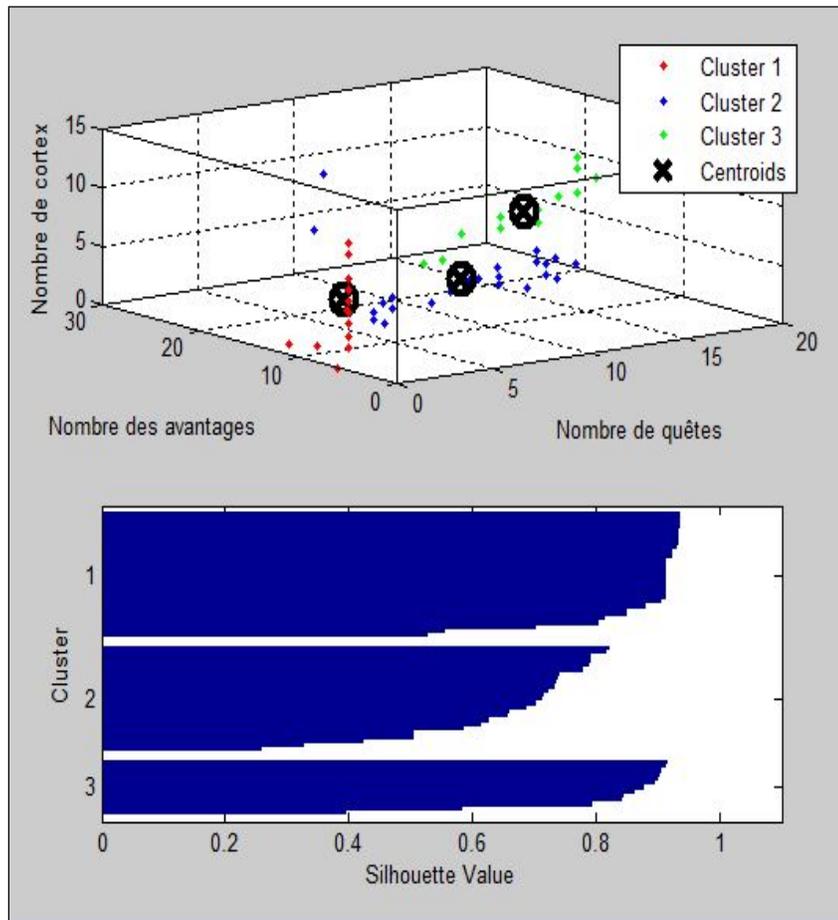


Figure 4.8 Cortex, quêtes et avantages dansVs70

4.4.2.4 Résultats du regroupement pour la durée totale des sessions

Le groupe englobant la majorité des joueurs (*Cluster 3*) représente ceux qui ont visité le jeu moins de 5 fois et qui ont participé au plus à une douzaine de quêtes (Figure 4.9). Par contre, pour la sous-population Vs20, le groupe englobant le plus d'individus (*Cluster 2*) représente les joueurs qui ont participé à peu de quêtes (moins de 5 quêtes), ayant utilisé au plus 5 à 6 avantages et ayant passé moins de 24 heures à jouer (*Voir ANNEXE VI Figure-A VI-20*).

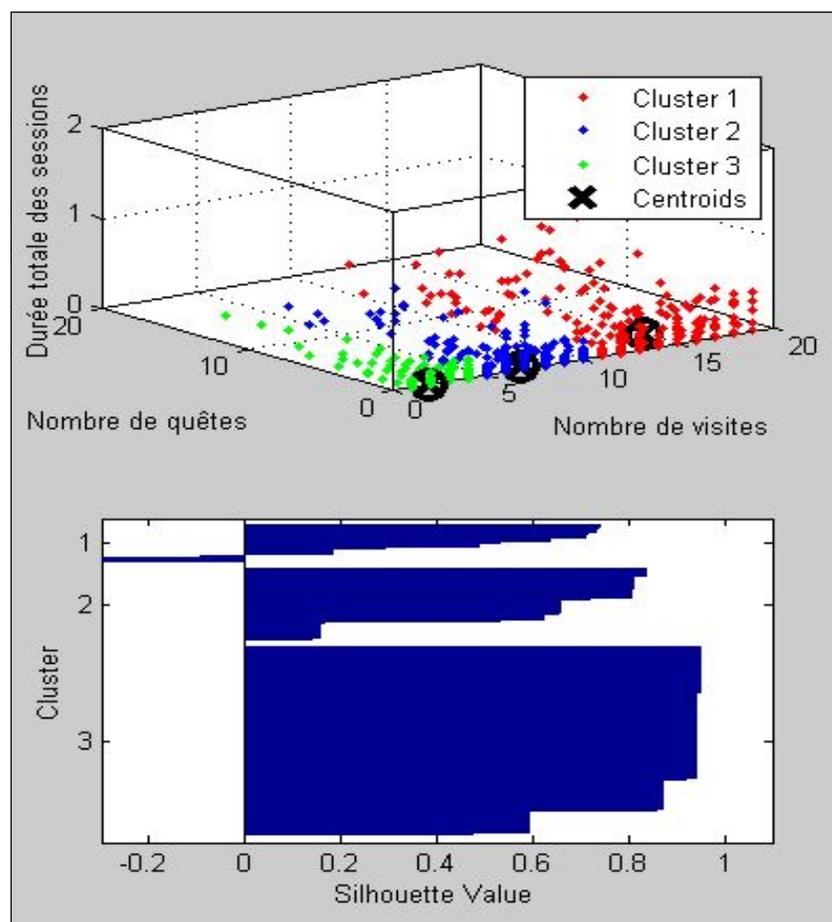


Figure 4.9 Durée totale, visites et quêtes dans Vs01

Pour la sous-population Vs70, nous avons noté la présence d'un grand groupe (*Cluster 3*) englobant presque la totalité des individus (Figure 4.10). Ce groupe représente les joueurs (filles et garçons), ayant visité moins de 200 visites, et dont la durée totale des sessions de jeu est inférieure à 48 heures. Aussi, nous avons remarqué que les filles sont plus nombreuses dans les deux premiers groupes et ont des durées de sessions plus longues (*Voir ANNEXE VI, Figure-A VI-21*) en comparant avec les garçons.

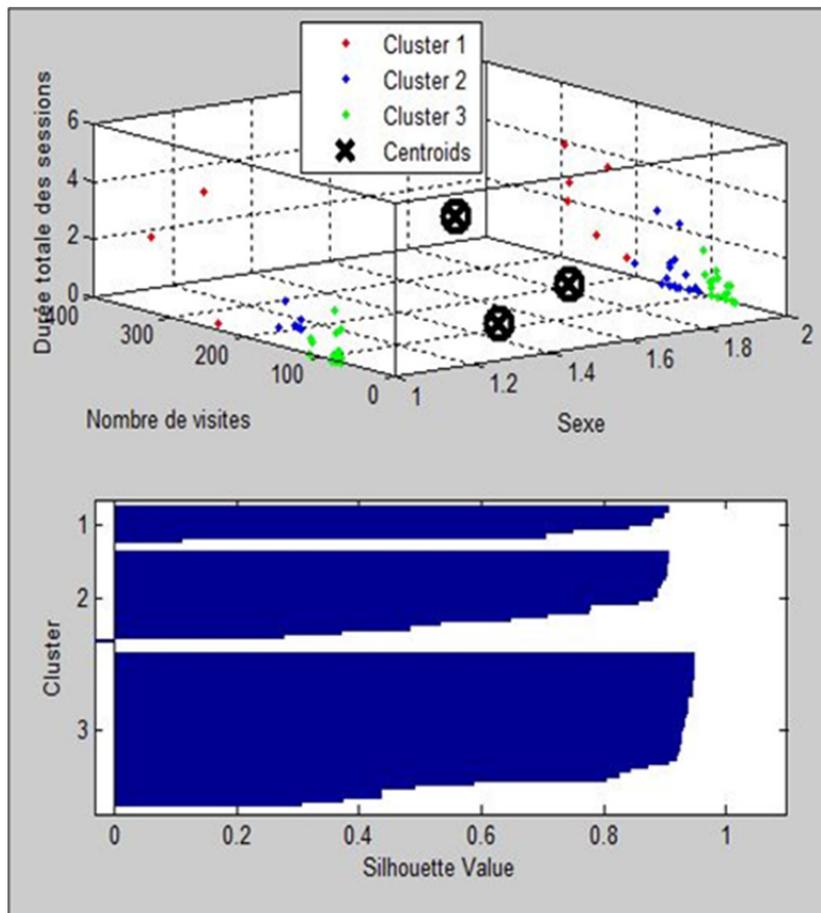


Figure 4.10 Durée totale, sexe et visites dans Vs70

4.4.3 Analyse des résultats

L'analyse des résultats obtenus avec la méthode K-moyennes (*Kmeans*) pour les variables représentant les scores des joueurs (neurones, talents et cortex) et dans chaque sous-population, nous a permis de constater qu'au début de l'apprentissage du jeu ($Nb_visite < 20$), la majorité des joueurs (formant un grand groupe) avait peu de quêtes complétées, peu de neurones et de talents accumulés et par conséquent peu d'avantages utilisés et de niveaux atteints. Dans ce cas, il n'y avait pas de distinction entre le jeu et les performances des garçons et des filles. Après cette phase d'apprentissage (Nb_visite entre 20 et 70), nous avons remarqué que les joueurs se différenciaient par le nombre de quêtes qu'ils avaient réalisé (formant deux groupes). Le premier groupe représentait des joueurs qui choisissaient de participer à peu de quêtes (ou les mêmes quêtes) à chaque fois selon leurs intérêts, leurs préférences ou la facilité du jeu. Tandis que le deuxième groupe représentait des joueurs qui avaient tendance à jouer à plus de quêtes et à accumuler plus de gains. Également, nous avons relevé qu'une fois habitués au jeu ($Nb_visite > 70$), les joueurs choisissaient soit de participer à moins de quêtes, utiliser moins d'avantages et gagner plus de niveaux, soit au contraire privilégier l'accès aux quêtes et aux avantages et obtenir moins de niveaux.

Concernant la durée totale des sessions, nous avons noté que pendant l'apprentissage du jeu, le groupe principal était formé de nouveaux joueurs probablement motivés, car ils ont exploré plusieurs quêtes en peu de visites et peu de temps. Mais avec une durée totale inférieure à 24 heures, la majorité des joueurs (1 groupe principal) n'avaient pas une grande participation au jeu (peu de quêtes et peu d'avantages) malgré un nombre de visites compris entre 20 et 70. Il s'agirait possiblement de joueurs qui participaient au jeu occasionnellement. Au-delà de 70 visites, les joueurs se distinguaient par leurs nombres de visites et leurs nombres d'avantages utilisés.

Nous avons noté aussi que la majorité des participants au jeu jouent en français et leurs intérêts scientifiques sont variés.

4.4.4 Interprétation des résultats

En analysant les regroupements formés dans les différentes sous-populations et testés sur plusieurs variables, nous avons déduit trois principaux aspects de participation des joueurs :

- Joueurs débutants: Avec peu de visites à leur effectif, ces joueurs participent aux quêtes, accumulent les scores et utilisent moins les avantages. Dans ce groupe, nous ne distinguons pas des différences de cheminement dans le jeu entre les filles et les garçons ni entre les joueurs de différents âges.
- Joueurs intermédiaires: Ceux sont des joueurs qui participent souvent aux quêtes du jeu et gagnent plus de points afin d'atteindre des niveaux supérieurs. Certains choisissent de participer à peu de quêtes ou possiblement aux mêmes quêtes selon leurs préférences ou l'accessibilité (facilité) du jeu. D'autres joueurs choisissent de découvrir plus de quêtes, y jouer et augmenter leurs scores.
- Joueurs avancés (habitués): Ceux sont des joueurs habitués au jeu, ils le visitent souvent et selon les cas choisissent d'accéder à plus de quêtes et d'utiliser plus d'avantages ou au contraire jouent à moins de quêtes et d'avantages en tentant de gagner le maximum de niveaux (cortex). Parmi ces joueurs, les filles ont tendance à jouer plus longtemps.

Les résultats du regroupement, semblent mener à des conclusions logiques et prévisibles. Néanmoins, ces conclusions permettent de consolider notre approche qui se base sur le partitionnement des individus en fonction des attributs (variables) choisis selon leurs effets significatifs sur les scores et la durée des sessions de jeu (déduits des modèles de régression linéaire multiple). En effet, cette approche nous a permis de retrouver des regroupements distincts de joueurs que nous avons expliqués en termes de comportement ou actions réalisées dans le jeu dans la perspective d'obtenir un premier aperçu sur les profils des joueurs.

4.5 Conclusion

Après avoir étudié la BD relative au jeu et préparé les données pertinentes aux objectifs de notre étude (transformation, élimination des valeurs aberrantes ou extrêmes), nous avons étudié la corrélation entre ces données. Cela nous a permis de sélectionner les variables (caractéristiques et scores du joueur, durée de ses sessions de jeu) pour l'étape de modélisation dans laquelle les techniques de régression linéaire multiple et de regroupement ont été appliquées.

Les résultats du modèle de régression multiple ont montré principalement l'influence du nombre d'accès au jeu (nombre de visites), des quêtes visitées et des avantages utilisés sur les performances du joueur (scores) et sur la durée passée à jouer. Toutefois, l'âge et le sexe des joueurs ne semblaient pas différencier leur expérience de jeu sauf dans le cas de nouveaux joueurs où les filles et les plus jeunes joueurs atteignaient davantage de niveaux; et dans le cas des joueurs habitués au jeu où les filles jouaient plus longtemps que les garçons.

D'autre part, les résultats de la méthode de regroupement ont révélé trois formes de participation des joueurs : (1) les joueurs débutants qui explorent les quêtes et accumulent les points; (2) les joueurs intermédiaires, qui selon leur préférences, choisissent de participer à plus ou moins de quêtes et gagner plus de niveaux; (3) les joueurs avancés qui tentent de gagner le maximum de niveaux, parmi eux, les filles ont tendance à jouer plus longtemps.

CONCLUSION

Grâce à leur composante ludique, les jeux sérieux offrent un support d'entraînement et de formation motivant et attrayant. Leur adaptabilité est un critère important de leur succès et de leur efficacité. Elle permet de maintenir l'intérêt du joueur, individualiser les sessions d'entraînement ainsi qu'offrir un support d'aide efficace. Cette adaptation se base essentiellement sur le suivi et l'évaluation de la progression du joueur (apprenant/patient) et l'analyse de ses caractéristiques et ses performances. Ces données sur les joueurs et leurs parcours sont très nombreuses et très variées, elles posent des problématiques concernant les approches à adopter, les techniques d'analyse à employer, et surtout l'extraction de connaissances interprétables et utilisables. Notre principale contribution s'inscrit dans ce contexte et tente d'apporter des éléments de réponses à ces problématiques à travers l'étude d'un jeu sérieux en particulier et sa BD, à savoir : « Science en jeu ». Cette étude contribue aussi à une meilleure compréhension du cheminement des apprenants/patients dans un environnement virtuel, qui permettrait l'amélioration des jeux sérieux par le biais de leur personnalisation (utilité industrielle).

L'objectif de notre travail est d'identifier les données pertinentes et les méthodes de forage de données appropriées pour les analyser et la déduction des caractéristiques des profils des joueurs. Notre approche est exploratoire, car d'une part la BD de « Science en jeu » n'a pas été étudiée auparavant, et d'autre part nous n'avons pas l'expertise pour définir les profils des joueurs à priori.

Pour cela, nous avons défini, en premier lieu, les JS et nous avons passé en revue des études de recherche sur les JS dans les domaines de l'éducation et de la stimulation cognitive ainsi que des études traitant le sujet du forage de données des jeux vidéo et des jeux sérieux. Cette revue nous a aidés dans l'élaboration de notre scénario d'analyse, dans la première sélection de l'ensemble des données à étudier (caractéristiques du joueur et de son parcours dans le jeu, scores et durée) et dans le choix des techniques d'analyse (modèle de régression linéaire multiple et regroupement). Ensuite, nous avons étudié la BD relative au jeu et nous avons

préparé les données pour la modélisation. Pour chaque modèle testé, nous avons présenté les résultats, suivis d'une analyse et d'une interprétation.

Les résultats du modèle de régression linéaire multiple montrent principalement les effets significatifs du nombre d'accès au jeu (nombre de visites), des quêtes visitées et des avantages utilisés sur les performances du joueur (scores) et sur la durée passée dans le jeu. Tandis que l'âge et le sexe des joueurs ne semblent pas différencier leur expérience de jeu sauf dans le cas de nouveaux joueurs où les filles et les plus jeunes joueurs atteignent davantage de niveaux; et dans le cas des joueurs habitués au jeu où les filles jouent plus longtemps que les garçons.

Concernant le modèle de regroupement testé (K-moyennes), nous avons noté que le comportement des joueurs est assez similaire pendant les premiers accès au jeu. Mais, il a tendance à se différencier avec l'acquisition de plus d'expérience. Plus précisément, les résultats révèlent trois formes de participation des joueurs: (1) les « débutants » qui explorent les quêtes et accumulent les points; (2) les « intermédiaires », qui selon leur préférence, participent à plus (ou moins) de quêtes pour gagner plus de niveaux; (3) les « avancés » qui privilégient le gain du maximum de niveaux, parmi eux, les filles ont tendance à jouer plus longtemps.

Compte tenu des résultats obtenus, nous suggérons à la compagnie créatrice les améliorations suivantes afin de favoriser une meilleure expérience du jeu et offrir un outil efficace pour l'apprentissage et l'entraînement :

- adapter les sessions du jeu au cheminement du joueur, à son style de jeu et à son rythme d'apprentissage grâce à l'analyse de ses performances et de ses réponses. Il serait possible, par exemple, de proposer aux joueurs (des garçons selon les résultats), qui se déconnectent souvent du jeu, des mini-jeux variés et de courtes durées qui retiendraient leur attention et leur permettraient de persévérer dans les quêtes;

- intégrer dans le jeu une partie « suivi et évaluation » qui permettrait à des enseignants, des parents ou des thérapeutes de suivre l'évolution de l'apprentissage ou de l'entraînement;
- renforcer la dimension sociale en offrant des jeux permettant la coopération et/ou la compétition;
- rentabiliser la partie du jeu offerte en anglais en faisant connaître la plateforme du jeu dans le milieu scolaire anglophone (campagnes publicitaires, promotions, compétitions, etc.);
- fidéliser sa clientèle, par exemple en encourageant les joueurs à se procurer des abonnements,
- encourager l'utilisation du jeu dans le cadre d'études de recherche en établissant des relations de coopération avec le milieu universitaire, scolaire et de la santé.

Plus généralement, les résultats obtenus nous fournissent une vue globale du comportement des joueurs et une première lecture de leurs profils. Seulement, ces profils manquent de précision et ne permettent pas de : 1) connaître le type d'activité dans laquelle le joueur a performé; 2) interpréter les scores en termes de compétences acquises ou à renforcer et par conséquent la personnalisation du jeu se retrouve compromise et limitée. Ce constat nous conduit à proposer des pistes de recherche pour :

- explorer les données issues des tests d'évaluation des compétences tels que les résultats d'examens dans le domaine de l'éducation, ou des batteries de tests neuropsychologiques dans le domaine de la stimulation cognitive en vue d'enrichir le profil du joueur;
- valider les profils des joueurs trouvés avec des classifications de références fournis par des experts en éducation/neurosciences;
- expérimenter les méthodes de classification supervisée en vue de construire un modèle prédictif basé sur le profil du joueur et l'historique de ses performances;
- expérimenter le jeu dans le contexte de l'éducation/santé en assurant toutes les conditions requises et à plus long terme pour valider les modèles prédictifs;
- développer des outils d'évaluation fiable avec des critères mesurables et quantifiables pour les jeux sérieux;

- explorer la dimension sociale qu'offrent certains jeux qui permettent la coopération et/ou la compétition entre les joueurs et les nouvelles connaissances qui peuvent y découler;
- développer des architectures avec des composants intelligents (IA) pour l'adaptation des JS selon le profil du joueur et les objectifs fixés par les spécialistes (enseignants/thérapeutes).

RECOMMANDATIONS

A travers cette étude, nous avons constaté que des travaux de recherche et des contextes d'application ainsi que la richesse des connaissances produites dans le domaine des JS sont très diversifiés. Il apparaît aussi clairement le rôle prometteur que peuvent jouer ces outils en éducation et en santé. Toutefois, il est important de renforcer toutes ces connaissances en effectuant plus d'analyses afin de cerner les exigences et les perspectives, et diriger cette voie de recherche de manière efficace en coordonnant les efforts des équipes multidisciplinaire qui œuvrent dans ce domaine à travers le monde. Ces efforts devront se concentrer sur les aspects suivants :

- Développement de JS attrayants, immersifs et répondants aux besoins de la clientèle;
- Personnalisation des jeux pour accroître la convivialité et l'efficacité du jeu et offrir une meilleure expérience au joueur;
- Élaboration de stratégies d'enseignement ou de thérapie incluant les JS afin de permettre aux intervenants (enseignants ou thérapeutes) d'adopter ce nouveau type d'outils et de les utiliser efficacement;
- Intégration de l'évaluation de l'apprentissage ou thérapie dans les JS dans le but de suivre la progression du joueur et d'orienter son apprentissage ou thérapie;
- Expérimentation des JS plus étendue et à plus long terme pour prouver leur pertinence et leur efficacité de manière scientifique et rigoureuse;
- Utilisation des techniques de forage de données pour une analyse plus poussée et plus précise des données des JS et des joueurs.

ANNEXE I

CODIFICATION DE L'ATTRIBUT INTÉRÊT SCIENTIFIQUE

Tableau-A I-1 Codes pour l'attribut Intérêt scientifique

Intérêt scientifique	Code
Aéronautique	1
Architecture	2
Astronomie	3
Archéologie/Paléontologie/Histoire	4
Art/Fiction/Sport/Langue	5
Biologie/Biochimie	6
Botanique	7
Chimie/Physique (Sciences pures)	8
Écologie/Météorologie/Sciences de la nature	9
Génie/Électricité et électronique/Mécanique et Robotique/Sciences et technologie	10
Génétique	11
Géologie	12
Géographie	13
Informatique et multimédia/Jeux vidéo	14
Mathématiques	15
Médecine et Corps humain	16
Océanographie	17
Sciences humaines/Sciences juridiques	18
Sofad	19
Transports	20
Zoologie	21
Tous	22
Aucun	23

ANNEXE II

GRAPHIQUES DE DISTRIBUTION DES VALEURS DES ATTRIBUTS

- Distribution des attributs avant la suppression des valeurs extrêmes et la subdivision en sous-populations

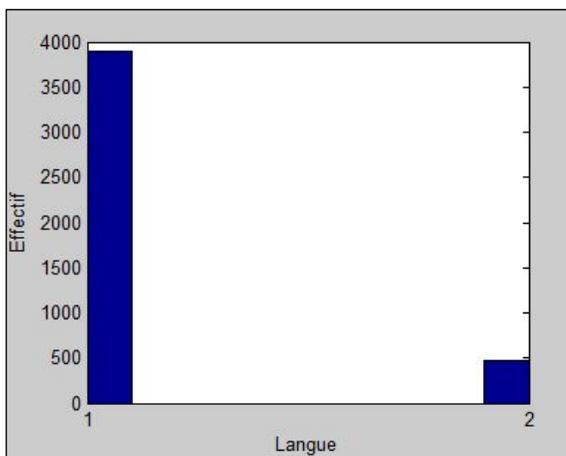


Figure-A II-1 Distribution de l'attribut langue

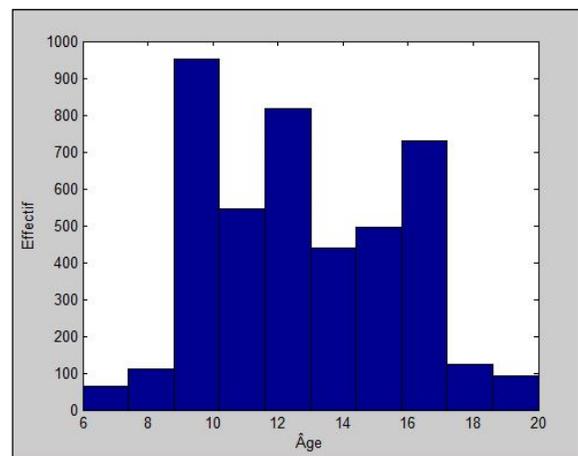


Figure-A II-2 Distribution de l'attribut âge

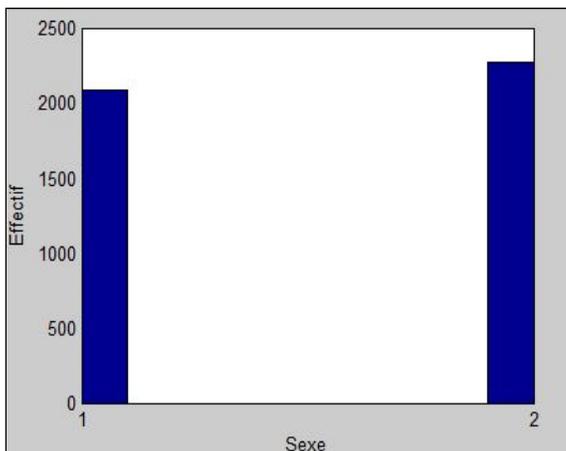


Figure-A II-3 Distribution de l'attribut sexe

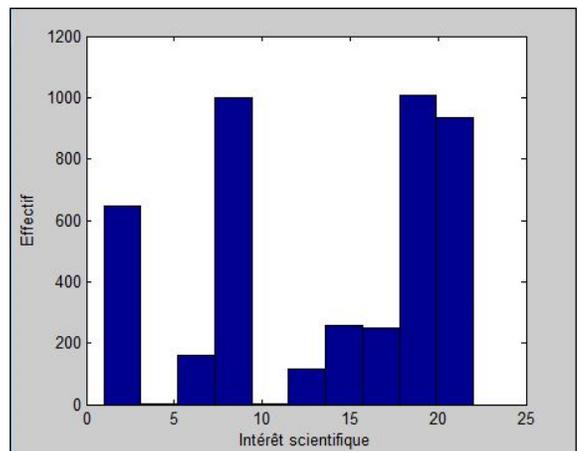


Figure-A II-4 Distribution de l'attribut intérêt scientifique



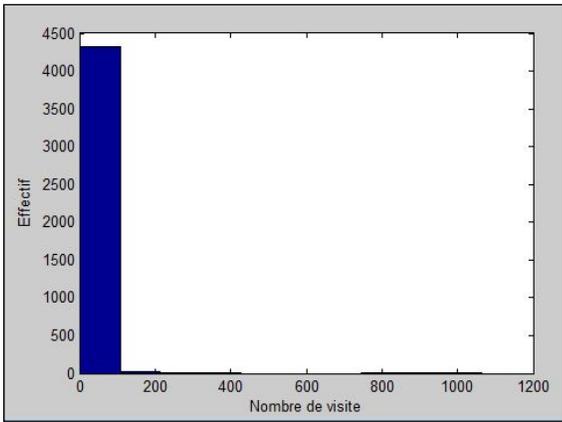


Figure-A II-5 Distribution de l'attribut nombre de visites (histogramme)

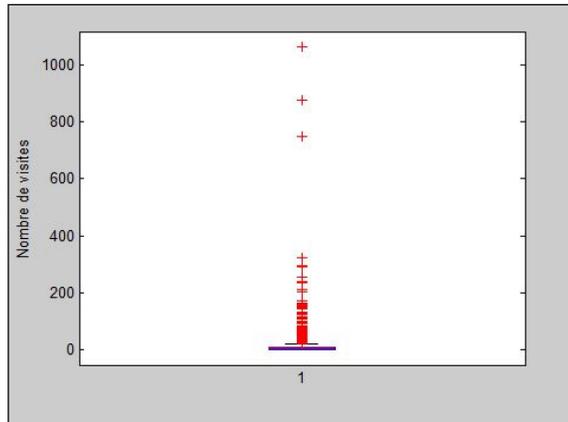


Figure-A II-6 Distribution de l'attribut nombre de visites (boite à moustaches)

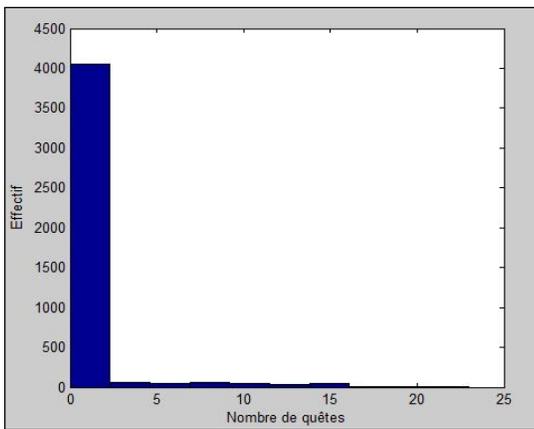


Figure-A II-7 Distribution de l'attribut nombre de quêtes (histogramme)

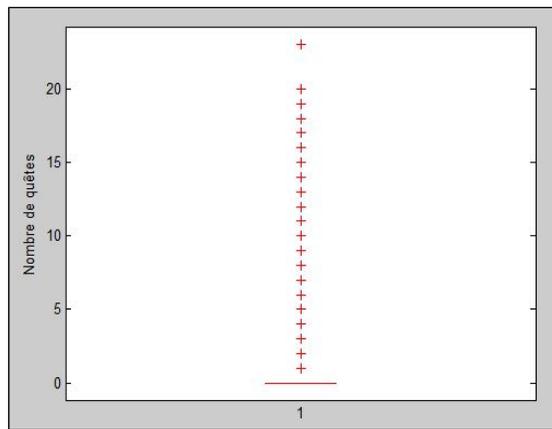


Figure-A II-8 Distribution de l'attribut nombre de quêtes (boite à moustaches)

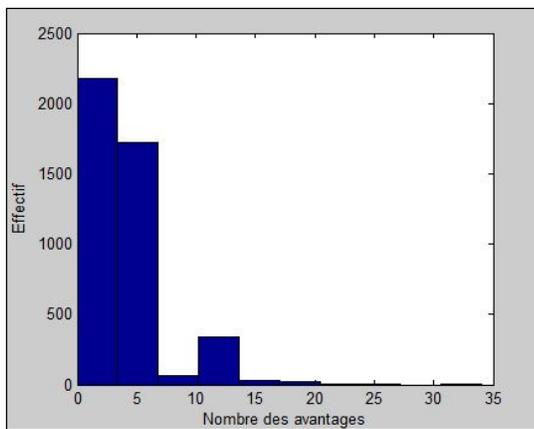


Figure-A II-9 Distribution de l'attribut nombre d'avantages (histogramme)

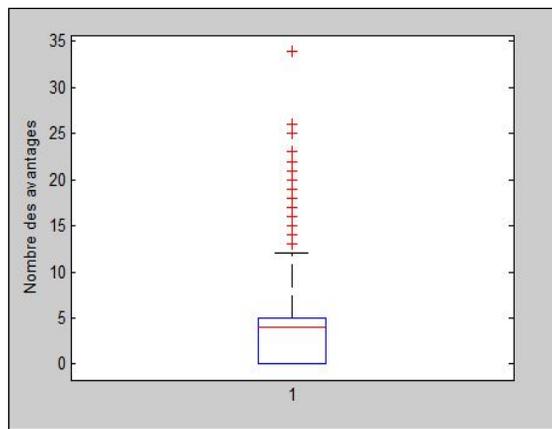


Figure-A II-10 Distribution de l'attribut nombre d'avantages (boite à moustaches)

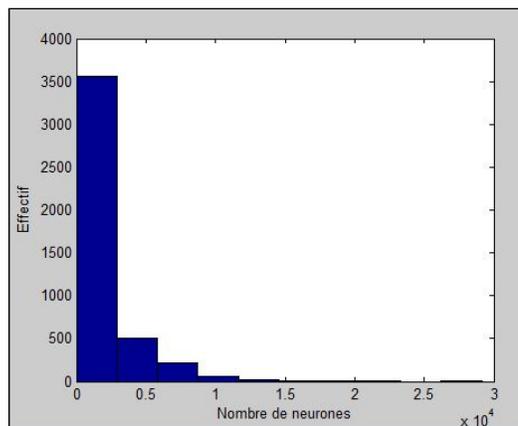


Figure-A II-11 Distribution de l'attribut nombre de neurones (histogramme)

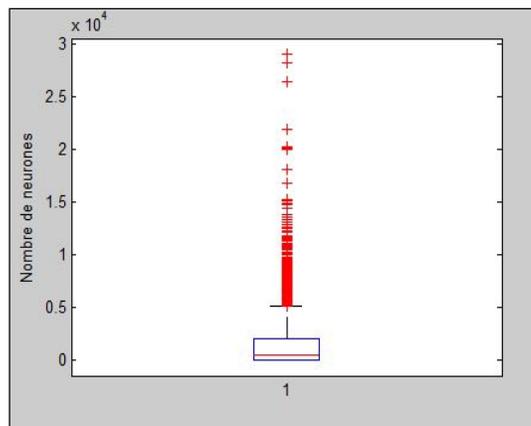


Figure-A II-12 Distribution de l'attribut nombre de neurones (boite à moustaches)

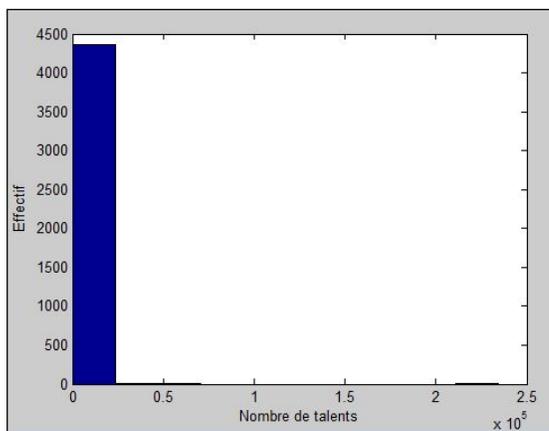


Figure-A II-13 Distribution de l'attribut nombre de talents (histogramme)

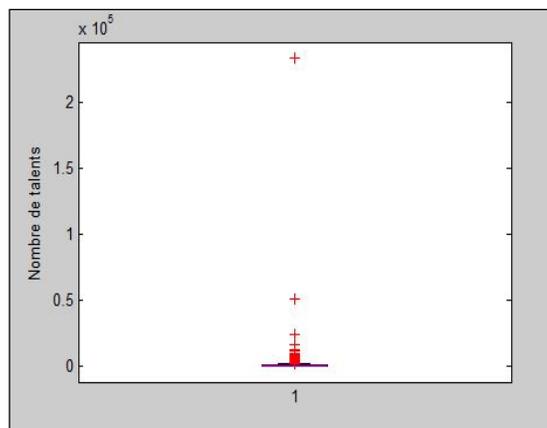


Figure-A II-14 Distribution de l'attribut nombre de talents (boite à moustaches)

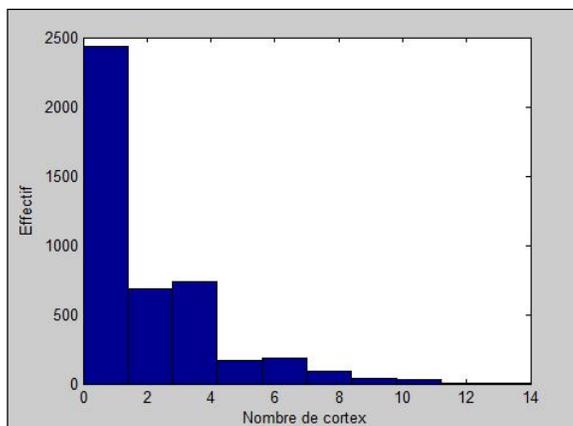


Figure-A II-15 Distribution de l'attribut nombre de cortex (histogramme)

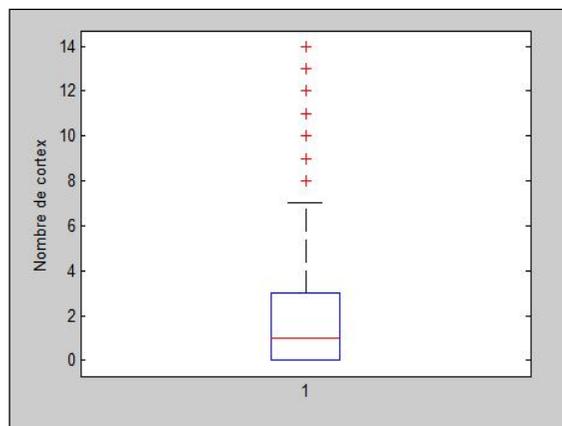


Figure-A II-16 Distribution de l'attribut nombre de cortex (boite à moustaches)

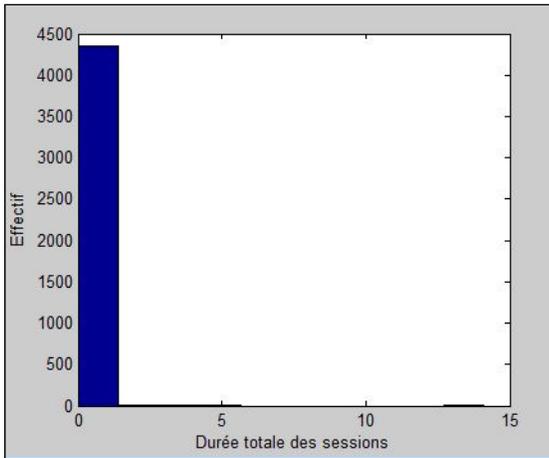


Figure-A II-17 Distribution de l'attribut durée totale des sessions (histogramme)

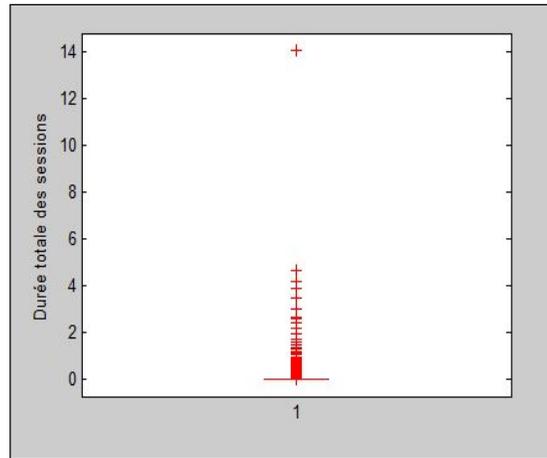


Figure-A II-18 Distribution de l'attribut durée totale des sessions (boite à moustaches)

- Distribution des attributs après la suppression des valeurs extrêmes et la subdivision en sous-populations

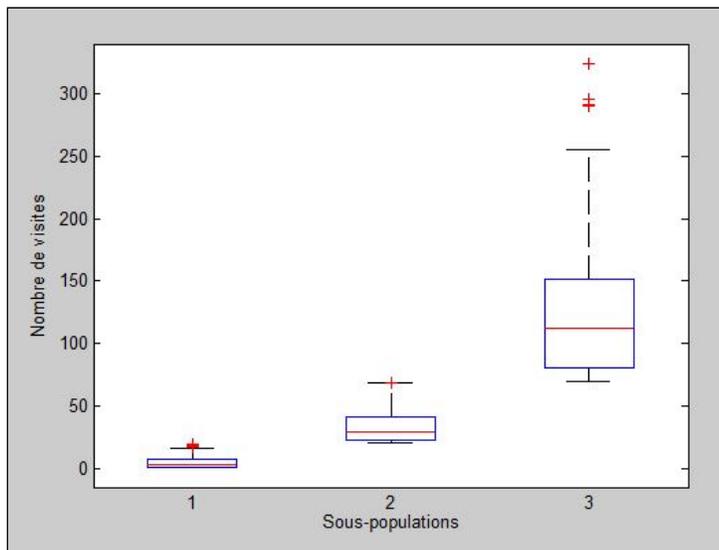


Figure-A II-19 Distribution de l'attribut nombre de visites dans les sous-populations Vs01, Vs20 et Vs70

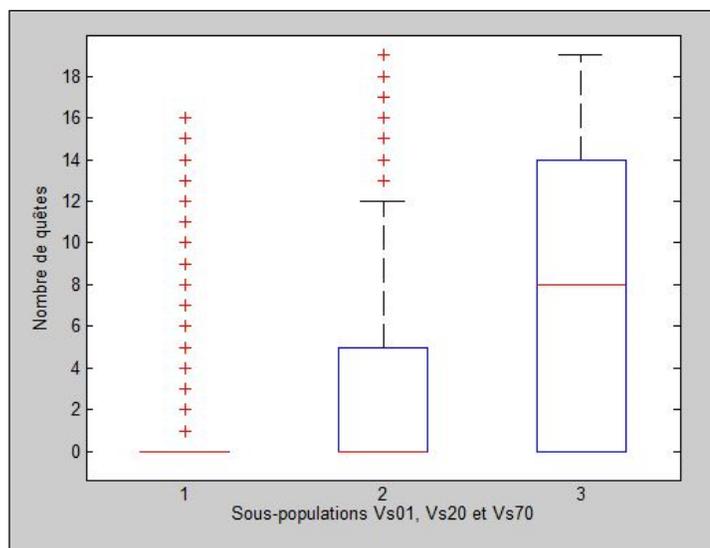


Figure-A II-20 Distribution de l'attribut nombre de quêtes dans les sous-populations Vs01, Vs20 et Vs70

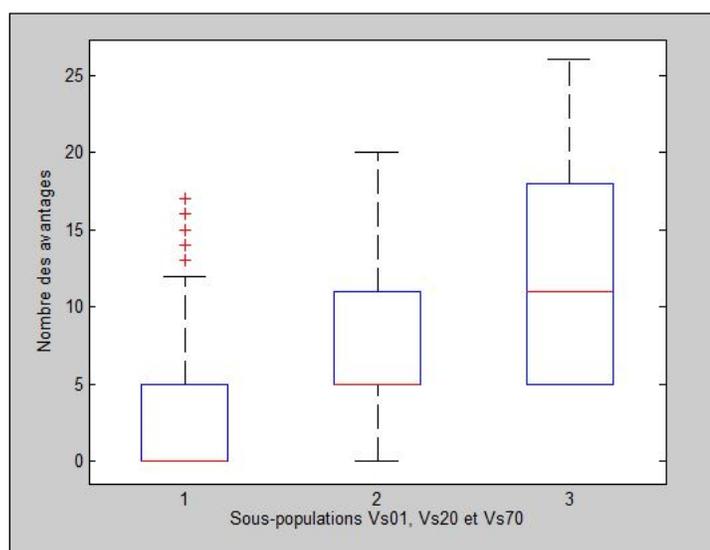


Figure-A II-21 Distribution de l'attribut nombre d'avantages dans les sous-populations Vs01, Vs20 et Vs70

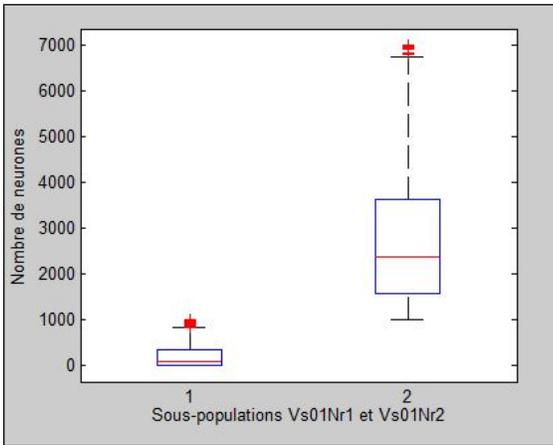


Figure-A II-22 Distribution de l'attribut nombre de neurones dans les sous-populations Vs01Nr1 et Vs01Nr2

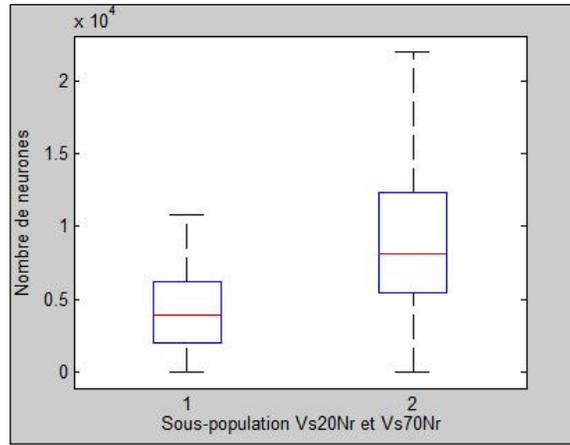


Figure-A II-23 Distribution de l'attribut nombre de neurones dans les sous-populations Vs20Nr et Vs70Nr

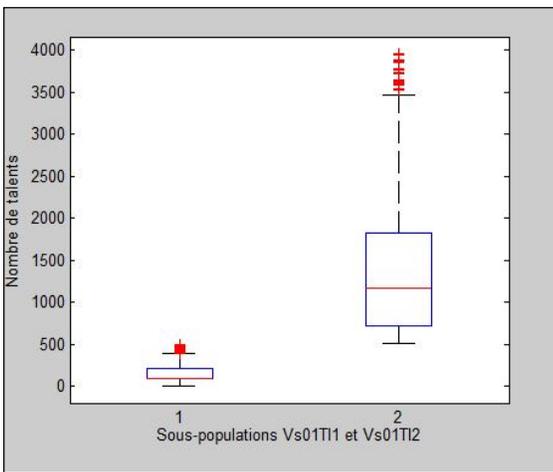


Figure-A II-24 Distribution de l'attribut nombre de talents dans les sous-populations Vs01TI1 et Vs01TI2

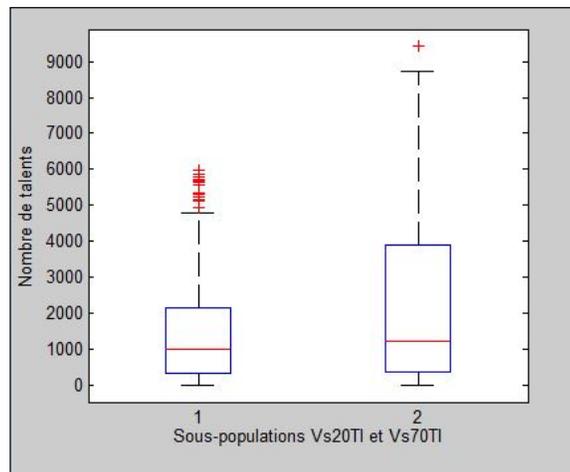


Figure-A II-25 Distribution de l'attribut nombre de talents dans les sous-populations Vs20TI et Vs70TI

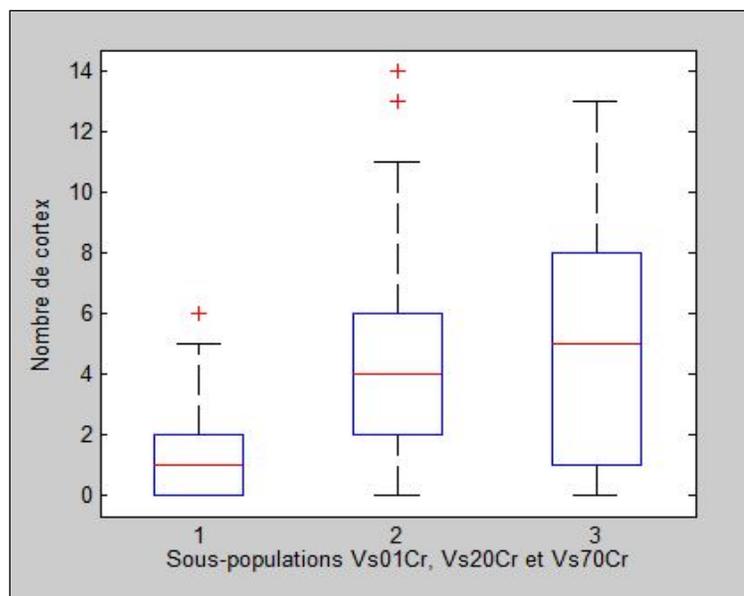


Figure-A II-26 Distribution de l'attribut nombre de cortex dans les sous-populations Vs01Cr, Vs20Cr et Vs70Cr

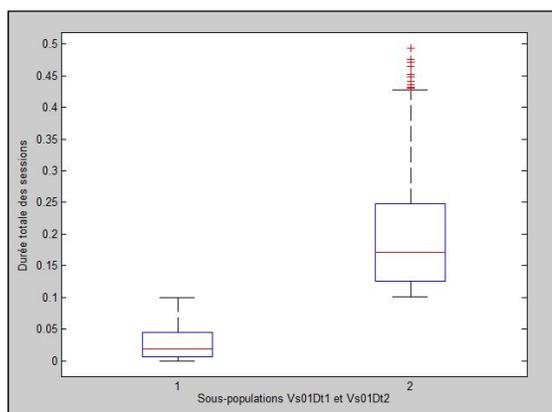


Figure-A II-27 Distribution de l'attribut durée totale des sessions dans les sous-populations Vs01Dt1 et Vs01Dt2

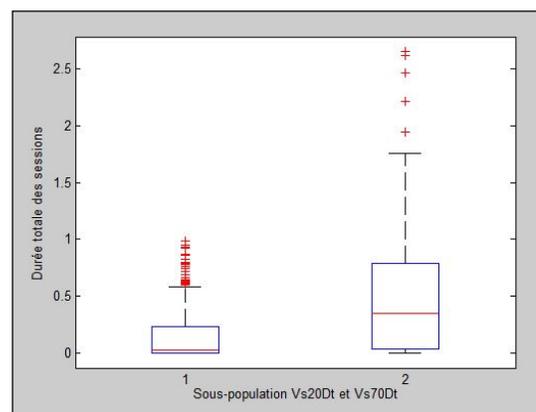


Figure-A II-28 Distribution de l'attribut durée totale des sessions dans les sous-populations Vs20Dt et Vs70Dt

ANNEXE III

RÉSULTATS DE L'APPLICATION DU MODÈLE DE RÉGRESSION LINÉAIRE MULTIPLE (SUITE)

- Variable Nb_neurones en fonction des variables indépendantes

Tableau-A III-1 MRLM avec Nb_neurones pour la sous-population Vs01Nr2
(Nombre d'observations de l'échantillon : 111)

Coefficients estimés du modèle								
Variables	Constante	Langue	Age	Sexe	Intérêt	visite	quête	avantage
Coefficients	1994,5	54,26	47,028	-86,95	-12,28	24,497	328,83	2,8436
p-value	0,1643	0,9503	0,3376	0,7654	0,5799	0,3827	0,0104	0,9715
Paramètres du modèle								
R²			F-stat			p-value		
0,143			2,45			0,0229		

Tableau-A III-2 MRLM avec Nb_neurones pour la sous-population Vs20Nr
(Nombre d'observations: 413)

Coefficients estimés du modèle								
Variables	Constante	Langue	Age	Sexe	Intérêt	visite	quête	avantage
Coefficients	4868,6	-1951	-5,6796	85,529	-59,09	36,583	321,75	18,3
p-value	< 0,000	<0,000	0,8891	0,66878	<0,000	<0,000	<0,000	0,6518
Paramètres du modèle								
R²			F-stat			p-value		
0,454			48,2			< 0,000		

- Variable Nb_talents en fonction des variables indépendantes

Tableau-A III-3 MRLM avec Nb_talents pour la sous-population Vs01T12
(Nombre d'observations de l'échantillon 113)

Coefficients estimés du modèle								
Variables	Constante	Langue	Age	Sexe	Intérêt	visite	quête	avantage
Coefficients	1466,7	-97,675	-37,42	-80,77	-21,29	31,256	-70,97	131,18
p-value	0,0146	0,7156	0,1229	0,5825	0,0571	0,0456	0,2043	0,0064
Paramètres du modèle								
R²			F-stat			p-value		
0,202			3,79			0,001		

Tableau-A III-4 MRLM avec Nb_talents pour la sous-population Vs20T1
(Nombre d'observations: 419)

Coefficients estimés du modèle								
Variables	Constante	Langue	Age	Sexe	Intérêt	visite	quête	avantage
Coefficients	1255,2	-483,28	17,769	111,8	-0,195	6,7807	59,592	-10,43
p-value	0,0237	0,0848	0,5293	0,4201	0,9849	0,2414	0,0118	0,7095
Paramètres du modèle								
R²			F-stat			p-value		
0,0526			3,26			0,0022		

Tableau-A III-5 MRLM avec Nb_talents pour la sous-population Vs70T1
(Nombre d'observations: 65)

Coefficients estimés du modèle								
Variables	Constante	Langue	Age	Sexe	Intérêt	visite	quête	avantage
Coefficients	-598,26	1239,2	56,859	319,22	-109,6	7,3888	-137,3	172,99
p-value	0,8037	0,4232	0,5773	0,6299	0,0315	0,2182	0,1145	0,0802
Paramètres du modèle								
R²			F-stat			p-value		
0,23			2,44			0,0296		

- Variable Nb_cortex en fonction des variables indépendantes

Tableau-A III-6 MRLM avec Nb_cortex pour la sous-population Vs01Cr
(Nombre d'observations de l'échantillon: 376)

Coefficients estimés du modèle								
Variables	Constante	Langue	Age	Sexe	Intérêt	visite	quête	avantage
Coefficients	3,0298	-0,1418	-0,073	-0,559	-0,002	0,0422	0,0847	0,0734
p-value	< 0,000	0,5858	0,0081	0,0002	0,8946	0,0216	0,2071	0,0219
Paramètres du modèle								
R²			F-stat			p-value		
0,153			9,46			< 0,000		

Tableau-A III-7 MRLM avec Nb_cortex pour la sous-population Vs20Cr
(Nombre d'observations: 432)

Coefficients estimés du modèle								
Variables	Constante	Langue	Age	Sexe	Intérêt	visite	quête	avantage
coefficients	7,3236	-1,4929	0,004	-0,293	-0,049	0,0356	0,402	-0,401
p-value	< 0,000	0,0034	0,932	0,245	0,0082	0,0007	<0,000	<0,000
Paramètres du modèle								
R²			F-stat			p-value		
0,245			19,7			< 0,000		

- Variable Duree_totale en fonction des variables indépendantes

Tableau-A III-8 MRLM avec Duree_totale pour la sous-population Vs01Dt1
(Nombre d'observations de l'échantillon: 82)

Coefficients estimés du modèle								
Variables	Constante	Langue	Age	Sexe	Intérêt	visite	quête	avantage
Coefficients	0,0137	0,0066	-0,0001	-0,001	-0,0004	0,00204	0,0013	0,0005
p-value	0,4662	0,316	0,9422	0,7803	0,2323	0,0037	0,4069	0,5421
Paramètres du modèle								
R²			F-stat			p-value		
0,207			2,76			0,0131		

Tableau-A III-9 MRLM avec Duree_totale pour la sous-population Vs01Dt2
(Nombre d'observations: 193)

Coefficients estimés du modèle								
Variables	Constante	Langue	Age	Sexe	Intérêt	visite	quête	avantage
Coefficients	0,18583	-0,0238	0,0016	-0,021	-0,0006	0,0078	0,0057	-0,0051
p-value	0,00025	0,25154	0,50473	0,1292	0,58367	<0,000	0,0067	0,062905
Paramètres du modèle								
R²			F-stat			p-value		
0,211			7,05			< 0,000		

Tableau-A III-10 MRLM avec Duree_totale pour la sous-population Vs20Dt
(Nombre d'observations: 422)

Coefficients estimés du modèle								
Variables	Constante	Langue	Age	Sexe	Intérêt	visite	quête	avantage
Coefficients	-0,0892	0,0218	0,0023	-0,0198	-0,002	0,001	0,0087	0,049
p-value	0,18944	0,1512	0,5079	0,24812	0,12174	0,1722	0,0022	< 0,000
Paramètres du modèle								
R²			F-stat			p-value		
0,37			34,7			< 0,000		

ANNEXE IV

GRAPHIQUES DE L'APPLICATION DU MODÈLE DE RÉGRESSION LINÉAIRE MULTIPLE SUR LES SOUS-POPULATIONS

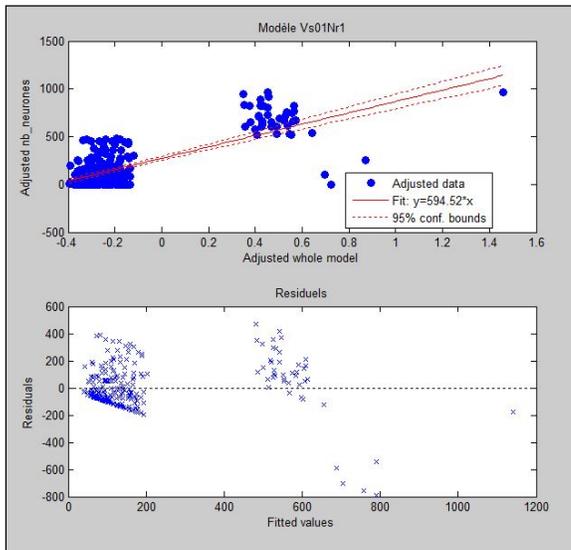


Figure-A IV-1 MRLM et résiduels pour Nb_neurones dans Vs01Nr1

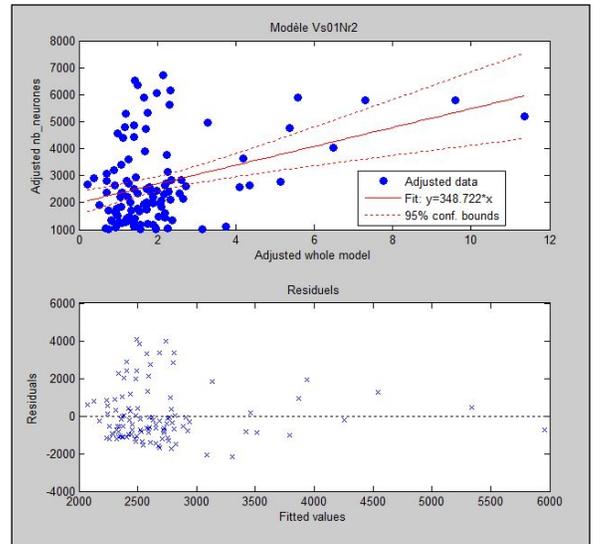


Figure-A IV-2 MRLM et résiduels pour Nb_neurones dans Vs01Nr2

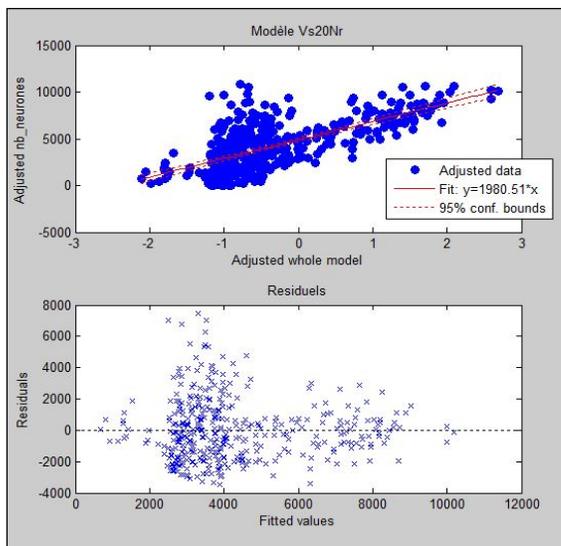


Figure-A IV-3 MRLM et résiduels pour Nb_neurones dans Vs20Nr

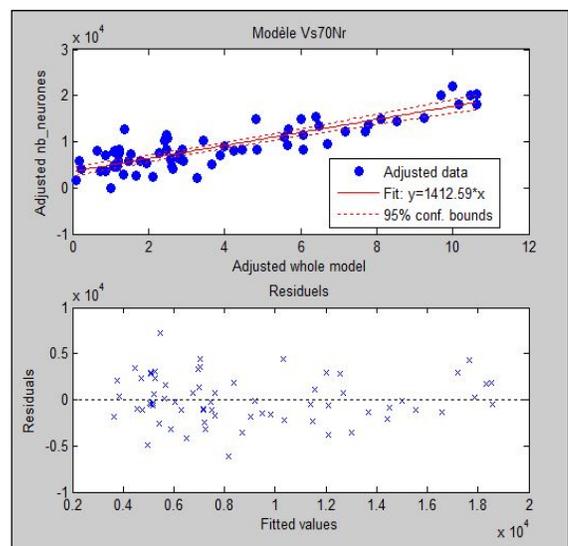


Figure-A IV-4 MRLM et résiduels pour Nb_neurones dans Vs70Nr

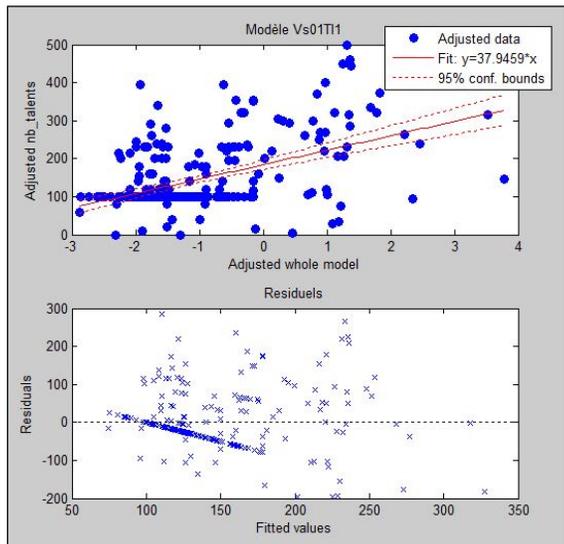


Figure-A IV-5 MRLM et résiduels pour Nb_talents dans Vs01T11

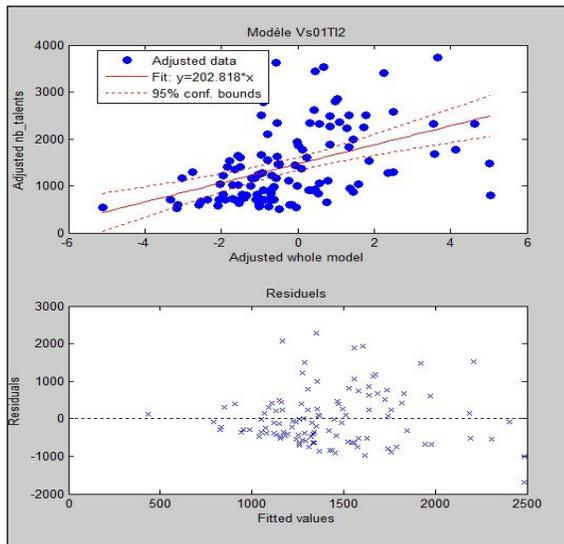


Figure-A IV-6 MRLM et résiduels pour Nb_talents dans Vs01TI2

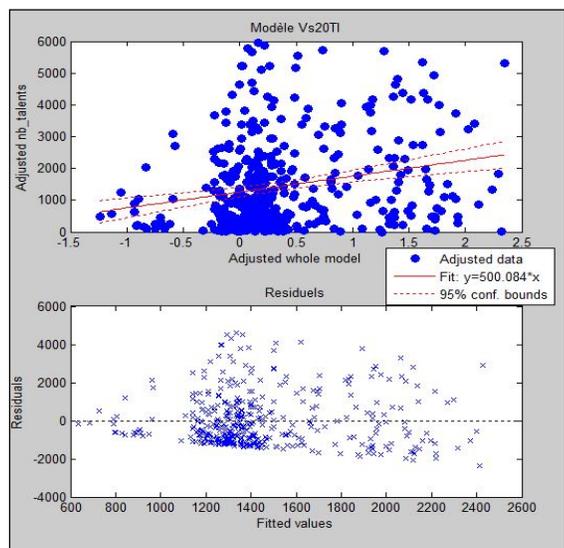


Figure-A IV-7 MRLM et résiduels pour Nb_talents dans Vs20T1

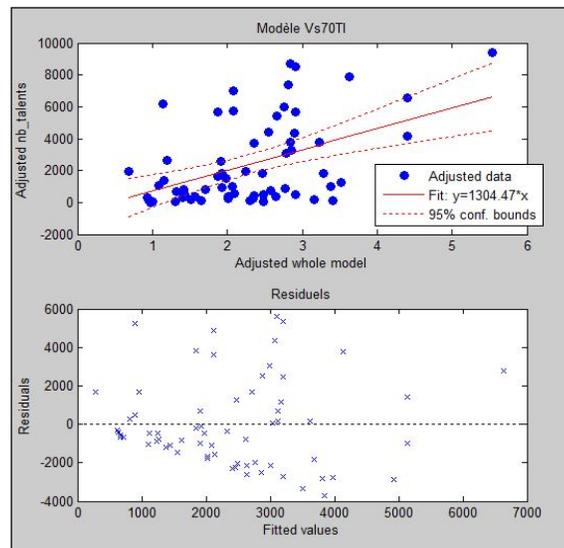


Figure-A IV-8 MRLM et résiduels pour Nb_talents dans Vs70T1

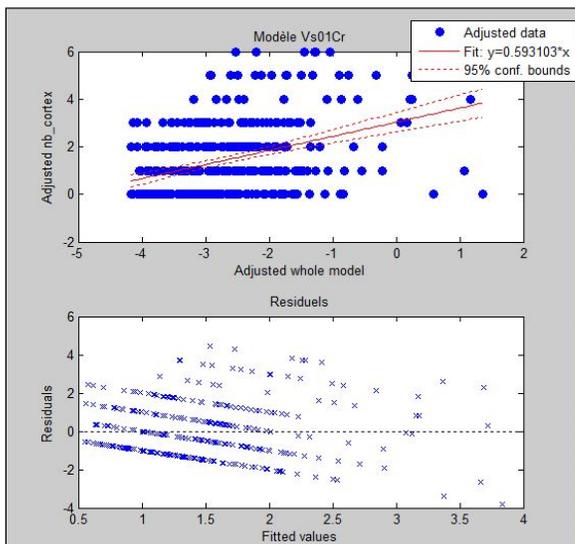


Figure-A IV-9 MRLM et résiduels pour Nb_cortex dans $Vs01Cr$

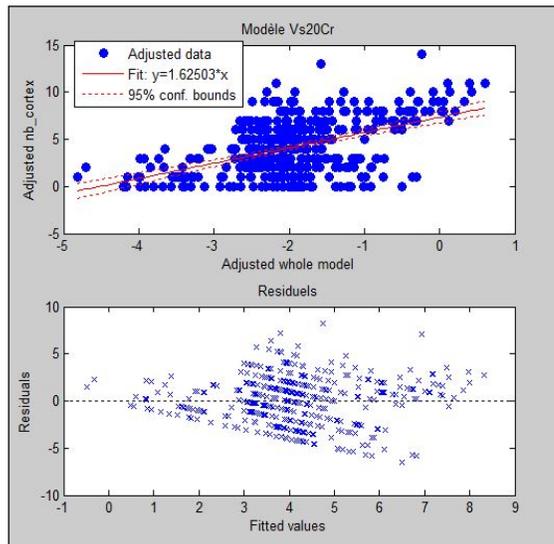


Figure-A IV-10 MRLM et résiduels pour Nb_cortex dans $Vs20Cr$

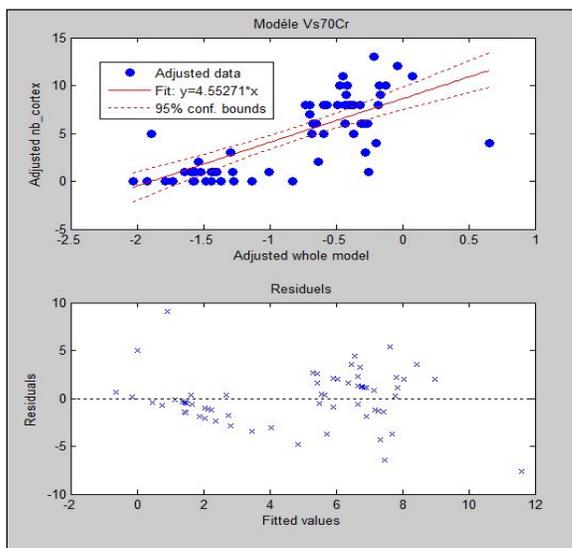


Figure-A IV-11 MRLM et résiduels pour Nb_cortex dans $Vs70Cr$

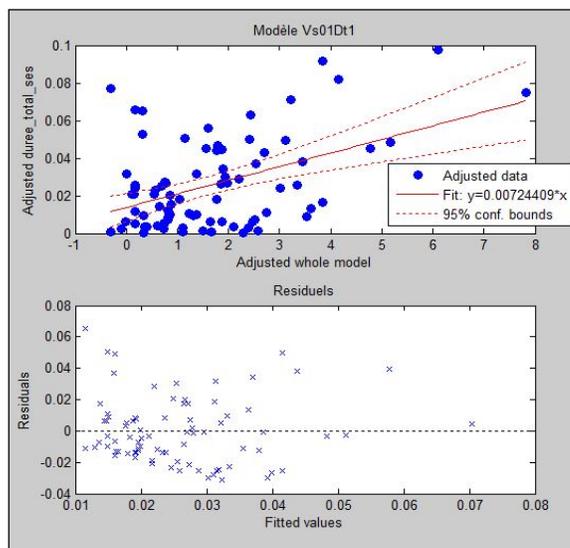


Figure-A IV-12 MRLM et résiduels pour $Duree_totale$ dans $Vs01Dt1$

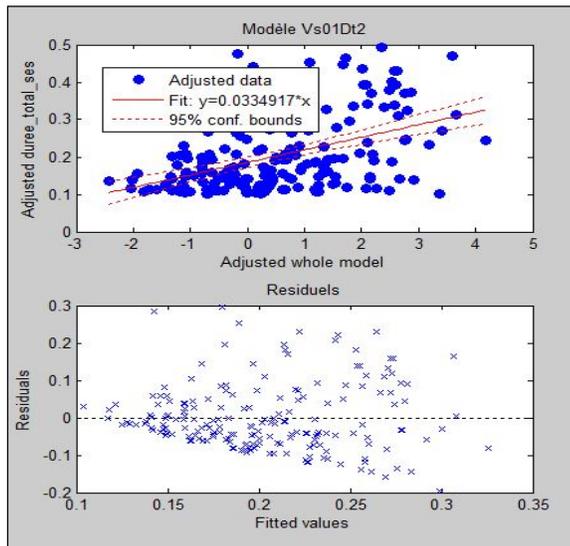


Figure-A IV-13 MRLM et résiduels pour Duree_totale dans Vs01Dt2

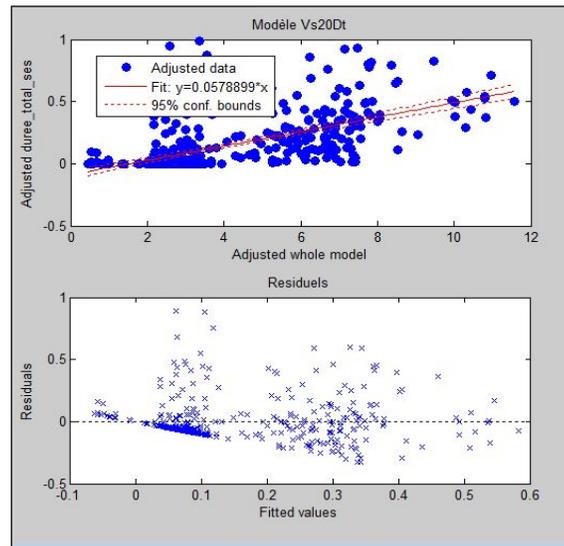


Figure-A IV-14 MRLM et résiduels pour Duree_totale dans Vs20Dt

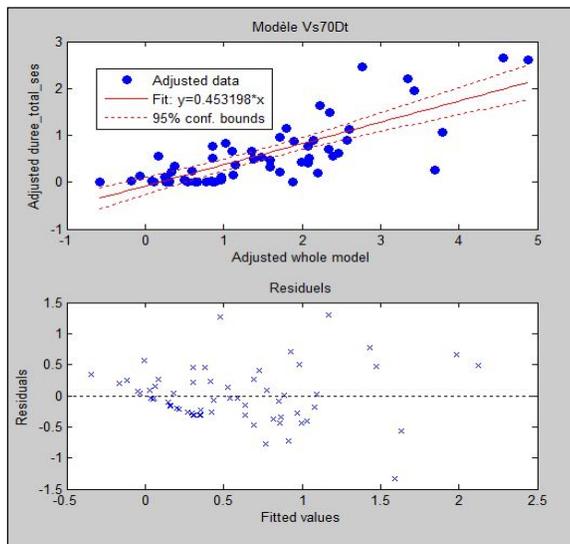


Figure-A IV-15 MRLM et résiduels pour Duree_totale dans Vs70Dt

ANNEXE V

RÉSULTATS DE L'APPLICATION DU MODÈLE DE RÉGRESSION LINÉAIRE MULTIPLE SUR PLUSIEURS ÉCHANTILLONS

- Variable Nb_neurones en fonction des variables indépendantes dans les sous-populations Vs01Nr1 et Vs01Nr2

Tableau-A V-1 MRLM avec Nb_neurones pour la sous-population Vs01Nr1

	Échantillon1	Échantillon 2	Échantillon 3	Échantillon 4
R^2	0,559	0.546	0.661	0.62
F-stat	47,4	45.1	73	61.1
p-value	< 0,000	< 0,000	< 0,000	< 0,000

Tableau-A V-2 MRLM avec Nb_neurones pour la sous-population Vs01Nr2

	Échantillon1	Échantillon 2	Échantillon 3	Échantillon 4
R^2	0,143	0.28	0.252	0.173
F-stat	2,45	5.73	4.96	3.08
p-value	0,0229	< 0,000	< 0,000	0.00544

- Variable Nb_talents en fonction des variables indépendantes dans les sous-populations Vs01T11 et Vs01T12

Tableau-A V-3 MRLM avec Nb_talents pour la sous-population Vs01T11

	Échantillon1	Échantillon 2	Échantillon 3	Échantillon 4
R^2	0,242	0.103	0.101	0.192
F-stat	12	4.28	4.19	8.9
p-value	< 0,000	0.000167	0.000216	< 0,000

Tableau-A V-4 MRLM avec Nb_talents pour la sous-population Vs01Tl2

	Échantillon1	Échantillon 2	Échantillon 3	Échantillon 4
R²	0,202	0.159	0.205	0.141
F-stat	3,79	2.84	3.87	2.47
p-value	0,001	0.00942	0.000872	0.0218

- Variable Nb_cortex en fonction des variables indépendantes dans la sous-population Vs01Cr

Tableau-A V-5 MRLM avec Nb_cortex pour la sous-population Vs01Cr

	Échantillon1	Échantillon 2	Échantillon 3	Échantillon 4
R²	0,153	0.249	0.157	0.155
F-stat	9,46	17.4	9.82	9.62
p-value	< 0,000	< 0,000	< 0,000	< 0,000

- Variable Duree_totale en fonction des variables indépendantes dans la sous-population Vs01Dt1

Tableau-A V-6 MRLM avec Duree_totale pour la sous-population Vs01Dt1

	Échantillon1	Échantillon 2	Échantillon 3	Échantillon 4
R²	0,207	0.211	0.18	0.188
F-stat	2,76	2.82	2.32	2.45
p-value	0,0131	0.0116	0.0338	0.0258

ANNEXE VI

RÉSULTATS DE L'APPLICATION DU MODÈLE DE REGROUPEMENT

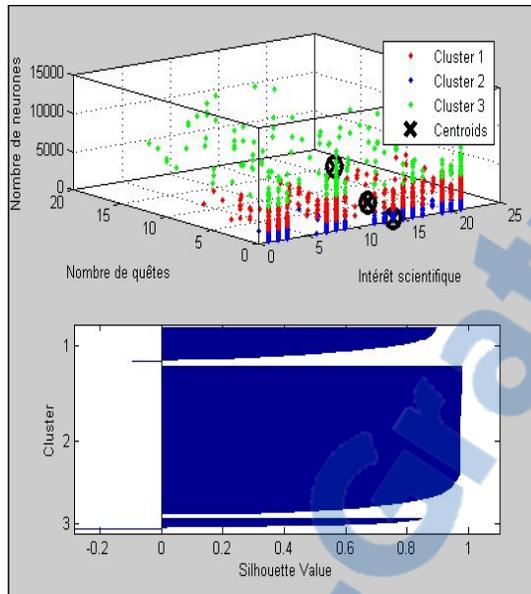


Figure-A VI-1 Neurones, intérêt scientifique et quêtes dans Vs01

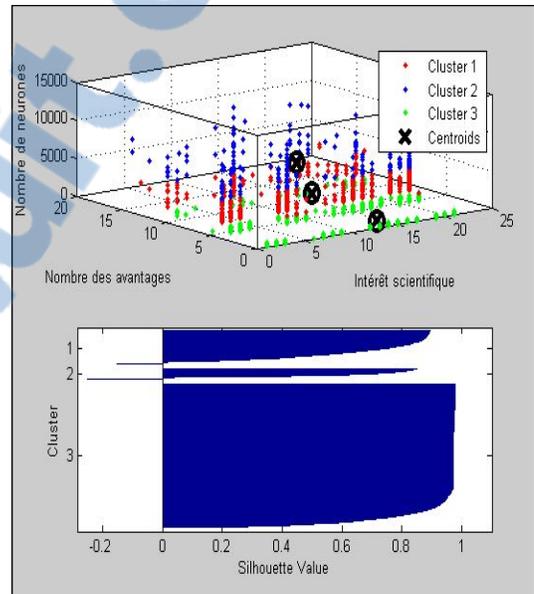


Figure-A VI-2 Neurones, intérêt scientifique et avantages dans Vs01

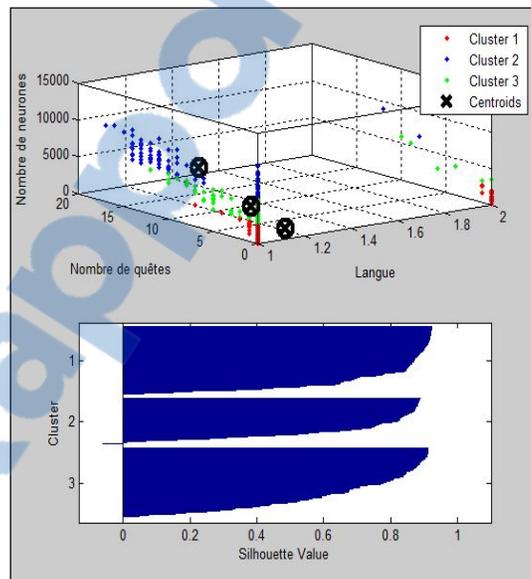


Figure-A VI-3 Neurones, langue et quêtes dans Vs20

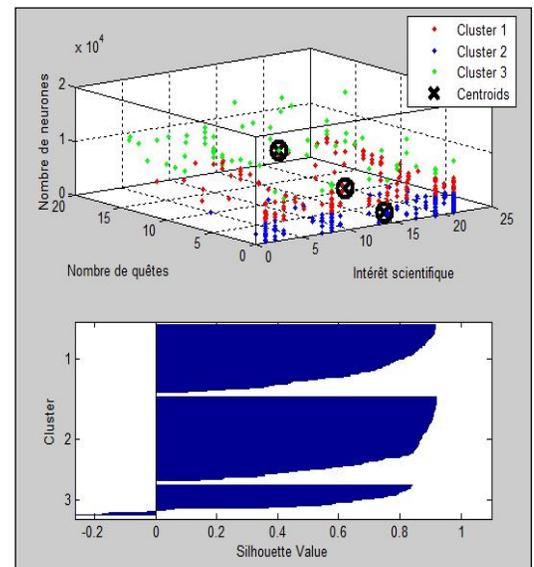


Figure-A VI-4 Neurones, intérêt scientifique et quêtes dans Vs20

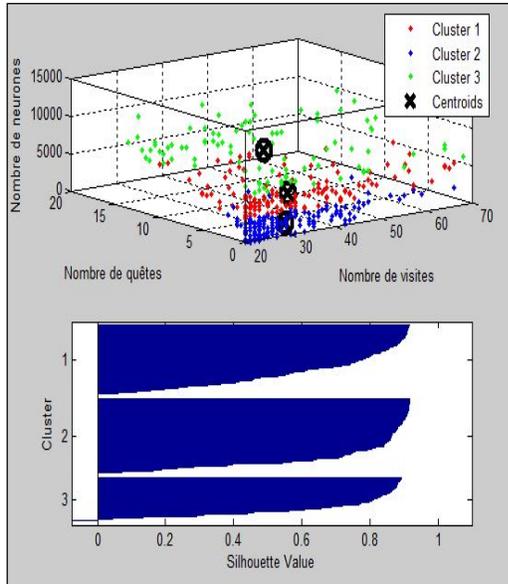


Figure-A VI-5 Neurones, visites et quêtes dans Vs20

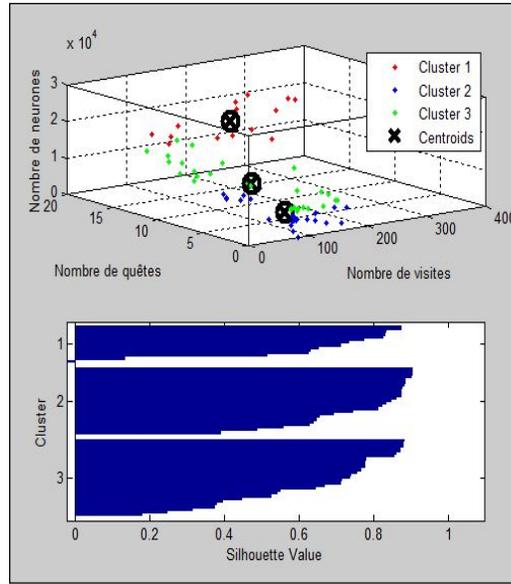


Figure-A VI-6 Neurones, visites et quêtes dans Vs70

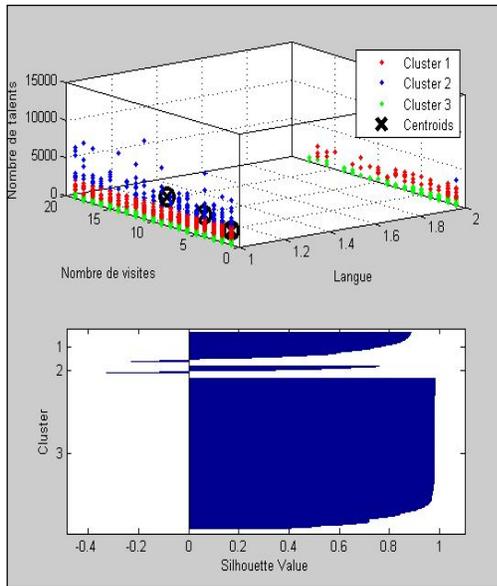


Figure-A VI-7 Talents, langue et visites dans Vs01

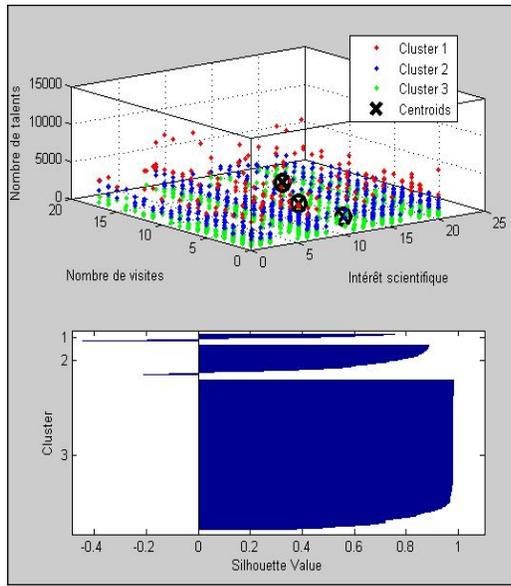


Figure-A VI-8 Talents, intérêt scientifique et visites dans Vs01

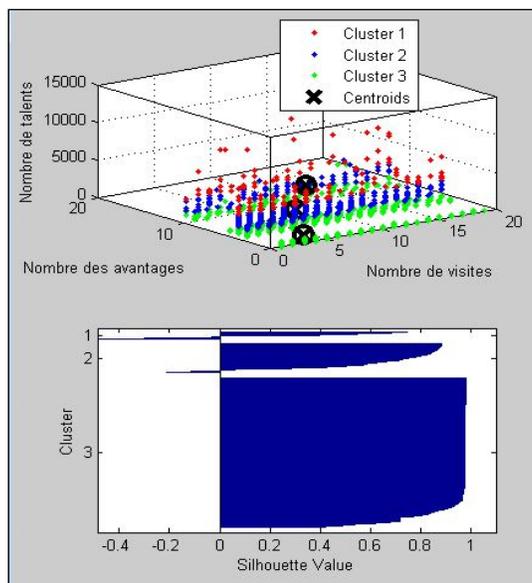


Figure-A VI-9 Talents, visites et avantages dans Vs01

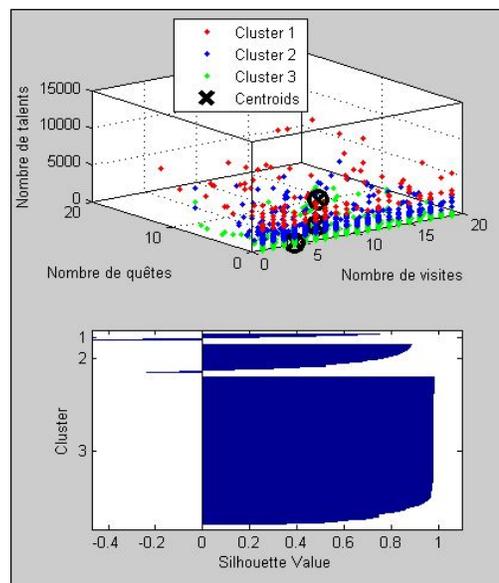


Figure-A VI-10 Talents, visites et quêtes dans Vs01

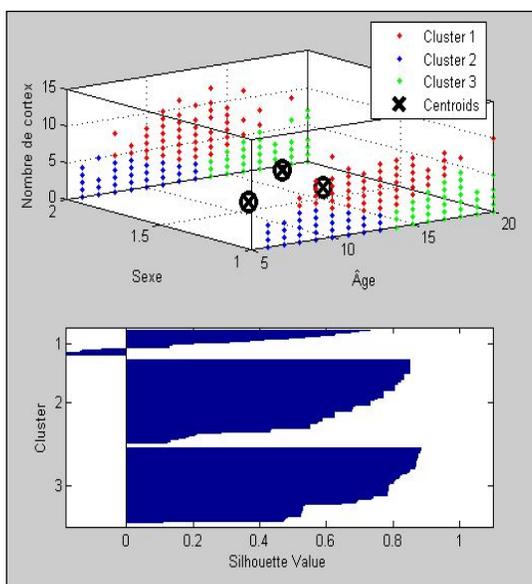


Figure-A VI-11 Cortex, âge et sexe dans Vs01

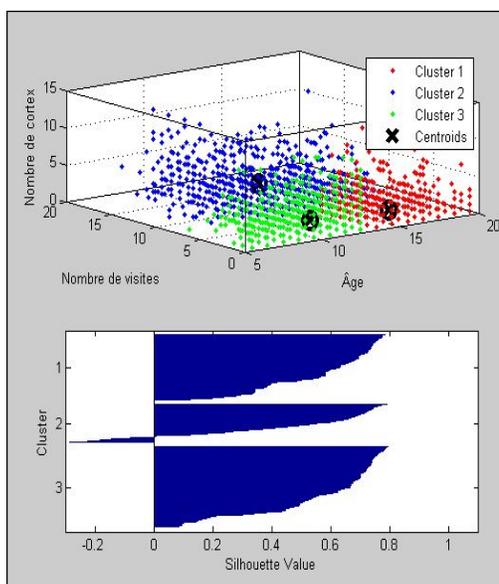


Figure-A VI-12 Cortex, âge et visites dans Vs01



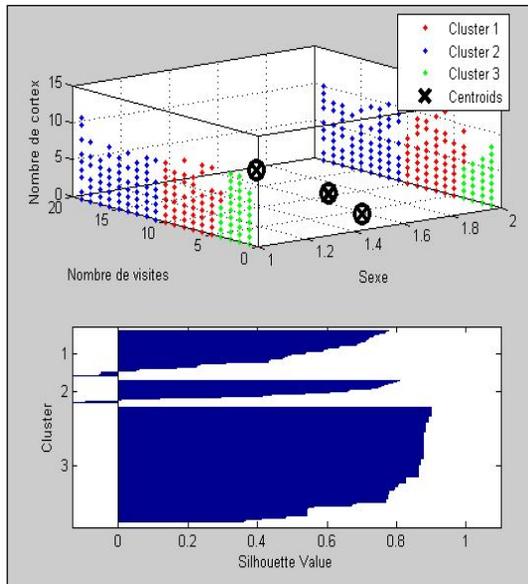


Figure-A VI-13 Cortex, sexe et visites dans Vs01

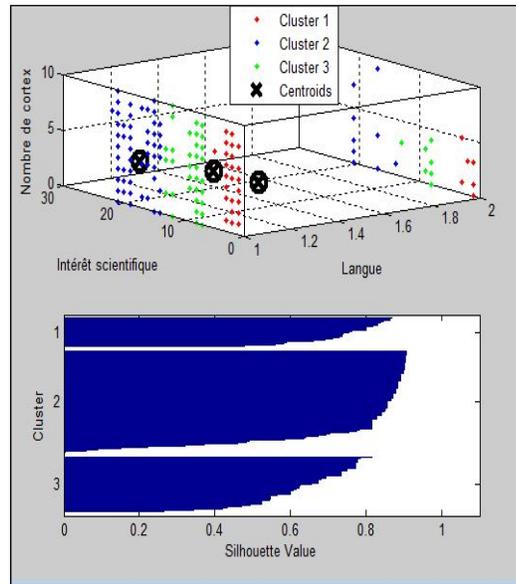


Figure-A VI-14 Cortex, langue et intérêt scientifique dans Vs20

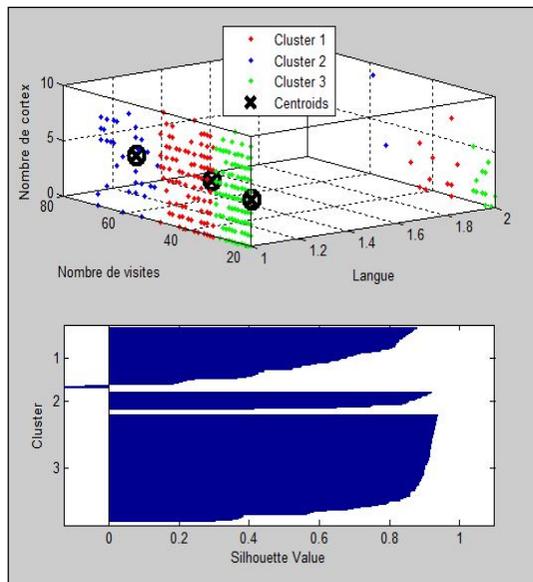


Figure-A VI-15 Cortex, langue et visites dans Vs20

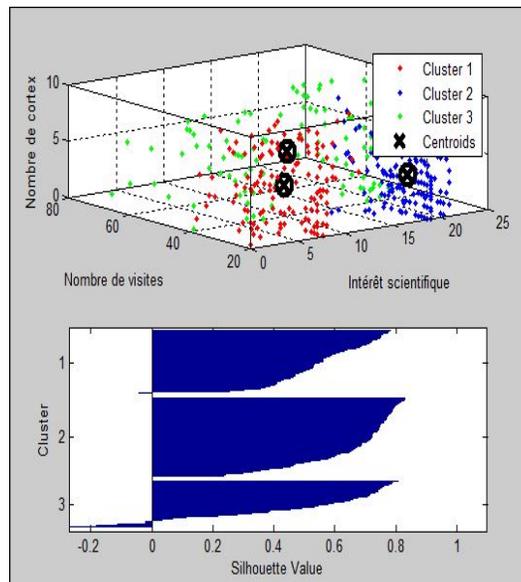


Figure-A VI-16 Cortex, intérêt scientifique et visites dans Vs20

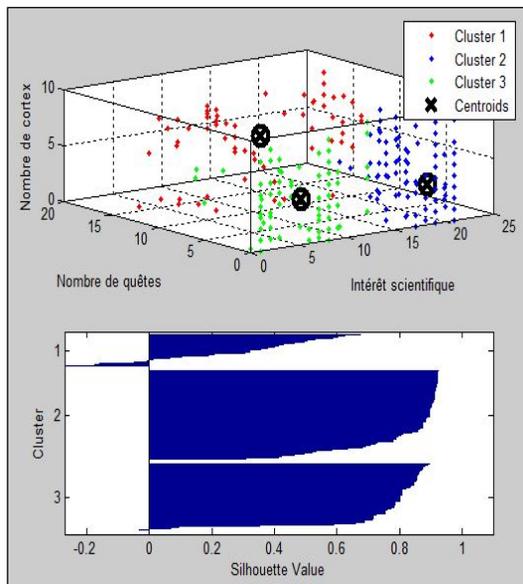


Figure-A VI-17 Cortex, intérêt scientifique et quêtes dans Vs20

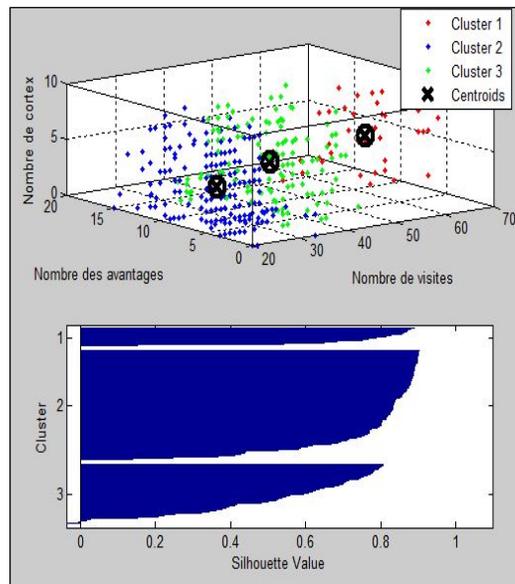


Figure-A VI-18 Cortex, visites et avantages dans Vs20

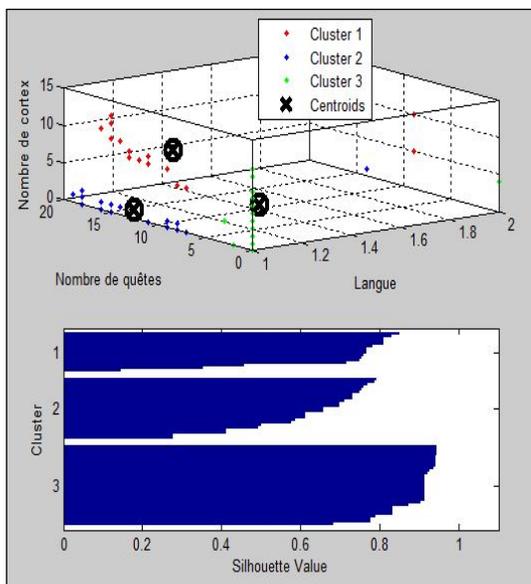


Figure-A VI-19 Cortex, langue et quêtes dans Vs70

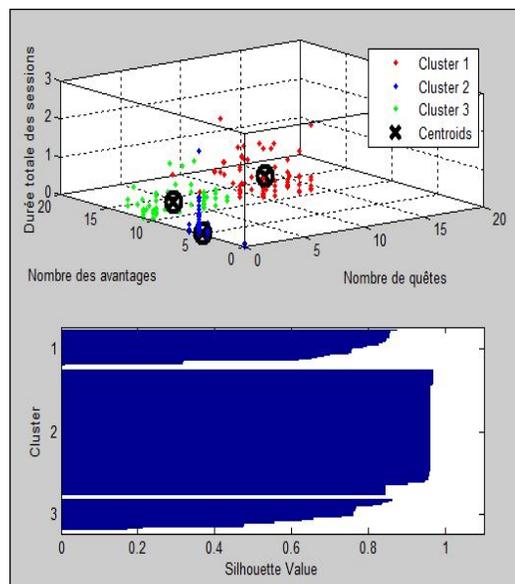


Figure-A VI-20 Durée totale, quêtes et avantages dans Vs20

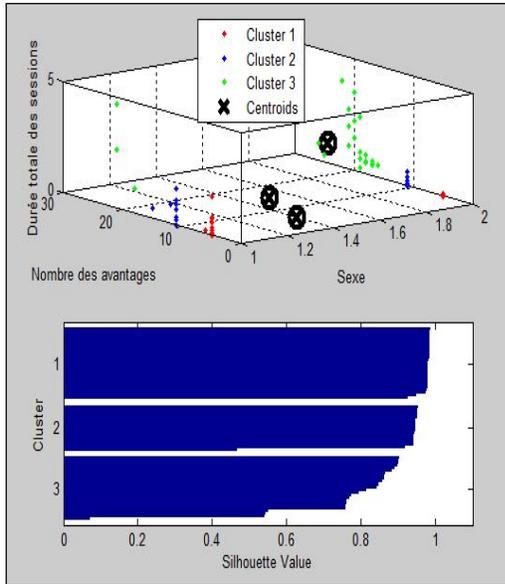


Figure-A VI-21 Durée totale, sexe et avantages dans Vs70

ANNEXE VII

DÉSCRIPTION DU JEU SÉRIEUX « SCIENCE EN JEU »

« Science en jeu » est une plateforme de jeux en ligne (<http://www.scienceenjeu.com/>) créée par la compagnie CREO (Figure-A VII-1). Elle prend la forme d'un monde virtuel destiné à promouvoir l'apprentissage des sciences et des technologies chez des enfants et adolescents de 9 à 14 ans. Elle offre également des exercices de stimulation cognitive (Gym Cérébral) pour les joueurs de tout âge voulants entrainer leurs fonctions cognitives (mémoire, attention, etc.).



Figure-A VII-1 Portail de « Science en jeu »
Tirée de CREO (2014)

Pour pouvoir participer au jeu, les joueurs choisissent un avatar (personnage fictif représentant le joueur dans un monde virtuel) qui leur permettra d'explorer librement le monde virtuel en tant que visiteur (accès limité au contenu des jeux) ou en tant que joueur inscrit en fournissant quelques informations (accès à tout le contenu des jeux) (Figure-A VII-2 et Figure-A VII-3).



Figure-A VII-2 Authentification après inscription ou accès comme visiteur
Tirée de CREO (2014)



Figure-A VII-3 Personnalisation de l'avatar
Tirée de CREO (2014)

Sur les îles virtuelles thématiques, le joueur découvre divers contenus: des jeux, des informations, des nouvelles, des photos, des vidéos, des capsules humoristiques, ainsi que des objets d'apprentissage interactifs (Figure-A VII-4). Plusieurs thèmes scientifiques sont traités (mathématiques, physique, chimie, environnement, etc.) dans ces îles virtuelles à travers des missions (quêtes) à accomplir, des activités à réaliser, des questionnaires à répondre, etc. (Figure-A VII-5 et Figure-A VII-6).



Figure-A VII-4 Îles virtuelles thématiques
Tirée de CREO (2014)



Figure-A VII-5 Quête Mathématicos
Tirée de CREO (2014)

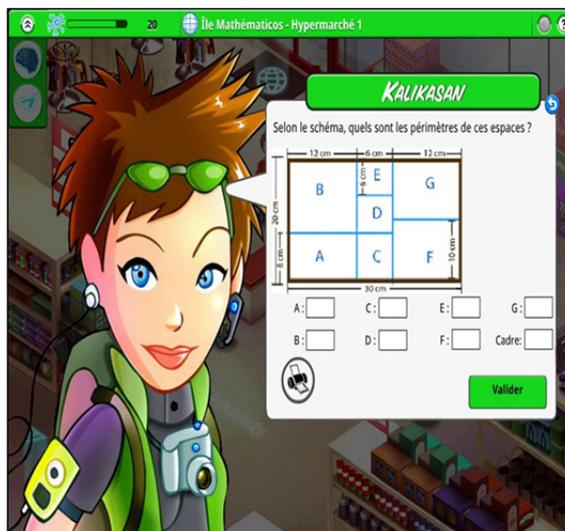


Figure-A VII-6 Activité dans la quête
Mathématicos
Tirée de CREO (2014)

Chaque joueur a pour but d'accumuler des points afin de faire évoluer son avatar du statut de simple apprenti-scientifique jusqu'au stade de super-savant. Trois types de points sont accordés tout au long de l'expérience du jeu (Figure-A VII-7 et Figure-A VII-8), à savoir:

- Les neurones qui représentent les points d'expérience gagnés en accomplissant les missions et les activités du jeu;
- Les talents qui représentent la monnaie dans le monde virtuel. Ils sont gagnés en même temps que les neurones, mais aussi en effectuant certaines activités spécifiques (répétition de tâches, mini-jeux, etc.). Ils servent à se procurer différents items; changer l'apparence de son avatar, acheter des potions magiques, etc. ;
- Les cortex qui représentent les niveaux atteints par le joueur. Ils permettent au joueur d'acquérir des pouvoirs tels que l'interaction avec les autres joueurs, et l'accès à plus d'options et de quêtes dans le jeu.

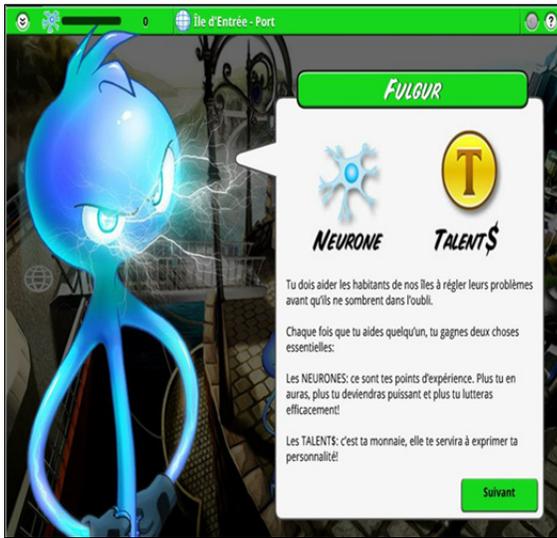


Figure-A VII-7 Neurones et talents
Tirée de CREO (2014)



Figure-A VII-8 Résultats du joueur
Tirée de CREO (2014)

LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES

- Alvarez, Julien. 2007. «Du jeu vidéo au serious game : Approches culturelle, pragmatique et formelle». Thèse – spécialité science de la communication et de l’information, Toulouse, Université Toulouse II – Toulouse le Mirail, 428 p.
- Alvarez, Julien et Damien Djaouti. 2010. *Introduction au serious game*. Paris: Questions théoriques, 227 p.
- Anagnostou, Kostas et Manolis Maragoudakis. 2009. « Data mining for player modeling in videogames ». In *13th Panhellenic Conference on Informatics (PCI)*. (Corfu, September 10-12 2009), p. 30-34. Piscataway : IEEE Computer Society.
- Basak, Chandramallika *et al.* 2008. «Can training in a real-time strategy video game attenuate cognitive decline in older adults? ». *American Psychological Association, Psychology and aging*, vol. 23, n° 4, p. 765-777.
- Bellack, Alan S. *et al.* 2005. «The development of a Computer-Assisted Cognitive Remediation Program for Patients schizophrenia». *Isr. J. Psychiatry Relat Sci*, vol. 42, n° 1, p. 5-14.
- Boughzala, Imed, Ikram Bououd et Hélène Michel. 2013. « Characterization and Evaluation of Serious Game : A perspective of their use in higher education ». In *46th Hawaii International Conference on System Sciences (HICSS)*. (Wailea, Jan. 7-10 2013), p. 844-852. Los Alamitos: IEEE Computer Society.
- Brisson, A. *et al.* 2012. «Artificial Intelligence and Personalization Opportunities for Serious Games». In *Eighth Artificial Intelligence and Interactive Digital Entertainment Conference (AIIDE)*. (Stanford, October 8-12 2012), p. 51-57.
- Caglio, Marcella *et al.* 2009. « Video Game Play Changes Spatial and Verbal Memory: Rehabilitation of a Single Case With Traumatic Brain Injury ». *Journal of Cognitive Processing*, vol. 10, p. 195-197.
- Chen, Yinpeng, *et al.* 2006. «The design of a real-time, multimodal biofeedback system for stroke patient rehabilitation». In *14th annual ACM international conference on Multimedia*. (Santa Barbara, October 23-27 2006), p. 763-772. New York: Association for Computing Machinery.

- Conati, M. et M. Manske. 2009. « Evaluating adaptive feedback in an educational computer game ». In *Intelligent Virtual Agents. 9th International Conference, IVA 2009*. (Amsterdam, Sept. 14-16 2009), p. 146-158. Berlin: Springer Verlag.
- Conconi, Alex *et al.* 2008. « PlayMancer: A Serious Gaming 3D Environment ». In *International Conference on Automated Solutions for Cross Media Content and Multi-channel Distribution*. (Piscataway, Nov. 17-19 2008), p. 111-117. Piscataway (USA) : IEEE.
- Coyle, David *et al.* 2005. « Personal Investigator: A Therapeutic 3D Game for Adolescent Psychotherapy ». *Interactive Technology and Smart Education*, vol. 2, n° 2, p. 73-88.
- Crawford, Chris. 1984. *The art of Computer Game Design*. Berkeley, Calif. Osborne/McGraw-Hill, 113 p.
- CREO. 2014. « Science en jeu: la puissance de la connaissance ». In *Science en jeu*. En ligne <<http://www.scienceenjeu.com/index.php>>. Consulté janvier-septembre 2014.
- Delmas, Guylain, Ronan Champagnat et Michel Augeraud. 2007. « Plot monitoring for interactive narrative games ». In *4th International Conference on Advances in Computer Entertainment Technology, ACE 2007*. (Salzburg, June 13-15 2007), p.17-20. New York: Association for Computing Machinery.
- Derbali, Lotfi et Claude Frasson. 2010. « Prediction Of Players Motivational States Using Electrophysiological Measures During Serious Game Play ». In *IEEE 10th International Conference on Advanced Learning Technologies (ICALT)*. (Sousse, July 5-7 2010), p. 498-502. Los Alamitos : IEEE Computer Society.
- Djaouti, Damien, Julian Alvares et Jean-Pierre Jessel. 2010. « Can Gaming 2.0 Help Design Serious Game?: A Comparative study ». In *5th ACM SIGGRAPH Symposium on Video Games*. p. 11-18.
- Djaouti, Damien. 2011. « Serious Game Design – considérations théoriques et techniques sur la création des jeux vidéo à vocation utilitaire ». Thèse en informatique, université de Toulouse III- Paul Sabatier, 330 p.
- El-Nasr, M. S. et Brian K. Smith. 2006. « Learning through game modding ». *Computers in Entertainment (CIE)*, vol. 4, n° 1, p. 1-20

- Green, Shawn C. et Daphne Bavelier. 2003. «Action Video Game Modifies Visual Selective Attention». *Nature Publishing Group*, vol. 423, p. 534-537.
- Hocine, Nadia *et al.* 2011. «Techniques d'adaptation dans les jeux ludiques et sérieux : État de l'art». *Revue de l'intelligence artificielle*, vol. 25, n° 2, p. 253-280.
- Imbeault, Frédérick, Bruno Bouchard et Abdenour Bouzouane. 2011. « Serious Games in Cognitive Training for Alzheimer's Patients ». In *IEEE 1st International Conference on Serious Games and Applications for Health, SeGAH 2011*. (Braga, November 16-18, 2011), p. 8. Piscataway (USA): IEEE Computer Society
- Ismailović, Damir *et al.* 2012. «Adaptive serious game development». In *2012 2nd International Workshop on Games and Software Engineering (GAS 2012)*. (Zurich, June 9 2012), p. 23-26. Piscataway (USA): IEEE.
- Jackson, G. Tanner, Natalie L. Davis et Danielle S. McNamara. 2011. « Students' Enjoyment of a Game-Based Tutoring System ». In *15th International Conference, AIED 2011*. (Auckland, June 28 – July 2011), p. 475-477. Berlin: Springer Berlin Heidelberg.
- Jequier, Sophie et Maxime Beaugeois. 2012. «Science Express, le "serious game" d'UNISCIEL comme vecteur de motivation». *TICE*, p. 242-247.
- Johnson, W. Lewis, Hannes Vilhjalmsson et Stacy Marsella. 2005. « Serious Game for Language Learning development: How Much Game, How Much AI? ». In *International Conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology*. (Amsterdam, July 2005), p. 306-313.
- Kebritchi, Mansourah et Atsusi "2c" Hirumi. 2008. « Examining the pedagogical foundations of modern educational computer games ». *Computers & Educations*, vol. 51, p. 1729-1743.
- Kickmeier-Rust, Michael D. *et al.* 2007. «Immersive Digital Games: The Interfaces for Next-Generation E-Learning? ». In *4th International Conference on Universal Access in Human-Computer Interaction, UAHCI 2007*. (Beijing, July 22-27, 2007), p. 647-656. Berlin: Springer Berlin Heidelberg.
- Manning, Christopher D., Prabhakar Raghavan et Hinrich Schütze. 2008. «Introduction to Information Retrieval: Evaluation of clustering». En ligne. <<http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html>>. Consulté le 19 juin 2014.

- Marfisi-Schottman, Iza *et al.* 2012. « Comment évaluer la qualité d'un Learning Game pendant sa conception? ». In *8ème Colloque Technologies de l'Information et de la Communication pour l'Enseignement Tice 2012*. (Lyon, 11-13 décembre 2012), p. 80-90.
- MathWorks-1. 2014. « Fuzzy Clustering – What is Data Clustering ». In *Fuzzy Clustering – MATLAB & Simulink*. En ligne <http://www.mathworks.com/help/fuzzy/fuzzy-clustering.html#bq_z_k1>. Consulté le 14 avril 2014.
- MathWorks-2. 2014. «clustering.evaluation.CalinskiHarabaszEvaluation class». In *Calinski-Harabasz criterion clustering evaluation object - MATLAB*. En ligne <<http://www.mathworks.com/help/stats/clustering.evaluation.calinskiharabaszevaluation-class.html>>. Consulté le 22 mai 2014.
- MathWorks-3. 2014. «clustering.evaluation.SilhouetteEvaluation class». In *Silhouette criterion clustering evaluation object - MATLAB*. En ligne <<http://www.mathworks.com/help/stats/clustering.evaluation.silhouetteevaluation-class.html>>. Consulté le 14 avril 2014.
- MathWorks-4. 2014. « kmeans ». In *K-means clustering – MATLAB kmeans*. En ligne <<http://www.mathworks.com/help/stats/kmeans.html> >. Consulté le 14 avril 2014.
- Matlab. 2014. « Product Description: The Language of Technical Computing ». In *Help: Getting Started with MATLAB*. Consulté le 18 février 2014.
- McAtamney, H., B. O'Shea, F. Mtenzi. 2005. « Using the Crytek game engine in the Dublin Institute of Technology ». In *7th International Conference on Computer Games: AI, Animation, Mobile, Educational & Serious Games, CNBDI*. (Angoulême, November 28-30th. 2005), 5 p.
- Mills, Chris et Barney Dalgarno. 2007. « A conceptual model for game-based intelligent tutoring systems ». *Australasian Society for Computers in Learning in Tertiary Education*, p. 692-702.
- Montserrat, Baptiste, Élise Lavoué et Sébastien George. 2012. « Learning game 2.0 : Support à l'Apprentissage Collaboratif par la Modification de Jeux Vidéo ». In *8ème Colloque Technologies de l'Information et de la Communication pour l'Enseignement Tice 2012*. (Lyon, 11-13 décembre 2012), p. 98-109.

- Muratet, Mathieu. 2010. « Conception, réalisation et évaluation d'un jeu sérieux de stratégie temps réel pour l'apprentissage des fondamentaux de la programmation ». Thèse de doctorat en informatique, Toulouse, Université Toulouse III - Paul Sabatier, 197 p.
- OQFL : Office québécois de la langue française. 2009. « Fiche terminologique : Avatar ». In *Le grand dictionnaire terminologique*. En ligne <http://www.gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=8357969>. Consulté le 09 juin 2014.
- Peirce, N., O. Conlan, V. Wade. 2008. « Adaptive Educational Games: Providing Non-invasive Personalised Learning Experiences ». In *Second IEEE International Conference on Digital Games and Intelligent Toy Enhanced*. (Banff, Nov. 17-19 2008), p. 28-35. Piscataway (USA): IEEE.
- Ratté, Sylvie (Ratté-1). 2013. *Forage de données*. [Notes du cours MTI830 Forage de textes et de données audiovisuelles]. Montréal : École de Technologie Supérieure.
- Ratté, Sylvie (Ratté-2). 2013. *Clustering*. [Notes du cours MTI830 Forage de textes et de données audiovisuelles]. Montréal : École de Technologie Supérieure.
- Rego, Paula, Pedro Miguel Moreira et Luis Paulo Reis. 2010. « Serious games for rehabilitation: A survey and a classification towards a taxonomy ». In *Fifth Iberian Conference on Information Systems and Technologies*. (Santiago de Compostela, June 16-19 2010), p. 1-6. Piscataway (USA): IEEE.
- Roche, Mathieu. 2008. « Le langage naturel et la fouille de données ». En ligne. 21 p. <http://www.lirmm.fr/~teisseir/SUPPORTCOURS/2007_FDA_et_TAL.pdf>. Consulté le 19 juin 2014.
- Salmeron, Raquel *et al.* 2011. « *AWARD^{prime}*: An Adaptive Web Based-Tool Prototype for Neurocognitive Individualized Assessment and Training ». In *Foundations on Natural and Artificial Computation. 4th International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 201*. (La Palma, May 30-June 3 2011), p. 540-549. Berlin (Germany): Springer-Verlag.
- Santos, Fábios E. G *et al.* 2011. « Assessment of ADHD through a Computer Game: An Experiment with a Sample of Students ». In *Third International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*. (Athens, May 4-6 2011), p. 104-111. IEEE.

- Sehaba, Karim, Pascal Estraillier et Didier Lambert. 2005. «Interactive educational games for autistic children with agent-based system». In *Entertainment Computing - ICEC 2005, 4th International Conference Proceedings*. (Sanda, Sept. 19-21 2005), p. 422-432. Berlin (Germany): Springer-Verlag.
- Shim, Kyong Jin, Hsu Kuo-Wei et Jaideep Srivastava. 2011. «Modeling Player Performance in Massively Multiplayer Online Role-Playing Games: The Effects of Diversity in Mentoring Network». In *Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. (Kaohsiung, July 25-27 2011), p. 438-442. Los Alamitos: IEEE Computer Society.
- Sottolare, Robert A. et Stephen Gilbert. 2011. « Considerations for adaptive tutoring within serious games: authoring cognitive models and game interfaces ». In *15th International Conference Artificial Intelligence in Education*. (Auckland, June 28-July 1 2011), 10 p.
- SPSS. 2014. « Boite à moustache : Rappel Théorique ». In *SPSS de l'UdeS – Boite à moustache*. En ligne <<http://spss.espaceweb.usherbrooke.ca/pages/stat-descriptives/boite-a-moustaches.php>>. Consulté le 03 février 2014.
- Technologysource. 2003. « Definition//Modding ». In *Definition: Modding*. En ligne <<http://technologysource.org/extra/231/definition/5/>>. Consulté le 09 juin 2014.
- TermiumPlus. 2013. « Autostigmatisation ». In *Self Stigma [1 fiche]-Termium Plus*. En ligne <http://www.btb.termiumplus.gc.ca/tpv2alpha/alphafra.html?Lang=fra&i=1&index=alt&__index=alt&srchtxt=SELF+STIGMA>. Consulté le 04 juillet 2014.
- TermiumPlus. 2014. « Visuomoteur ». In *Visuomoteur [1 fiche]-Termium Plus*. En ligne <http://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-fra.html?lang=fra&i=&index=alt&__index=alt&srchtxt=Visuomoteur&comencsrch.x=0&comencsrch.y=0>. Consulté le 19 juin 2014.
- Tufféry, Stéphane. 2010. *Data mining et statistique décisionnelle: l'intelligence des données*. Paris: TECHNIP, 705 p.
- Wartena, Bard O. et Hylke W. van Dijk. 2013. «Bias Blaster – Aiding Cognitive Bias Modification- Interpretation through a bubble shooter induced gameflow». In *Games for Health : Proceedings of the 3rd European Conference on Gaming and Playful Interaction in Health Care*,(Amsterdam, October 28-29 2013), p. 47-60. Wiesbaden: Springer Fachmedien.

Witten, Ian H., Eibe Frank et Mark A. Hall. 2011. *Data mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Burlington: Morgan Kaufmann Publishers, 629 p.

Wolf, Mark J. P. 2001. *The Medium of the Video Game*. Austin: University of Texas Press, 223 p.

