PROJET
STT-3400

Prédiction de la récidive chez les délinquants sexuels au Québec

Pascale AUBIN
Pierre GAGNON

27 avril 2016

Résumé

En criminologie, lors d'une demande de libération conditionnelle, il est important de connaître le risque de récidive d'un individu. Ceci permet de prendre la meilleure décision et de l'encadrer adéquatement advenant le cas que sa demande soit acceptée. Ainsi, les coûts et les risques pour la société sont minimisés. Dans ce contexte, une collecte d'information a été réalisée chez les délinquants sexuels au Québec afin de permettre différentes études sur le risque de récidive de ces individus. L'objectif de ce projet est de s'intéresser à la prédiction de la récidive de libérés à l'aide d'outils qui pourraient être utilisés lors de processus judiciaires.

Le projet est divisé en trois objectifs. Premièrement, on s'intéresse aux variables qui caractérisent la récidive violente ou sexuelle des individus. Pour ce faire, à l'aide des données disponibles, un modèle de régression logistique et un arbre de classification sont construits. Les deux modèles contiennent les mêmes quatre variables. Les taux d'erreur sont respectivement d'environ 11% et 14%. À l'issue de ce premier objectif, pour la facilité de son utilisation, l'arbre de classification est le modèle jugé le plus pratique. Le second objectif vise à tester trois outils de prédiction de la récidive actuellement utilisés par les intervenants qui agissent auprès des libérés soient : l'ERRRS, la Statique-99 et la Statique-2002. Celui des trois qui possède le meilleur Kappa de Cohen est la Statique-2002. Il sera alors possible de le comparer à l'arbre de classification afin de déterminer quel outil est le meilleur. Le dernier objectif consiste à former différents groupes dans la population à l'étude pour ensuite vérifier si des outils performent mieux pour certains groupes. La classification non supervisée effectuée permet de former quatre groupes. Le plus grand groupe contient 38% des individus et le plus petit contient 13% des individus. Il est intéressant de remarquer que les proportions de récidivistes au sein d'un groupe ne sont pas les mêmes. Elles varient de 3 à 25%. Pour tous les groupes ce sont les modèles créés qui prédisent le mieux la présence ou l'absence de récidive.

Table des matières

1 Description de l'étude

Les individus ayant commis un crime représentent une partie de la population qui nécessite une attention particulière de la part de leurs concitoyens. Plusieurs disciplines et institutions s'affairent à bien comprendre ces individus et à les encadrer adéquatement afin qu'ils ne constituent plus une menace pour la population ou encore mieux, qu'ils s'y réintègrent. Pour que leur encadrement soit efficace, il est nécessaire de bien cerner les besoins de ces individus et pourquoi ils en sont arrivés à commettre un délit. La criminologie est la science qui étudie, entre autres, le comportement des criminels et les facteurs qui le déterminent.

Le chercheur qui chapeautera notre projet est monsieur Patrick Lussier, professeur titulaire de criminologie à la Faculté des sciences sociales de l'Université Laval. Il a complété son doctorat au centre de recherche sur la violence sexuelle à l'Université Simon Fraser. Les travaux de recherche de M. Lussier visent majoritairement à expliquer pourquoi et comment certains éléments physiques et comportementaux prédisposent les délinquants sexuels à récidiver. Par exemple, plusieurs de ses projets, réalisés avec divers collaborateurs, sont fondés sur la même base de données que celle utilisée dans le projet. Cette base de données contient de l'information sur les délinquants sexuels au Québec. Des facteurs comme l'âge ont étés étudiés sous plusieurs angles afin de connaître leur effet sur la propension à commettre un autre crime à la suite de leur libération [1]. On s'intéresse au risque de récidive d'un individu, puisqu'il permet d'aider à accepter ou refuser une demande de libération conditionnelle. Il détermine aussi l'importance du suivi auprès d'un individu qui retourne dans la communauté.

La principale tâche du projet consiste à essayer de prédire la récidive violente ou sexuelle chez les délinquants sexuels au Québec. Dans le cadre du projet, la récidive est définie comme étant une nouvelle condamnation pour un crime violent ou sexuel à la suite d'une libération. La prédiction de la récidive chez les délinquants sexuels, sujet de plusieurs publications de M. Lussier, s'effectue actuellement à l'aide de différents outils basés sur quelques variables autant individuelles que psychosociales. Les outils les plus couramment utilisés pour prédire la récidive sont l'ERRRS, la Statique-99 et la Statique-2002. La première tâche consiste à

Tableau 1 – Description des outils de prédiction de la récidive qui seront analysés

Outil	Nombre de populations utilisées pour la construction de l'outil	Nombre d'items
ERRRS	3	4
Statique-99	4	10
Statique-2002	10	14
Modèle Objectif 1	1	Inconnu pour le moment

construire de nouveaux outils permettant de prédire la récidive violente ou sexuelle seront construits. Contrairement aux autres outils, ils seront développés à partir d'une seule population, celle des délinquants sexuels au Québec (tableau 1). La deuxième tâche est de valider les outils actuellement utilisés en criminologie, en plus de celui créé lors du premier objectif, au moment de leur application sur la population des délinquants sexuels au Québec. La troisième tâche est de distinguer les différents profils de délinquants dans la population québécoise et de valider la prédiction de la récidive à l'aide des différents outils pour chaque profil.

2 Description de la collecte des données

Les individus condamnés au Québec à une peine fédérale (sentence de plus de deux ans) sont d'abord évalués au Centre Régional de Réception de Sainte-Anne-des-Plaines à la suite de leur procès. Le résultat de cette évaluation a pour but de déterminer le niveau de sécurité approprié du pénitencier (établissement de détention fédéral) où les individus purgeront leur peine. Entre avril 1994 et juin 2000, il a été proposé à toutes les personnes passant par ce centre qui ont été condamnées pour un délit sexuel au Québec de faire partie de l'étude. Cette collecte de données est une initiative de M. Jean Proulx, criminologue à l'Université de Montréal. L'unité de cette étude observationnelle, que l'on peut qualifier de recensement des délinquants sexuels au Québec entre 1994 et 2000, est le délinquant sexuel.

L'information pour chaque libéré était recueillie à l'aide d'une entrevue semi-dirigée effectuée par des assistants de recherche. Ce type d'entrevue consiste à entretenir une conversation pendant laquelle l'assistant de recherche oriente la conversation dans le but d'obtenir les renseignements dont il a besoin. Il ne s'agit pas de répondre systématiquement à une liste de questions. Ces assistants étaient tous des étudiants gradués en criminologie ou en psychologie. L'information obtenue lors de l'entrevue était validée en la comparant au contenu des dossiers criminels officiels. En cas de discordance entre l'information révélée par l'individu et les sources officielles, ce sont ces dernières qui avaient priorité [1]. Les variables avec lesquelles le projet est effectué sont en majorité des variables vérifiables avec les informations administratives, mais certaines des variables sont nécessairement obtenues lors de l'entrevue, par exemple la variable cohabitation qui n'est pas inclue dans le dossier criminel. La liste des variables du jeu de données se trouve à l'annexe A.

Un total de 553 individus ont accepté de participer, ce qui représente un taux de participation de 93%. Il s'agit exclusivement d'hommes, car les femmes ne passent pas par le centre de Sainte-Anne-des-Plaines. De plus, le nombre de femmes qui reçoivent une peine fédérale pour délit sexuel est très faible, il a donc été décidé de ne pas les inclure dans l'étude. Il faut mentionner que la participation à l'étude se faisait sur une base strictement volontaire, il n'y avait aucune compensation en échange de la participation. Le faible taux de non-réponse, en l'occurrence 7%, s'explique par le fait que les individus admis à Sainte-Anne-des-Plaines sortent à peine d'un processus judiciaire long et éprouvant. Le stress du procès étant tombé, ils ressentent en général le besoin de parler et de se confier, d'où le taux de participation élevé.

Des suivis ont été effectués à différents moments dans le temps à la suite de leur libération afin de vérifier si les individus avaient récidivé. Le dernier suivi remonte à 2007. Étant donné que les individus ne terminent pas tous leur peine en même temps, la période à risque de récidive des individus n'est pas la même pour tous. C'est un élément important dont il faudra tenir compte lors de l'analyse des données. Une autre caractéristique de ces données qui imposera la prudence lors de l'interprétation des résultats est la fraction des personnes interrogées qui ont refusé de participer à l'étude, soit le 7% de non-réponse. En plus de la

non-réponse, il est possible qu'une partie des individus ne soit pas répertoriée en raison de la formulation de leur défense pendant leur procès. Par exemple, si un individu commet un crime sexuel avec violence, il se peut que des négociations concernant les chefs d'accusation entraînent l'exclusion de la mention de l'aspect sexuel du crime dans l'intérêt de l'accusé. Ainsi, il n'est pas offert à quelqu'un dans cette situation de faire partie de l'étude lors de son arrivée au Centre Régional de Réception.

3 Description des outils existants pour prédire la récidive

3.1 ERRRS

Le premier prédicteur à étudier est l'outil d'Évaluation Rapide du Risque de Récidive Sexuelle (ERRRS) publié en 1997 [2]. La construction de cet outil est basée sur trois échantillons de délinquants sexuels, soit un provenant de l'Institut Philippe Pinel (Québec), un provenant de la prison provinciale Millbrook (Ontario) et un de l'Institut psychiatrique Oak Ridge (Ontario). L'ERRRS est constituée de quatre items (un item peut contenir plusieurs variables) présenté au tableau 2.

La grille détaillée de l'outil se situe à l'annexe B. Un pointage est associé à chaque item selon la valeur de la/des variable(s) qui le compose(nt). Le pointage des quatre items est additionné afin d'avoir un pointage total pour l'individu. Le pointage total varie de 0 à 6, où 0 indique un faible risque de récidive [2].

3.2 La Statique-99

Le deuxième prédicteur est la Statique-99 publié en 2000[2]. Tel qu'il est possible de voir dans le tableau 2, tous les éléments de l'ERRRS sont inclus ainsi que six autres items. La Statique-99 est l'outil le plus utilisé actuellement pour la prédiction de la récidive sexuelle. La construction de la Statique-99 est basée sur quatre échantillons de délinquants sexuels, les trois premiers sont ceux énumérés pour l'ERRRS et le quatrième provient de la prison Her Majesty's Prison Service (Royaume-Uni).

La grille détaillée pour l'attribution des pointages pour la Statique-99 est présentée à l'annexe C. Comme pour l'ERRRS, un pointage est associé à chaque item selon la valeur de la/des variable(s) qui le compose(nt). Les pointages des dix items sont additionnés afin d'avoir un pointage total pour l'individu. Le pointage total varie de 0 à 12. Les pointages sont convertis en quatre catégories de risque. Les pointages plus élevés sont associés à un risque de récidive plus élevé[2].

Tableau 2 – Variables contenues dans les outils actuels utilisés pour la prédiction de la récidive

Item	ERRRS	Statique-99	Statique-2002
Âge	√	√	√
Infraction sexuelles antérieures	√	√	√
Victime de sexe masculin	√	√	√
Victime sans lien de parenté	√	√	√
Délit sexuel sans contact		√	√
Victime inconnue		√	√
Condamnation antérieure		√	√
Violence non sexuelle antérieure		√	√
Violence non sexuelle répertoriée		√	
Cohabitation		√	
Démêlé avec la justice			√
Victime jeune			√
Fréquence des délits sexuels			√
Bris de condition			√
Délit sexuel jeune et adulte			√
Année en liberté avant condamnation actuelle			√

3.3 La Statique-2002

Le troisième prédicteur est la Statique-2002 publié en 2003 [3]. Il comprend quatorze items dont tous ceux de la Statique-99 sauf les variables cohabitation et infractions répertoriées avec violence non sexuelle. Le pointage des items provenant de la Statique-99 n'est pas toujours attribué de la même façon dans la Statique-2002. La construction de la Statique-2002 est basée sur dix échantillons de délinquants sexuels dont sept au Canada, deux aux États-Unis et un au Royaume-Uni. Les quatorze items de la Statique-2002 sont présentés au tableau 2.

La grille détaillée pour l'attribution des pointages est présentée à l'annexe D. Les quatorze items sont classés en cinq sous-sections. Un pointage est associé à chaque item selon les caractéristiques de l'individu. Ensuite, le total de chaque sous-section est converti en pointage partiel. Les cinq pointages partiels sont additionnés pour obtenir le pointage total. Les pointages finaux varient de 0 à 14 et sont convertis en cinq catégories de risque. Un pointage total plus élevé est associé à un risque de récidive plus élevé [3].

4 Préparation des données

La description des variables disponibles a permis de réduire le nombre de variables à considérer pour les différents objectifs. Les raisons qui ont guidé l'élimination de certaines variables sont présentées dans les prochaines sous-sections. De plus, toutes les valeurs manquantes ont reçu le même code.

4.1 Reprise d'information dans les variables

Des variables ont été rejetées parce qu'elles sont en fait une combinaison de plusieurs variables. Par exemple, la variable prisex99 (score lié aux infractions sexuelles antérieures) a été rejetée, car elle dépend de deux autres variables à notre disposition, soit prichsex (nombre de chefs d'accusation pour délit sexuel) et pricvsex (nombre de condamnations pour délit sexuel). Les variables prisoany et proant concernent toutes deux le nombre de peines anté-

rieures toutes causes confondues. Donc, posséder une de ces deux variables est suffisant pour avoir cette information. Dans le cas présent la variable proant est conservée. De même, les variables prisoso, pricvsex et sexcon mesurent le nombre de condamnations antérieures pour infractions sexuelles de l'individu. La variable sexcon, qui contient le nombre de peines avec au moins un chef d'accusation de nature sexuelle, est conservée.

4.2 Nombre d'observations trop faible pour une variable

Pour la variable juvsex2, qui détermine la présence d'arrestation à l'âge mineur pour infraction sexuelle, 77 % des observations sont manquantes. La proportion de réponse est très faible étant donné que les antécédents judiciaires à l'âge mineur ne sont pas accessibles lors de la consultation du dossier criminel. Cette variable a donc été retirée ainsi que juvsex pour laquelle les données manquantes de juvsex2 avaient simplement été remplacées par des 0.

4.3 Création d'une variable

Une version continue de la variable age est créée à partir de la variable age2002 en attribuant la valeur centrale de la classe d'âge de l'individu décrite dans la grille de la Statique-2002 de l'annexe D. L'exemple standard est qu'un individu dans la classe d'âge 18-25 se fera attribuer 21,5 ans. Pour la classe d'âge de 50 ans et plus, la valeur de 65 ans sera attribuée. Ce choix a été fait, car l'espérance de vie pour les hommes en Amérique du Nord est environ 80 ans. L'intérêt de créer cette variable est de pouvoir accéder à d'autres valeurs lors de l'imputation.

5 Imputation

L'imputation est possible si les données manquantes sont réparties aléatoirement (MAR) dans le jeu de données. Le test de Little [4] a été effectué afin de déterminer si la distribution des valeurs manquantes est complètement aléatoire (MCAR), si tel est le cas les données manquantes sont nécessairement MAR. L'hypothèse nulle de ce test a été rejetée significativement (valeur de p <0.0001), donc les données ne sont pas MCAR. Ainsi, à cette étape on ne peut rien conclure à propos de leur statut MAR.

Tableau 3 – Patrons les plus fréquents des valeurs manquantes dans les données non-imputées

Patron	**Variable(s) manquante(s)**	**Fréquence**
1	Uniquement breach	21
2	Uniquement lives	31
3	Toutes les variables suivantes : age25, age2002, lives, prichany, breach, privio99, privio02,indexvio, prichsex, notouch, males, extra, stranger	20

Pour le patron 1 du tableau 3, lorsque la variable breach (bris des conditions de la surveillance communautaire) est manquante, en moyenne la valeur de la variable prichsex (nombre de chefs d'accusation antérieurs pour infractions sexuelles) est plus faible. Dans le patron 2, lorsque la variable lives (cohabitation avec un conjoint pendant au moins 2 ans) est manquante, en moyenne la valeur de la variable pricvsex (nombre de condamnations antérieures pour infractions sexuelles) est plus élevée. Finalement, pour le patron 3, la variable période à risque vaut majoritairement 0, c'est-à-dire que l'individu n'a pas terminé sa peine lors du dernier suivi. Cependant, puisqu'il est très complexe de travailler avec des données manquantes de façon non-aléatoire (MNAR), l'imputation sera tout de même réalisée en supposant qu'elles sont MAR.

Puisque plusieurs variables sont manquantes, l'imputation multiple est réalisée à l'aide de la méthode FCS de la procédure MI du logiciel SAS (version 9.4). Étant donné que toutes les valeurs à imputer doivent être positives, il faut spécifier, avec l'option *MIN*=, une valeur minimale de 0 pour les variables numériques discrètes sauf pour la variable age qui doit valoir au minimum 18 étant donné le contexte.

Les variables utiles à la construction de l'outil ERRRS et Statique-99 ont toutes été imputées. Dans le cas de la Statique-2002, lorsque l'on possède toutes les variables qui constituent un sous-score, ses variables sont imputées et ensuite il est calculé. Lorsque les variables pour créer un sous-score ne sont pas toutes disponibles, alors le sous-score est imputé. Un nombre de dix jeux avec données imputées ont été créés. Les manipulations pour répondre aux objectifs se feront à l'aide de ces dix jeux de données.

6 Définition des outils statistiques utilisés

Tout au long du présent ouvrage, plusieurs statistiques seront utilisées. Dans cette section, les définitions des statistiques en terme de récidive sont présentées. Le tableau suivant sera utilisé comme exemple.

Tableau 4 – Tableau de contingence croisant deux évaluateurs (0 = absence de récidive, 1 = présence de récidive)

		Évaluateur 2	
		0	1
Évaluateur 1	0	a	b
	1	c	d

Les statistiques calculées :

- Spécificité : probabilité que l'évaluateur 2 prédise l'absence de récidive conditionnellement à ce que l'évaluateur 1 prédise l'absence de récidive $\left(\frac{a}{a+b}\right)$
- Sensibilité : probabilité que l'évaluateur 2 prédise une récidive conditionnellement à ce que l'évaluateur 1 prédise la présence de récidive $\left(\frac{d}{c+d}\right)$
- Kappa de Cohen : mesure l'accord entre la prédiction de récidive de l'évaluateur 1 et l'évaluateur 2

Le Kappa de Cohen se calcule ainsi :

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

où

- $P(A)$ représente la proportion d'accord entre les deux évaluateurs $\left(\frac{a+d}{a+b+c+d}\right)$
- $P(E)$ représente la probabilité d'un accord aléatoire $\left(\frac{(a+b)(c+d)+(c+d)(d+b)}{(a+b+c+d)^2}\right)$

Plus la valeur du Kappa de Cohen est grande, plus l'accord entre les deux évaluateurs est bonne. Pour ce faire, dans le logiciel SAS, l'option *agree* de l'énoncé *table* de la procédure FREQ a été utilisée.

7 Objectif 1 : Prédire la récidive

7.1 Modèle de régression logistique

Afin de prédire le risque de récidive chez les délinquants sexuels au Québec, un modèle de régression logistique est ajusté. Cette section présente le modèle mathématique, la méthode de sélection des variables, l'interprétation du meilleur modèle choisi ainsi qu'une discussion sur la performance de ce dernier.

Pour tenir compte de la période à risque de récidive de chacun, cette variable est forcée à être dans le modèle. Elle mesure le temps entre la libération et la fin de l'étude ou jusqu'à une récidive. De plus, avant de faire cette analyse, la forme fonctionnelle entre chaque variable explicative numérique de l'annexe E et la variable réponse, la récidive violente ou sexuelle, a été vérifiée. Pour ce faire, le graphique du logit (logarithme($\frac{p}{1-p}$) où p indique la probabilité de récidive) permet d'étudier le lien entre une variable explicative quelconque et la probabilité de récidive de l'individu. Un exemple où il y a un problème avec la forme fonctionnelle est présenté à la figure 13 de l'annexe F pour la variable age.

Les polynômes fractionnaires d'Altman [5] sont aussi utilisés afin de compléter l'analyse graphique. Cela consiste à imposer différentes transformations à la variable explicative et à analyser si la vraisemblance du lien entre cette variable et la variable réponse s'améliore. Dans un premier temps, il est suggéré d'utiliser les puissances suivantes comme transformation -2,-1,-0.5,0,0.5,1,2,3 où 0 représente le logarithme naturel. Dans le cas présent, les puissances -3 et -4 sont aussi vérifiées afin de bien voir le comportement lorsque le maximum de vraisemblance est atteint à la puissance -1 ou -2. Pour savoir si une transformation sur la variable est requise, il faut faire un test du Khi-deux entre l'absence de transformation et la meilleure. Lorsque le test est significatif, il faut appliquer la meilleure transformation.

Pour la variable age à la figure 1, la meilleure transformation suggérée par les dix jeux avec données imputées est l'exposant -2 ou -1. La transformation de -1 est choisie, car c'est cette transformation qui procure la meilleure linéarisation pour le graphique du logarithme($\frac{p}{1-p}$) présenté à la figure 14de l'annexe F.

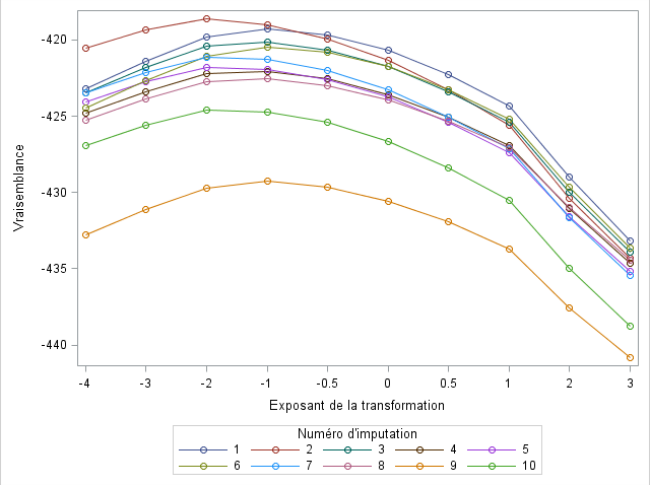


FIGURE 1 – Vraisemblance en fonction de la transformation de la variable age pour chacun des jeux de données imputées afin de déterminer la forme fonctionnelle à utiliser lors de la construction du modèle

En procédant de la même façon pour les autres variables continues, la transformation logarithmique est appliquée pour les variables viocon et noxcon et l'inverse de la racine carrée pour la variable proant. Pour chaque variable, il est possible de choisir une transformation qui n'est pas significativement différente de la meilleure, mais qui est plus facilement interpré-

table, c'est particulièrement le cas avec proant pour laquelle la transformation logarithmique est choisie au lieu de l'inverse de la racine carrée. Les transformations des variables age, viocon, noxcon et proant sont respectivement renommées agem1, log_viocon, log_noxcon et log_proant.

7.1.1 Modèle mathématique

La fonction de lien utilisée est le logit. Le modèle s'écrit comme suit :

$$Y_i \sim Bernoulli(\pi_i) \text{ où } \pi_i = \frac{e^{\beta_0+\beta_1 x_{1,i}+...+\beta_p x_{p,i}}}{1+e^{\beta_0+\beta_1 x_{1,i}+...+\beta_p x_{p,i}}}$$

- Y_i représente la présence de récidive pour l'individu i
- π_i représente la probabilité de récidiver pour l'individu i
- $x_{1,i}, x_{2,i}, ...x_{p,i}$ représentent les variables explicatives mesurées chez l'individu i
- $\beta_1, \beta_2, ...\beta_p$ sont les paramètres du modèle

7.1.2 Description de la sélection de variables

Les variables utilisées pour la construction du modèle sont présentées à l'annexe E. L'unité statistique est l'individu. La variable réponse du modèle qui sera créé est la présence de récidive violente ou sexuelle qui est une variable binaire.

La méthode principale de la sélection de modèle est la méthode du lasso réalisée avec le package glmnet du logiciel R. Le lasso permet d'ajuster une régression logistique sous la contrainte que la somme des coefficients estimés est inférieure au paramètre de régularisation ($\sum|\hat{\beta}| < \lambda$). Cette méthode est un processus efficace pour rejeter les variables les moins importantes en leur accordant un coefficient de 0. La fonction utilisée pour déterminer les coefficients des différentes variables est cv.glmnet de R qui permet de faire de la validation croisée et d'indiquer la contrainte qui minimise la déviance. Ainsi, pour chacun des jeux de données imputées, on possède un ensemble de paramètres estimés sous contrainte. Ces dix ensembles indiquent les variables à conserver dans le modèle. Un résumé de ces 10 ensembles est présenté au tableau 5.

Tableau 5 – Ensembles de variables retenues avec la méthode du lasso pour chacun des dix jeux de données imputées

Variable	**Nombre de fois sur 10 que la variable est retenue**
breach	10
lives	10
agem1	10
log_proant	10
log_viocon	10
période à risque	10
privio99	10
indexvio	8
notouch	8
prichany	6
prichsex	4
Autres variables	0

Dans un premier temps, un modèle qui contient les variables présentes au moins une fois dans le tableau précédent est créé. Il est à noter que certaines variables définies comme importantes dans la prédiction de la récidive en criminologie, comme la variable stranger pour la présence d'au moins une victime inconnue, sont absentes dans les dix imputations. Le modèle global des dix ensembles de jeux de données imputées nous suggère ainsi de conserver onze variables. Afin de voir si ces onze variables sont significatives :

1. Une régression logistique est appliquée sur chacun des dix jeux de données
2. Les dix ensembles de coefficients sont combinés avec la procédure MIANALYZE.
3. À partir de ces résultats on retire la variable la moins significative.

On répète ces trois étapes jusqu'à ce que toutes les variables soient significatives. À la suite de cet algorithme "backward", les variables période à risque, agem1, log_viocon et breach sont retenues. En comparaison, la méthode stepwise de la procédure LOGISTIC de SAS avec le BIC comme critère de sélection [6] procure ces mêmes quatre variables huit fois sur dix. De plus, toutes les interactions possibles entre ces quatre variables ont été testées et seule l'interaction entre les variables breach et agem1 s'est révélée significative. Une étude approfondie de l'ajout de cette interaction est présentée au tableau 6.

Tableau 6 – Statistiques pour comparer l'ajout de l'interaction entre les variables breach et agem1 au modèle de régression logistique final

	Point de coupure	**Sensibilité** (%)	**Spécificité** (%)	**Taux d'erreur** (%)	**Kappa de Cohen**
Modèle sans interaction	0.14	74.35	71.78	27.85	0.2865
	0.15	71.79	74.73	25.68	0.3051
	0.46	28.21	98.32	11.57	0.3570
Modèle avec interaction	0.12	71.79	71.36	29.64	0.2680
	0.14	71.79	76.63	24.95	0.3279
	0.43	35.90	97.05	11.57	0.4082

Un point de coupure optimal est une valeur qui permet à la fois de maximiser la sensibilité et la spécificité et de minimiser le taux d'erreur lors de la prédiction de la récidive avec ce modèle. Selon le tableau 6, le meilleur point de coupure est de 0.14 pour le modèle avec interaction qui procure un taux d'erreur faible, une bonne spécificité et une bonne sensibilité. Par ailleurs, on peut tenter de seulement minimiser le taux d'erreur. Plusieurs points de coupure le minimisent à une valeur de 11.57% pour le modèle avec interaction. De ces points, celui qui permet le meilleur compromis entre la sensibilité et la spécificité se situe à 0.43. Ainsi le modèle final, présenté au tableau 7, sera composé des variables agem1, breach, la période à risque de récidive, log_viocon et de l'interaction entre breach et agem1. La moyenne des aires sous la courbe ROC pour les dix jeux de données imputées est de 0,808.

7.1.3 Interprétation du modèle choisi

Lorsqu'il y a présence de bris des conditions de la surveillance communautaire, un jeune individu a une plus grande probabilité de récidiver, comme présenté à la figure 2.

Ensuite, plus la période à risque est longue, plus la probabilité de récidive d'un individu est élevée. De même, plus l'individu a un grand nombre de peines avec au moins un chef d'accusation de nature violente, plus il est à risque de récidiver. Toutefois, étant donné l'échelle logarithmique pour cette variable, chaque condamnation supplémentaire n'a pas le même impact sur la prédiction de la récidive. Les premières peines avec au moins un chef d'accusation de nature violente font beaucoup augmenter le risque de récidive contrairement à un

Tableau 7 – Paramètres du modèle de régression logistique obtenu pour prédire la récidive violente ou sexuelle chez les délinquants sexuels au Québec

Paramètre	**Valeur estimée**	**Erreur-type**	**Augmentation relative dans la variance (%)**	**Valeur de p**
Ordonnée à l'origine	-4.1201	0.9290	7.68	<0.0001
Breach (Présence de bris de condition)	-1.7162	1.0711	5.18	0.1092
agem1	-0.1238	34.1550	17.02	0.9971
log_viocon (Logarithme du nombre de peines avec au moins un chef d'accusation de nature violente)	0.4120	0.1165	3.12	0.0004
Période à risque (Temps à risque de récidiver en mois)	0.0296	0.0058	1.98	<0.0001
Breach* agem1	91.5009	38.2186	10.92	0.0169

individu qui passe de dix à onze peines avec au moins un chef d'accusation de nature violente.

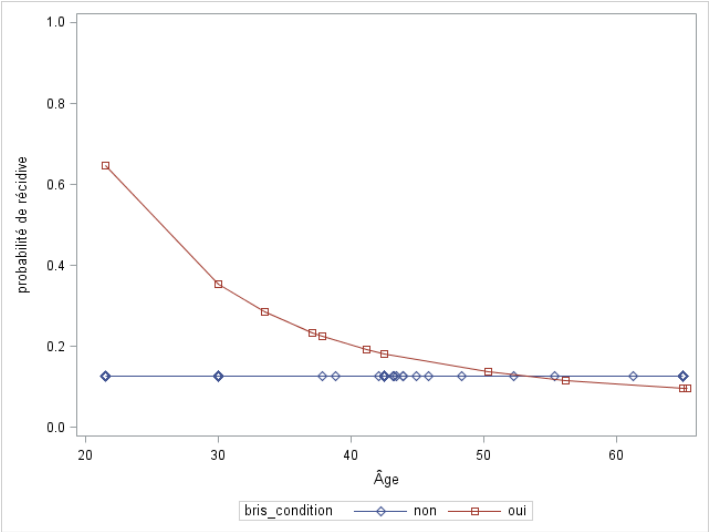


FIGURE 2 – Représentation de l'interaction entre la présence de bris des conditions de la surveillance communautaire et l'âge pour une période à risque de 60 mois et sans condamnation avec au moins un chef d'accusation de nature violente (viocon=0)

Pour valider l'hypothèse que le modèle s'ajuste bien aux données, on utilise le test d'Hosmer et Lemeshow qui est un test du Khi-carré. La statistique de test est obtenue avec l'option

lackfit du logiciel SAS. Une valeur de p inférieure à 0.05 indique que le modèle ne s'ajuste pas bien aux observations. Les valeurs de p pour chacun des ensembles de données imputées sont toujours supérieures à 0.52. Le modèle de régression logistique sélectionné pour ce premier objectif s'ajuste donc bien aux données. De plus, l'interprétation de ses paramètres est cohérente avec ce qui est déjà connu en criminologie.

Un des points de coupure intéressant a été choisi à partir de la figure 3 où le croisement des deux courbes indique le meilleur équilibre entre la sensibilité et la spécificité. Avec ce point de coupure, si la probabilité de récidive obtenue avec le modèle est supérieure à 0.14, alors on prédit une récidive pour l'individu. Différents points de coupure peuvent être intéressants dépendamment du critère que l'on prend. Un résumé des trois points les plus susceptibles d'être choisis sont présentés au tableau 8. Tout compte fait, le modèle présenté dans cette section permet d'identifier les facteurs les plus importants à considérer pour connaître le risque de récidive d'un individu. Avec le modèle choisi, la valeur de 0.35 est utilisée comme seuil pour prédire une récidive.

Tableau 8 – Statistiques des différents points de coupure pour la prédiction de la récidive violente ou sexuelle du modèle de régression logistique

Critère de sélection	**Point de coupure**	**Sensibilité**	**Spécificité**	**Taux d'erreur**	**Kappa de Cohen**
Équilibrer sensibilité et spécificité	0.14	71.79	76.63	24.05	0.3279
Maximiser Kappa de Cohen	0.35	47.43	93.05	13.38	0.4230
Minimiser Taux d'erreur	0.43	35.90	97.05	11.57	0.4082

7.1.4 Précisions sur le modèle de régression logistique

Le modèle de régression sélectionné permet de prédire la récidive seulement si l'on spécifie le temps à risque. Il peut sembler illogique que le modèle contienne cette variable, car contrairement aux autres, on ne connait pas sa valeur lors de la libération de l'individu. On peut songer à travailler avec la période à risque de récidive en la séparant en catégories. Une dichotomisation de la période à risque a été appliquée afin de déterminer si la probabilité de

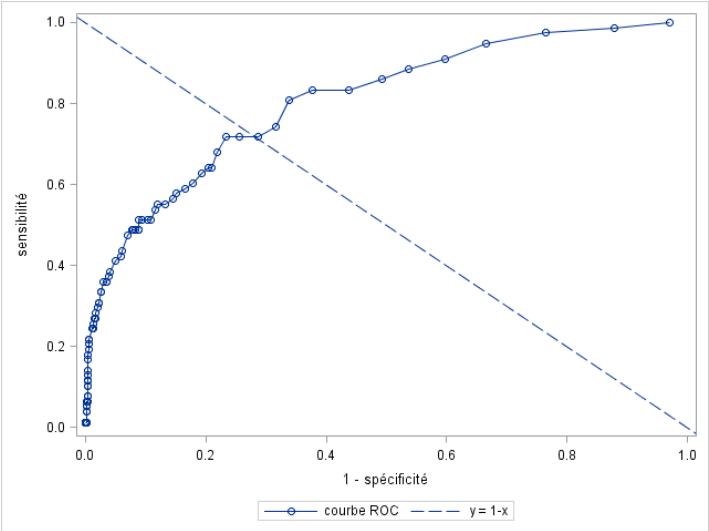


FIGURE 3 – Courbe ROC du modèle de régression logistique

récidive des individus possédant une période à risque inférieure à 48 mois (valeur identifiée par l'arbre de classification, voir section suivante) s'explique par les mêmes variables et interactions que si la période à risque est supérieure à 48 mois. L'interaction entre la période à risque et l'âge s'est alors révélée significative. La moyenne des AIC pour le modèle avec dichotomisation est de 372.30, alors qu'elle est de 363.58 pour le modèle sans dichotomisation. Comme un plus petit AIC indique un meilleur modèle, la dichotomisation ne semble pas améliorer le modèle. Il faut aussi mentionner que sans une dichotomisation de la période à risque, il y a une relation linéaire avec le logit de la probabilité de récidive. Ainsi, conserver le modèle avec la période à risque de récidive sans dichotomisation semble être la meilleure chose à faire. L'avantage de ce modèle est qu'il est possible de déterminer le risque de récidive de l'individu sur une période à risque au choix.

7.2 Arbre de classification

Il existe d'autres façons de répondre à cet objectif. Une d'entre elles consiste à construire un arbre de classification à l'aide de la fonction rpart du package portant le même nom en R [7]. La figure 4 montre l'arbre de classification obtenu pour prédire la récidive violente ou sexuelle en fonction de toutes les variables explicatives de l'annexe E.

On y retrouve les mêmes quatre variables que dans le modèle de régression logistique.

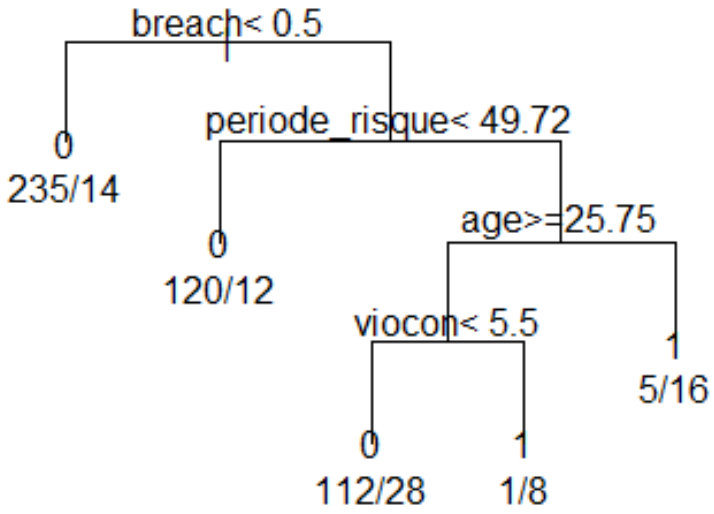


FIGURE 4 – Arbre de classification obtenu pour la prédiction de la récidive chez les délinquants sexuels au Québec(0= absence de récidive, 1=présence de récidive).

Lorsque la condition est respectée, on se dirige dans la branche de gauche. À la fin de chaque branche on retrouve un classement, l'individu est prédit récidiviste (valeur de 1) ou non récidiviste (valeur de 0). De cela, les bonnes classifications et les mauvaises classifications peuvent être obtenues. Pour chaque fin de branche, le chiffre de gauche indique le nombre d'individus qui en réalité n'ont pas récidivé et le chiffre de droite, le nombre d'individus qui en réalité ont récidivé. Par exemple, pour la variable breach, si un individu n'a pas commis de bris des conditions de la surveillance communautaire, alors l'arbre classe l'individu comme non-récidiviste correctement 235 fois et commet une mauvaise classification 14 fois. La sensibilité et la spécificité peuvent être calculée à l'aide du tableau 9 et elles sont respectivement de 30.77 % (24/78) et de 98.32 % (467/475). Dans ce tableau, les deux observations de moins qu'avec la régression logistique s'expliquent par l'absence de valeur pour la variable réponse pour deux individus. Dans le modèle de régression, ces valeurs manquantes avaient été imputées, mais ici la fonction rpart n'impute que les variables explicatives [8]. Toutefois dans l'éventualité de comparer deux modèles, l'arbre est utilisé pour donner une valeur à la variable réponse pour ces deux individus afin d'avoir un total de 553 prédictions. Il faut prendre en compte que cet ajustement ajoute deux bonnes classifications à l'arbre et aura pour effet d'améliorer ses statistiques. Le taux d'erreur obtenu par validation croisée sur la population

totale séparée en k groupes est d'environ 14 % (tableau 10).

Puisque l'arbre prend en compte la période à risque, encore une fois lors de la prédiction il est possible de spécifier la période à risque de récidive, c'est-à-dire de choisir l'horizon de la période à risque lors de la prise de décision.

Tableau 9 – Tableau de fréquence croisant la récidive réelle d'un délinquant sexuel et la récidive prédite par l'arbre de classification (0= absence de récidive, 1=présence de récidive)

		Récidive prédite		
		0	**1**	**Total**
Récidive réelle	**0**	467	6	473
	1	54	24	78
	Total	521	30	551

7.3 Comparaison des deux modèles issus de l'objectif 1

Puisque deux outils ont été développés pour répondre à l'objectif 1, il serait intéressant de les comparer avant la confrontation avec les outils déjà existants. Le taux d'erreur pour l'arbre de classification présenté au tableau 10 est obtenu par validation croisée sur la population totale séparée en k groupes. Les variables nécessaires au calcul de cette statistique sont la récidive réelle et la récidive prédite par l'outil, deux variables binaires.

Tableau 10 – Statistiques des deux outils créés à l'objectif 1 pour prédire la récidive violente ou sexuelle

Modèle	**Régression logistique (point de coupure=0.35)**	**Arbre de classification**
Sensibilité(%)	47.43	31.58
Spécificité (%)	93.05	98.68
Taux d'erreur (%)	13.38	14.63 (k=5) 14.05 (k=10)
Kappa de Cohen	0.4230	0.3970

Pour poursuivre à l'objectif 2, l'arbre de classification sera utilisé puisque les statistiques de prédiction des modèles sont comparables et qu'il est plus facile d'utiliser l'arbre.

7.4 Complément de l'objectif 1 : Modèle d'analyse de survie

Par le contexte du projet, il y a présence de données censurées à droite, c'est-à-dire que le suivi est terminé, mais qu'il est possible que des individus n'aient pas encore récidivé. En guise de complément aux deux modèles de ce premier objectif, un modèle d'analyse de survie est ajusté sur les mêmes variables que ces modèles.

7.4.1 Sélection de la distribution dans le modèle de survie

Le modèle ajusté est un modèle paramétrique et il est obtenu avec la procédure LIFEREG de SAS. Différentes distributions des termes d'erreur sont testées à tour de rôle dans l'option *dist*= de l'énoncé *model* de cette procédure. La distribution choisie est celle qui procure la meilleure vraisemblance. D'après le tableau 11, la distribution logistique est celle qui sera considérée pour le modèle de survie.

Tableau 11 – Vraisemblances pour différentes distributions des erreurs dans le modèle d'analyse de survie ajusté avec la procédure LIFEREG

Distribution mentionnée dans l'énoncé dist	**-2 log vraisemblance**
Weibull	899.2
lognormale	944.3
loglogistique	911.4
gamma	893.7
exponentielle	1024.2
normale	893.8
logistique	890.1

7.4.2 Modèle mathématique

Le modèle de survie pour la prédiction du temps avant une récidive violente ou sexuelle est :

$$Y_i = x_i'\beta + \sigma\varepsilon_i$$

avec $\varepsilon_i \sim$ logistique

- Y_i temps entre la libération et la récidive violente ou sexuelle pour l'individu i
- x_i' vecteur des variables explicatives pour l'individu i
- β vecteur des paramètres du modèle
- σ paramètre d'échelle
- ε_i terme d'erreur pour l'individu i

La fonction de survie, qui est la probabilité qu'un individu ne commette pas de récidive pendant y mois est $S(y) = 1 - F(y)$ où $F(y)$ est la fonction de répartition de la distribution des temps entre la libération et la récidive violente ou sexuelle. Pour le modèle de survie décrit précédemment, la fonction de survie est

$$S(y) = \frac{1}{1 + e^{(y - x^t\beta)/\sigma}}$$

7.4.3 Interprétation du modèle

En ajustant la distribution logistique aux termes d'erreur, on obtient les paramètres du modèle suivant :

Tableau 12 – Paramètres du modèle de survie pour prédire le temps (en mois) avant une récidive violente ou sexuelle chez les délinquants sexuels au Québec

Paramètre	**Estimation**	**Erreur-type**	**Valeur de p**
Ordonnée à l'origine	118.83	13.87	<0.0001
age	-0.1643	0.2814	0.5594
viocon	-3.4145	1.1820	0.0039
breach	-51.8102	16.343	0.0015
age* breach	1.0033	0.3728	0.0071
paramètre d'échelle	14.4256	1.2290	

Le modèle indique que plus il y a de peines avec au moins un chef d'accusation de nature violente (viocon), plus le temps avant qu'un individu récidive sera court. Si l'individu a commis un bris des conditions de la surveillance communautaire et qu'il est jeune, le temps avant qu'il récidive a plus de chance d'être court. Si l'individu est âgé et qu'il a déjà commis un bris de condition, le temps avant qu'il récidive aura tendance à être plus long. On peut aussi observer à la figure 5 l'interaction entre ces deux variables. Lorsqu'il n'y a pas de bris des conditions de surveillance communautaire (droite rouge), la pente négative n'est pas significative (valeur de p de 0.5594). Donc selon ce modèle, s'il n'y a pas de bris de condition, il n'y a pas de lien entre l'âge et le temps avant une récidive violente ou sexuelle. Pour obtenir cette figure, la valeur de la variable viocon a été fixée à la médiane de la distribution de ses valeurs soit 1.

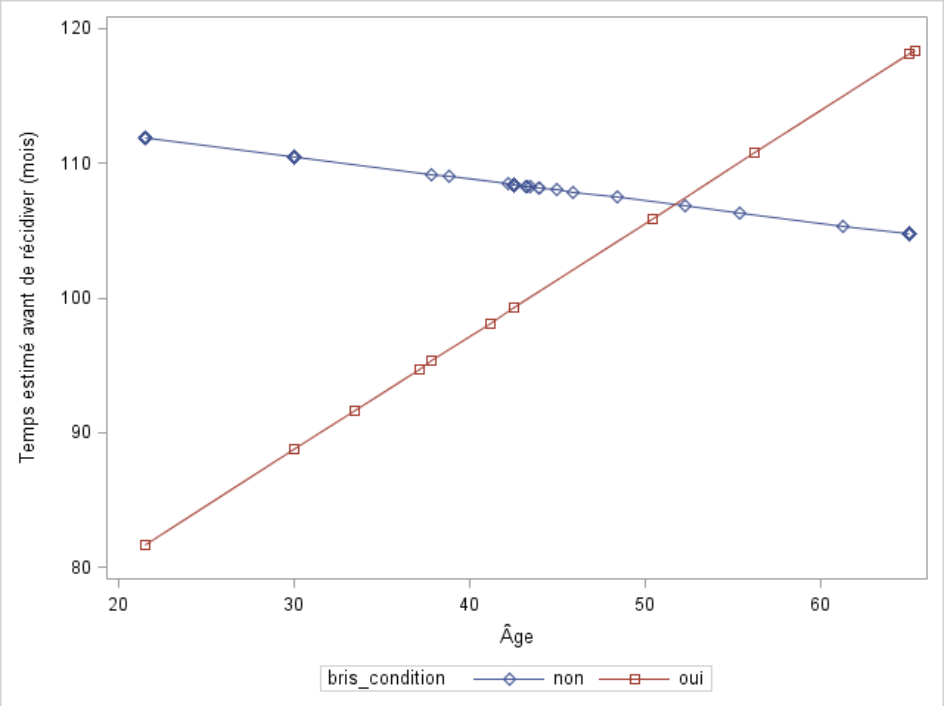


FIGURE 5 – Représentation du temps avant qu'un individu commette une récidive violente ou sexuelle estimé par le modèle d'analyse de survie (viocon = 1)

8 Objectif 2 : Validation des outils actuels et de l'outil créé pour prédire le risque de récidive

Plusieurs outils ont été développés pour estimer le risque de récidive à long terme (plus de cinq ans) et peuvent permettre de mettre sur pied certaines stratégies de traitement et de surveillance des individus [9]. Le deuxième objectif du projet est de vérifier la validité de trois outils déjà existants ainsi que l'outil choisi à l'objectif 1.

8.1 Précisions sur la méthodologie utilisée

Dans le premier objectif, les dix jeux de données imputées sont utilisés. Dans le présent objectif, une agrégation de ces dix jeux de données imputées est utilisée afin de calculer les scores des différents outils. Pour les variables binaires, une valeur agrégée inférieure ou égale à 0.4 est arrondie à 0, une valeur supérieure ou égale à 0.6 est arrondie à 1 et si la valeur est de 0.5, elle est aléatoirement ajustée à 0 ou 1.

Il faut rappeler que les variables nécessaires pour calculer le score de la Statique-2002 ne sont pas toutes disponibles. Donc, pour cet outil, certains sous-scores sont calculés à partir des variables qui le composent si toute ses variables sont disponibles. Sinon, le sous-score est imputé. Les sous-scores imputés sont ceux pour la criminalité générale et pour la persistance des infractions sexuelles.

Après ces ajustements, pour chacun des outils, il faut déterminer le niveau de risque à partir duquel une récidive est prédite. Pour ce faire, le seuil choisi pour un outil sera celui qui maximise le Kappa de Cohen.

8.2 ERRRS

L'aire sous la courbe ROC de l'ERRRS, présentée à la figure 8 de l'annexe B, est de 0.4835. Il est à noter qu'il n'y a pas de catégories de niveau de risque pour cet outil. Le score du point de coupure qui maximise le Kappa de Cohen est 5, procurant une valeur de 0.0118

pour cette statistique. L'accord est mauvais parce que cet outil est conçu et utilisé pour prédire la récidive sexuelle uniquement.

Tableau 13 – Statistiques de la prédiction pour chaque score possible pour l'ERRRS

Score du point de coupure	Sensibilité	Spécificité	Taux d'erreur	Kappa de Cohen
1	0.9231	0.0105	0.8608	-0.0190
2	0.6026	0.3789	0.5895	-0.0077
3	0.3333	0.6695	0.3779	0.0018
4	0.1667	0.8442	0.2513	0.0104
5	0.0897	0.9200	0.1971	0.0118
6	0.0256	0.9642	0.1682	-0.0148

8.3 La Statique-99

L'aire sous la courbe ROC de la Statique-99, présentée à la figure 10 de l'annexe C, est de 0.6093. Le point de coupure qui procure le meilleur Kappa de Cohen est le 3^e niveau de risque pour un Kappa de Cohen de 0.1102. Donc tous ceux pour qui l'outil indique un score supérieur ou égal à 4 sont prédits récidivistes.

Tableau 14 – Statistiques de la prédiction pour chaque niveau de risque pour la statique-99

Niveau de risque du point de coupure (Score)	Sensibilité	Spécificité	Taux d'erreur	Kappa de Cohen
Faible (0-1)	1	0	0.8590	0
Faible-modéré (2-3)	0.9102	0.1600	0.7342	0.0227
Modéré-élevé (4-5)	0.6410	0.5726	0.4177	0.1102
Élevé (6 et plus)	0.2435	0.8358	0.2477	0.0720

8.4 La Statique-2002

L'aire sous la courbe ROC de la Statique-2002, présentée à la figure 12 de l'annexe D, est de 0.6239. Le point de coupure qui procure le meilleur Kappa de Cohen est le 4^e niveau de risque ce qui donne un Kappa de Cohen de 0.1333. Donc tous ceux pour qui l'outil indique un score supérieur ou égal à 7 sont prédits récidivistes.

Tableau 15 – Statistiques de la prédiction pour chaque niveau de risque pour la statique-2002

Niveau de risque du point de coupure (Score)	Sensibilité	Spécificité	Taux d'erreur	Kappa de Cohen
Faible (0-2)	1	0	0.8590	0
Faible-moyen (3-4)	0.8077	0.3663	0.5714	0.0687
Moyen (5-6)	0.6154	0.5705	0.4231	0.0962
Moyen-élevé (7-8)	0.3718	0.7958	0.2640	0.1333
Élevé (9 et plus)	0.1410	0.9137	0.1953	0.0636

8.5 Discussion et comparaison des outils

Des trois outils existants, la Statique-2002 est celle qui possède le meilleur Kappa de Cohen. Selon les points de coupure choisis, c'est aussi cet outil qui procure le meilleur taux d'erreur. De plus, il comporte la plus grande aire sous la courbe ROC. Il est intéressant de remarquer que les outils sont de plus en plus performants.

Tableau 16 – Statistiques des outils de prédiction de la présence de récidive analysés

Outil	Sensibilité	Spécificité	Taux d'erreur	Kappa de Cohen (outil vs réalité)
ERRRS	0.0897	0.9200	0.1971	0.0118
Statique-99	0.6410	0.5726	0.4177	0.1102
Statique-2002	0.3718	0.7958	0.2640	0.1333
Arbre de classification	0.3158	0.9868	0.1197	0.3017

Comme l'arbre de classification est le modèle retenu à l'objectif 1, il reste à comparer les prédictions de l'arbre aux prédictions de la Statique-2002. Le Kappa de Cohen entre ces deux outils est 0.0831, ce qui est faible. La spécificité, le taux d'erreur et le Kappa de Cohen sont meilleurs pour l'arbre de classification, donc c'est cet outil qui sera utilisé à l'objectif 3.

Tableau 17 – Tableau de fréquences croisant les récidives prédites par l'arbre de classification et par la Statique-2002 (0= absence de récidive, 1= présence de récidive)

		Prédiction Statique-2002		
		0	**1**	**Total**
Prédiction de l'arbre	**0**	409	113	522
	1	18	13	31
	Total	427	126	553

9 Objectif 3 : Prédiction de la récidive pour les différents profils de délinquants sexuels

Le chercheur qui chapeaute le projet s'intéresse aux différents profils de délinquants sexuels. Certains profils ont déjà été dressés dans une des récentes publications de M. Lussier [10]. La tâche pour ce troisième objectif consiste à identifier les divers profils de délinquants sexuels au Québec indépendamment de ce qui a déjà été fait. Par la suite, il sera question de vérifier si certains outils sont plus précis pour un profil d'individus de la population à l'étude.

9.1 Classification des délinquants sexuels

9.1.1 Méthodologie

Pour regrouper les individus qui ont des caractéristiques similaires, la classification hiérarchique ascendante sera utilisée afin de déterminer le nombre de groupes. La classification hiérarchique ascendante consiste au départ à considérer chaque individu comme étant un groupe, pour ensuite fusionner itérativement les deux groupes les plus semblables jusqu'à l'obtention d'un seul groupe contenant tous les individus. Cette façon de faire permet de former des groupes d'individus qui sont à la fois semblables aux individus du même groupe et différents des individus des autres groupes.

La classification est réalisée à l'aide des variables utilisées à l'objectif 1 (annexe E), c'est-à-dire que l'on ne prend pas en compte les variables des scores et sous-scores des outils pour créer les différents groupes. Il faut tenir compte du fait qu'il n'existe pas de façon unique pour regrouper les observations. Les différentes méthodes de classification n'utilisent pas la même définition de distance entre les observations pour créer des groupes. Outre le fait que l'on désire obtenir un petit nombre de groupes et des effectifs raisonnablement grands, il n'y a pas de contraintes sur les groupes à obtenir.

Avant même de classer les individus, il faut éviter qu'une variable ayant une grande variabilité possède un grand poids lors de la classification. Ainsi, il faut s'assurer de standardiser

les variables. Pour ce faire, des composantes principales sont créées. Les composantes principales sont des variables orthogonales entre elles créées par des combinaisons linéaires des variables disponibles. Elles permettront d'expliquer la variation dans les données tout en étant indépendantes et standardisées.

La procédure PRINCOMP de SAS est utilisée pour obtenir les composantes principales. Ensuite, elles sont utilisées pour les différentes méthodes de classification hiérarchique ascendante. Avec la procédure CLUSTER, cinq méthodes qui calculent différemment les distances ont été spécifiées dans l'énoncé *method*=. Les critères utilisés pour déterminer le nombre de groupes à conserver, après que les observations aient été regroupées, sont la statistique du pseudo-t^2, la statistique du pseudo-F et le *cubic clustering criterion* (ccc). Les méthodes du centroïde, de la moyenne et de la médiane suggèrent de former quatre groupes. Ces trois méthodes ont l'avantage d'être robustes aux données aberrantes en formant toutefois des groupes généralement de tailles différentes. Les méthodes de Ward, qui est sensible au données aberrantes, et du plus proche voisin, qui permet de créer des groupes de forme très irrégulière, n'indiquent pas clairement un certain nombre de groupes.

Pour créer les groupes, la procédure FASTCLUS est utilisée avec l'option *maxc=4* qui force la procédure à former quatre groupes. Il s'agit du nombre proposé de groupes par la majorité des méthodes dans la procédure antérieure. La procédure FASTCLUS effectue une classification non-hiérarchique des individus, plus précisément en utilisant la méthode des k-moyennes [11]. Ensuite, on s'intéresse aux variables qui permettent de faire la discrimination entre les groupes. La procédure CANDISC permet ensuite d'analyser l'utilité individuelle de chacune des variables avec l'option *anova*. De cela, les variables pour lesquelles l'anova est significative sont conservées pour poursuivre l'analyse des caractéristiques des groupes.

Dans chacun des groupes, il y a un certain nombre de récidivistes présentés au tableau 18. Un test du Khi-deux a été réalisé afin de savoir si les proportions des récidivistes dans chacun des groupes sont les mêmes avec l'option *exact pchi* de la procédure FREQ. La valeur de p étant <0.0001, on rejette l'hypothèse que les proportions de récidivistes dans chacun des

groupes sont égales. D'une certaine façon le classement d'un individu dans un certain groupe est une méthode non paramétrique qui peut aider à prédire le risque de récidive chez un individu.

Tableau 18 – Groupes formés par la classification non-hiérarchique de la procédure FASTCLUS

Groupe	**Effectif**	**Récidiviste (% du groupe)**
1	74	4 (5.4)
2	151	5 (3.3)
3	117	17 (14.5)
4	211	52 (24.7)

9.1.2 Interprétation des différents groupes créés

Pour chacune des composantes canoniques, il est possible de déterminer les variables qui lui sont les plus reliées. Pour ce faire, il faut interpréter les coefficients canoniques normalisés et groupés intra-classe obtenus par la sortie SAS de la procédure CANDISC[12].

À l'aide du tableau 19, dans lequel se trouvent toutes les variables significatives dans une anova à un facteur pour prédire le groupe, il est possible de déterminer les caractéristiques associées à chaque composante canonique. Un score élevé sur la première composante implique que l'individu a beaucoup de peines antérieures, beaucoup de condamnations antérieures pour violence non sexuelle et peu de peines avec chef d'accusation de nature non-violente ou non-sexuelle. Un score élevé sur la deuxième composante implique que l'individu a au moins une victime de sexe masculin, a peu de peines antérieures et peu de chefs d'accusation antérieurs pour infractions sexuelles. Pour la troisième composante, un score élevé indique la présence de démêlées antérieurs avec le système de justice pénale et peu de condamnations antérieures pour violence non sexuelle. Ces trois composantes canoniques permettent de distinguer quatre groupes dans la population (figure 6 de la page 30).

Tableau 19 – Classement des variables discriminantes pour les composantes canoniques (CC)

CC1		CC2		CC3	
Variables	**Coefficient**	**Variables**	**Coefficient**	**Variables**	**Coefficient**
proant	0,820	males	1,089	prichany	0,716
privio99	0,757	proant	-0,665	privio99	-0,409
noxcon	-0,475	prichsex	-0,422	breach	0,329
sexcon	-0,344	noxcon	0,401	sexcon	0,271
breach	0,318	sexrate	-0,313	sexrate	0,219
males	-0,298	privio99	0,306	notouch	0,193
prichany	0,192	viocon	0,147	indexvio	-0,131
notouch	-0,115	prichany	0,143	viocon	-0,078
prichsex	0,092	extra	0,109	stranger	0,072
sexrate	0,071	twolt12	0,099	extra	0,066
stranger	0,067	age	0,081	age	-0,043
extra	0,066	indexvio	-0,076	prichsex	0,031
anyviosex	0,058	stranger	-0,054	males	0,029
indexvio	-0,050	notouch	0,047	twolt12	0,029
viocon	0,049	anyviosex	0,040	proant	-0,023
twolt12	-0,034	breach	0,036	anyviosex	0,019
age	-0,014	sexcon	0,022	noxcon	-0,005

Le groupe 1 se distingue par un faible score sur la première composante et un score élevé sur la deuxième composante. Ainsi les individus du groupe 1 ont peu de peines antérieures toutes causes confondues, ont presque tous au moins une victime de sexe masculin et ont peu de condamnations pour violence non-sexuelle.

Le groupe 2 se distingue par un faible score sur les deuxième et troisième composantes. Les individus de ce groupe n'ont généralement pas de victime de sexe masculin, ont peu de peines antérieures, peu de condamnations pour violence non sexuelle et n'ont pas de démêlées antérieurs avec le système de justice pénale. On peut dire qu'il s'agit d'un groupe d'individus qui en sont généralement à leur première condamnation.

Le groupe 3 se distingue par un faible score sur la deuxième composante et un score élevé sur la troisième composante. Les individus de ce groupe ont beaucoup de peines antérieures, un très grand nombre de chefs d'accusation antérieurs pour infractions sexuelles, ont presque tous déjà eu des démêlés avec le système de justice pénale et plusieurs d'entre eux ont un bris de condition de la surveillance communautaire.

Le groupe 4 se distingue presque uniquement par un score élevé sur la première composante canonique. Les individus de ce groupe ont un très grand nombre de peines antérieures, ont au moins une condamnation antérieure pour violence non sexuelle, ont presque tous des bris de conditions, ont peu de peines avec au moins un chef d'accusation de nature sexuelle. Donc, ce sont des individus qui commettent beaucoup de crimes dont quelques-uns qui sont à connotation sexuelle.

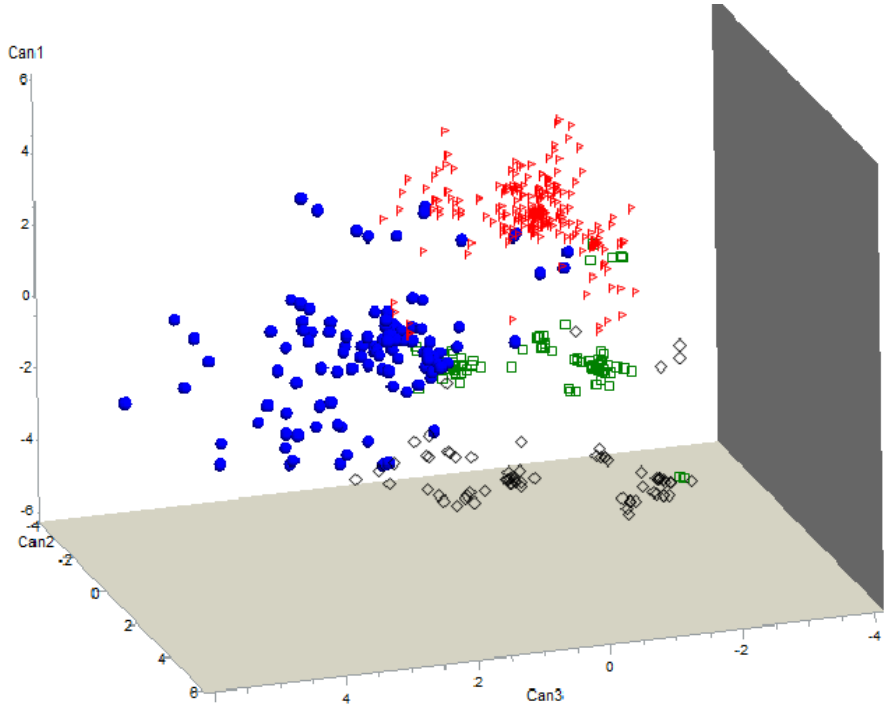


FIGURE 6 – Représentation graphique des différents groupes de délinquants sexuels sur les trois composantes canoniques. Les losanges noirs sont les individus du groupe 1, les carrés verts du groupe 2, les ronds bleus du groupe 3 et les drapeaux rouges du groupe 4

Pour visualiser dans quelle mesure les groupes se distinguent sur les variables males, proant et prichsex, consulter l'annexe G.

9.2 Performance des outils sur chacun des groupes

Pour chaque profil et pour chaque outil, un tableau de contingence des récidives prédites et réelles est construit, puis le Kappa de Cohen est calculé pour mesurer la capacité des outils à prédire la récidive. Le Kappa de Cohen est choisi, puisqu'il permet de minimiser le taux d'erreur en tenant compte du déséquilibre entre le nombre de récidivistes et de non-récidivistes.

9.2.1 Groupe 1

Les spécificités et les taux d'erreur sont significativement différents entre eux, mais les sensibilités ne sont pas différents au seuil de 5 % à cause du nombre de récidiviste trop faible. Pour ce groupe, le meilleur outil est l'arbre de classification, car il procure le meilleur Kappa de Cohen.

Tableau 20 – Statistiques à comparer pour chacun des outils analysés pour prédire la présence de récidive dans le groupe 1

Outil	**Sensibilité**	**Spécificité**	**Taux d'erreur**	**Kappa de Cohen (outil vs réalité)**
Régression logistique	0.25	0.9857	0.0541	0.3084
ERRRS	0	0.9000	0.1486	-0.0739
Statique-99	0.25	0.5571	0.4595	-0.0449
Statique-2002	0	0.9000	0.1486	-0.0739
Arbre de classification	0.25	1	0.0405	0.3867

9.2.2 Groupe 2

Les spécificités et les taux d'erreur sont significativement différents entre eux, mais les sensibilités ne sont pas différents au seuil de 5 % à cause du nombre de récidiviste trop faible.

Pour ce groupe, les meilleurs outils sont l'arbre de classification et le modèle de régression, car ils procurent le plus faible taux d'erreur et le meilleur Kappa de Cohen.

Tableau 21 – Statistiques à comparer pour chacun des outils analysés pour prédire la présence de récidive dans le groupe 2

Outil	**Sensibilité**	**Spécificité**	**Taux d'erreur**	**Kappa de Cohen (outil vs réalité)**
Régression logistique	0.20	1	0.0265	0.3259
ERRRS	0	1	0.0331	0
Statique-99	0	0.9658	0.0662	-0.0342
Statique-2002	0	0.9452	0.0861	-0.0425
Arbre de classification	0.20	1	0.0265	0.3259

9.2.3 Groupe 3

Pour ce groupe, les spécificités et les taux d'erreur sont significativement différents entre eux, mais les sensibilités ne sont pas différents au seuil de 5 % à cause du nombre de récidiviste relativement faible.Pour ce groupe, le meilleur outil est la régression logistique, car elle procure le meilleur Kappa de Cohen et un taux d'erreur comparable à celui de l'arbre de classification.

Tableau 22 – Statistiques à comparer pour chacun des outils analysés pour prédire la présence de récidive dans le groupe 3

Outil	**Sensibilité**	**Spécificité**	**Taux d'erreur**	**Kappa de Cohen (outil vs réalité)**
Régression logistique	0.4706	0.940	0.1282	0.4430
ERRRS	0.3529	0.730	0.3248	0.0596
Statique-99	0.5882	0.430	0.5470	0.0082
Statique-2002	0.4706	0.600	0.4188	0.0402
Arbre de classification	0.2941	0.980	0.1197	0.3626

9.2.4 Groupe 4

Pour ce groupe, au seuil de 5%, les sensibilités, les spécificités et les taux d'erreur sont significativement différentes entre eux. Pour ce groupe, le meilleur outil est la régression logistique, le Kappa de Cohen est le meilleur car le taux d'erreur n'est pas le plus faible, mais le compromis sensibilité-spécificité est meilleur.

Tableau 23 – Statistiques à comparer pour chacun des outils analysés pour prédire la présence de récidive dans le groupe 4

Outil	**Sensibilité**	**Spécificité**	**Taux d'erreur**	**Kappa de Cohen (outil vs réalité)**
Régression logistique	0.5182	0.8365	0.2417	0.3534
ERRRS	0.0192	0.9748	0.2607	-0.0085
Statique-99	0.750	0.3082	0.5829	0.0357
Statique-2002	0.7358	0.4038	0.3460	0.1304
Arbre de classification	0.3462	0.9748	0.1801	0.3983

9.2.5 Discussion sur l'objectif

Pour tous les groupes, selon le Kappa de Cohen, l'arbre de classification et la régression logistique sont comparables entre eux et meilleurs que les outils actuels. Les outils ne semblent pas en mesure de mieux prédire la récidive pour un groupe en particulier.

10 Discussion et conclusion

Ce projet avait comme but de prédire la récidive des délinquants sexuels au Québec. L'intérêt de créer un modèle indépendamment des outils actuellement utilisés lors de processus judiciaires était de confirmer l'importance des variables utilisées par les outils actuels.

En raison de la présence de données manquantes pour certains individus, l'imputation multiple a été réalisée pour obtenir dix jeux de données imputés. Il faut garder en tête que les résultats obtenus pour les modèles de l'objectif 1 dépendent de ces imputations. En imputant

à nouveau, de légers changements peuvent survenir sans toutefois apporter de modifications majeures aux modèles obtenus. Le fait que tous les individus de la population à l'étude aient commis leur délit au Québec pourrait compromettre les performances des modèles créés s'ils sont appliqués à d'autres populations. En effet, les libérés faisant partie de l'étude évoluent au sein d'une même société, donc ils partagent vraisemblablment des caractéristiques que des individus provenant d'ailleurs n'auraient pas. Par ailleurs, l'âge est particulièrement reconnue pour son importance dans la prédiction de la récidive. Dans le projet, la variable age est traitée comme étant continue, mais puisqu'elle a été créée à partir d'une variable catégorique elle ne prend généralement que quatre valeurs. Donc, ne pas connaître de façon précise les valeurs de cette variable a pu limiter les performances des modèles.

De plus, les modèles créés reposent sur une seule population qui est utilisée à la fois pour leur construction et pour la comparaison avec les autres outils qui eux sont construits à partir d'autres populations. C'est donc sans surprise qu'à l'objectif 2 ce sont les modèles créés à l'objectif 1 qui paraissent meilleurs. Il serait intéressant de comparer les divers outils avec une population n'ayant pas servi à la conception d'aucun d'entre eux. En procédant de cette façon, la comparaison serait moins biaisée.

Idéalement, la variable réponse des modèles aurait été la présence de récidive sexuelle. Par contre, le nombre insuffisant de récidivistes sexuels (27 récidivistes sexuels sur 553 individus) a imposé d'utiliser la récidive violente ou sexuelle afin d'obtenir un nombre raisonnable de récidivistes (78 récidivistes violents ou sexuels sur 553 individus). Procéder de cette façon permet d'aller chercher les individus ayant commis des crimes à la fois sexuels et violents qui ont pu faire retirer les chefs d'accusation de nature sexuelle lors de négociations pendant leur procès. Il ne faut pas oublier que 7% des individus n'ont pas accepté de participer à l'étude. Ce sont des individus qui ont potentiellement des caractéristiques particulières qui auraient pu influencer les modèles créés.

Pour terminer, même avec un nombre d'observations relativement faible, le projet a permis de confirmer la pertinence de plusieurs variables des outils actuels pour prédire la récidive

violente ou sexuelle. Plusieurs analyses ont été réalisées, mais il aurait aussi été intéressant d'utiliser un modèle de survie semi-paramétrique afin de traiter de la meilleure façon possible la période de suivi.

Bibliographie

[1] P. Lussier and J. Healey. Rediscovering quetelet, again : The "aging" offender and the prediction of reoffending in a sample of adult sex offenders. *Justice Quaterly*, 2009.

[2] K. Hanson and D. Thornton. Static 99 : Improving actuarial risk assessments for sex offenders. *Repéré sur le site du ministère de la Sécurité publique du Gouvernement du Canada : http ://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/sttc-mprvng-actrl/index-en.aspx*, 1999.

[3] K. Hanson and D. Thornton. Notes on the development of static-2002. *Repéré sur le site du ministère de la Sécurité publique du Gouvernement du Canada : http ://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/nts-dvlpmnt-sttc/index-en.aspx*, 2003.

[4] RJA Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 1998.

[5] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regressionm, 2nd edition*. John Wiley et Sons, Inc., 2000.

[6] K. P. Burnham and D. R Anderson. Multimodel inference : understanding aic and bic in model selection. *Sociological Methods & Research*, 2004.

[7] T. Therneau, B. Atkinson, and B. Ripley. *Rpart : Recursive Partitioning and Regression Trees*, 2015. R package version 4.1-10.

[8] T. M. Therneau and E. J. Atkinson. An introduction to recursive partitioning using the rpart routines, 29 juin 2015.

[9] A. Harris, A. Phenix, K. Hanson, and D. Thornton. Statique-99 : Règles de codage révisées - 2003. *Repéré sur le site du ministère de la Sécurité publique du Gouvernement du Canada : http ://www.securitepublique.gc.ca/cnt/rsrcs/pblctns/vldty-sttc-99/index-fr.aspx*, 2003.

[10] P. Lussier and G. Davies. A person oriented perspercitive on sexual offenders, offending trajectories and risk of recidivism. *Psychology, Public Policy, and Law*, 2011.

[11] SAS Institute. *SAS/STAT® 9.2 User's Guide. Second Edition*. Cary, Caroline du Nord, É-U : SAS Institute Inc, 2009.

[12] C. J. Huberty. *Applied discriminant analysis*. John Wiley et Sons, Inc. New York, 1994.

[13] K. Hanson and G. Davies. The developpement of a brief actuarial scale for sexual offense recidivism. *Repéré sur le site du ministère de la Sécurité publique du Gouvernement du Canada : http ://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/dvlpmnt-brf-ctrl/index-en.aspx*, 1997.

[14] A. Phoenix, D. Doren, L.Helmus, K. Hanson, and D. Thornton. Règles de codage pour l'échelle statique-2002. *Repéré sur le site du ministère de la Sécurité publique du Gouvernement du Canada : http ://www.securitepublique.gc.ca/cnt/rsrcs/pblctns/sttc-2002/sttc-2002-fra.pdf*, 2002.

Annexe A : Liste complète des variables

Tableau 24 – Variables disponibles pour l'étude

Variable	Explication en français	Type de la variable
age25	Moins de 25 ans à la libération	Dichotomique
age2002	Âge à la libération codé pour la Statique-2002	Catégorique
breach	Bris des conditions de la surveillance communautaire	Dichotomique
crim6	Score brut de la criminalité générale	Numérique discrète
crim3	Sous-total de la criminalité générale	Numérique discrète
devsex	Sous-total pour les intérêts sexuels déviants	Numérique discrète
education1	Études secondaires terminées	Dichotomique
ethnic	Origine ethnique (0 pour caucasien, 1 sinon)	Dichotomique
extra	Présence d'au moins une victime sans lien de parenté	Dichotomique
indexvio	Présence de condamnations répertoriées avec violence non sexuelle	Dichotomique
juvsex	Présence d'arrestation à l'adolescence pour une infraction sexuelle (données manquantes changées pour «non»)	Dichotomique
juvsex2	Présence d'arrestation à l'adolescence pour une infraction sexuelle (variable originale)	Dichotomique
lives	Cohabitation avec un conjoint pendant deux années consécutives	Dichotomique
males	Présence d'au moins une victime de sexe masculin dans le passé	Dichotomique
notouch	Présence de prononcé de peine pour des infractions sexuelles sans contact	Dichotomique
noxcon	Nombre de peines sans chef d'accusation de nature violence ou sexuelle	Numérique discrète
période à risque	Durée de la période de suivi (mois)	Numérique continue
persist5	Score brut de la persistance des infractions sexuelles	Numérique discrète

Variable	Explication en français	Type de la variable
persist3	Sous-total de la persistance des infractions sexuelles	Numérique discrète
prichany	Présence de démêlés antérieurs avec le système de justice pénale	Dichotomique
prichsex	Nombre de chefs d'accusation antérieurs pour infractions sexuelles	Numérique discrète
pricvsex	Nombre de condamnations antérieures pour infractions sexuelles	Numérique discrète
prior4	Quatre prononcés de peine antérieurs ou plus	Dichotomique
prisex99	Score lié aux infractions sexuelles antérieures pour Statique-99	Numérique discrète
prisex02	Score lié aux infractions sexuelles antérieures pour Statique-2002	Numérique discrète
prisoany	Nombre de prononcés de peine antérieurs	Numérique discrète
prisoso	Nombre de condamnations antérieures pour infractions sexuelles	Numérique discrète
privio99	Condamnation antérieure pour violence non sexuelle pour la Statique-99	Dichotomique
privio02	Condamnation antérieure pour violence non sexuelle pour la Statique-2002	Dichotomique
proant	Nombre de peines antérieures	Numérique discrète
rrasor	Score pour l'ERRRS	Numérique discrète
victims	Sous-total pour la relation avec les victimes	Numérique discrète
sexcon	Nombre de peines avec au moins un chef d'accusation de nature sexuelle	Numérique discrète
sexrate	Fréquence des infractions sexuelles - Un prononcé de peine ou plus tous les 15 ans	Dichotomique
static99	Score pour la Statique-99	Numérique discrète
static02	Score pour la Statique-2002	Numérique discrète
stranger	Présence d'au moins une victime qui était un inconnu	Dichotomique
twolt12	Présence de victimes jeunes sans lien de parenté avec l'individu	Dichotomique
viocon	Nombre de peines avec au moins un chef d'accusation de nature violente	Numérique discrète
anyviosex	Récidive violente ou sexuelle (variable réponse)	Dichotomique

Annexe B : Grille d'évaluation ERRRS

Rrasor rating form

Risk Factor	**Score**
Prior sex offenses (not including index offenses)	
None	0
1 conviction ; 1-2 charges	1
2-3 convictions ; 3-5 charges	2
4 convictions or more ; 6 or more charges	3
Age at release (current age)	
25 or more	0
Less than 25	1
Victim gender	
Only females	0
Any males	1
Relationship to victim	
Only related	0
Any non-related	1

FIGURE 7 – Grille d'évaluation pour l'ERRRS [13]

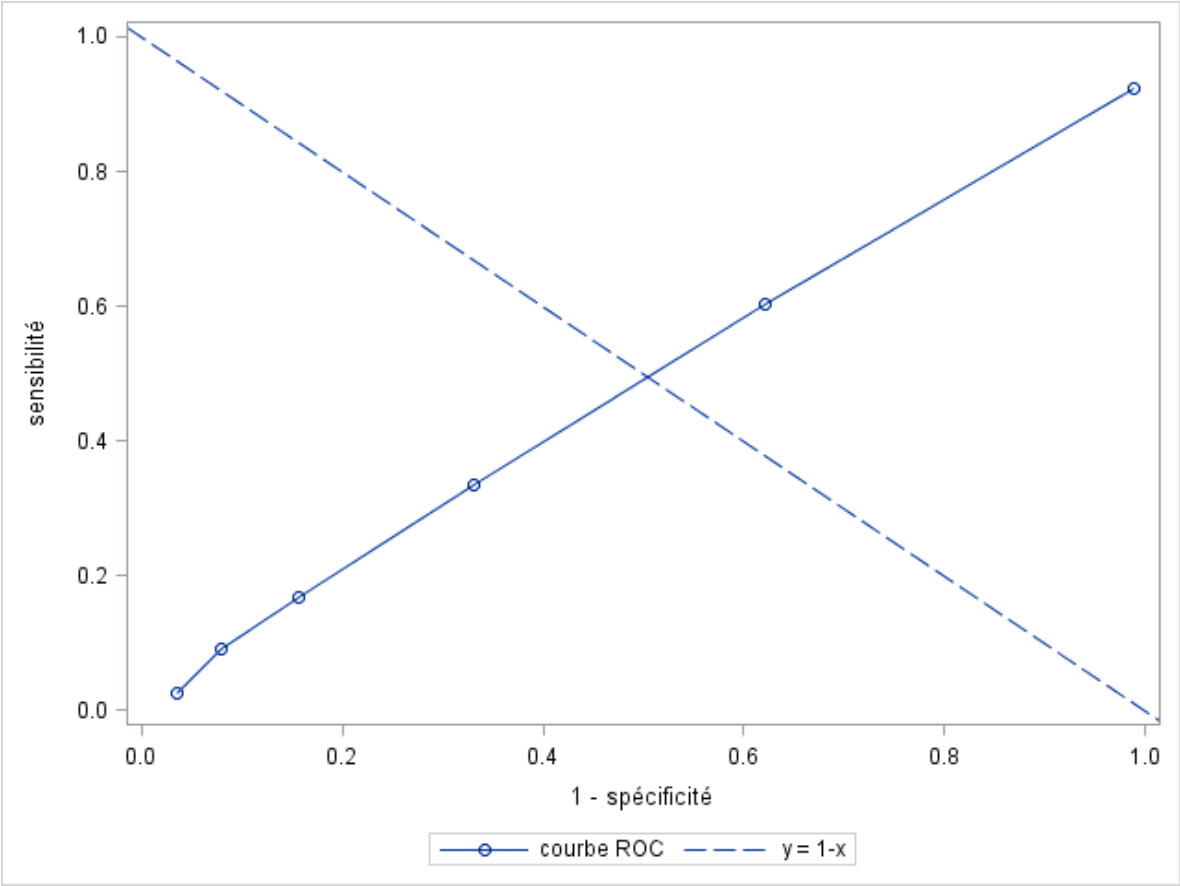


FIGURE 8 – Courbe Roc de l'ERRRS

Annexe C : Grille d'évaluation Statique-99

Annexe cinq

Formulaire de cotation de la Statique-99

Facteur	Facteur de risque	Codes	Score
1	Jeune (S9909)	25 ans ou plus 18 à 24,99 ans	0 1
2	Cohabitation (S9910)	Ce délinquant a-t-il cohabité avec un partenaire intime pendant au moins deux ans? Oui Non	 0 1
3	Infractions répertoriées avec violence non sexuelle Condamnations seulement (S9904)	Non Oui	0 1
4	Infractions antérieures avec violence non sexuelle Condamnations seulement (S9905)	Non Oui	0 1
5	Infractions sexuelles antérieures (S9901)	Accusations / Condamnations Aucune / Aucune 1-2 / 1 3-5 / 2-3 6+ / 4+	 0 1 2 3
6	Prononcés de peine antérieurs (sauf celui visant l'infraction répertoriée) (S9902)	3 ou moins 4 ou plus	0 1
7	Condamnations pour infractions sexuelles sans contact (S9903)	Non Oui	0 1
8	Au moins une victime sans lien de parenté avec le délinquant (S9906)	Non Oui	0 1
9	Au moins une victime qui était un inconnu (S9907)	Non Oui	0 1
10	Au moins une victime de sexe masculin (S9908)	Non Oui	0 1
	Score total	**Faire la somme des scores obtenus pour les différents facteurs de risque**	

CONVERSION DES SCORES OBTENUS SELON LA STATIQUE-99 EN CATÉGORIES DE RISQUE

Score	Catégorie de risque
0, 1	Faible
2, 3	Faible-modéré
4, 5	Modéré-élevé
6 et plus	Élevé

FIGURE 9 – Grille d'évaluation pour la Statique-99 et barème de conversion des pointages [9]

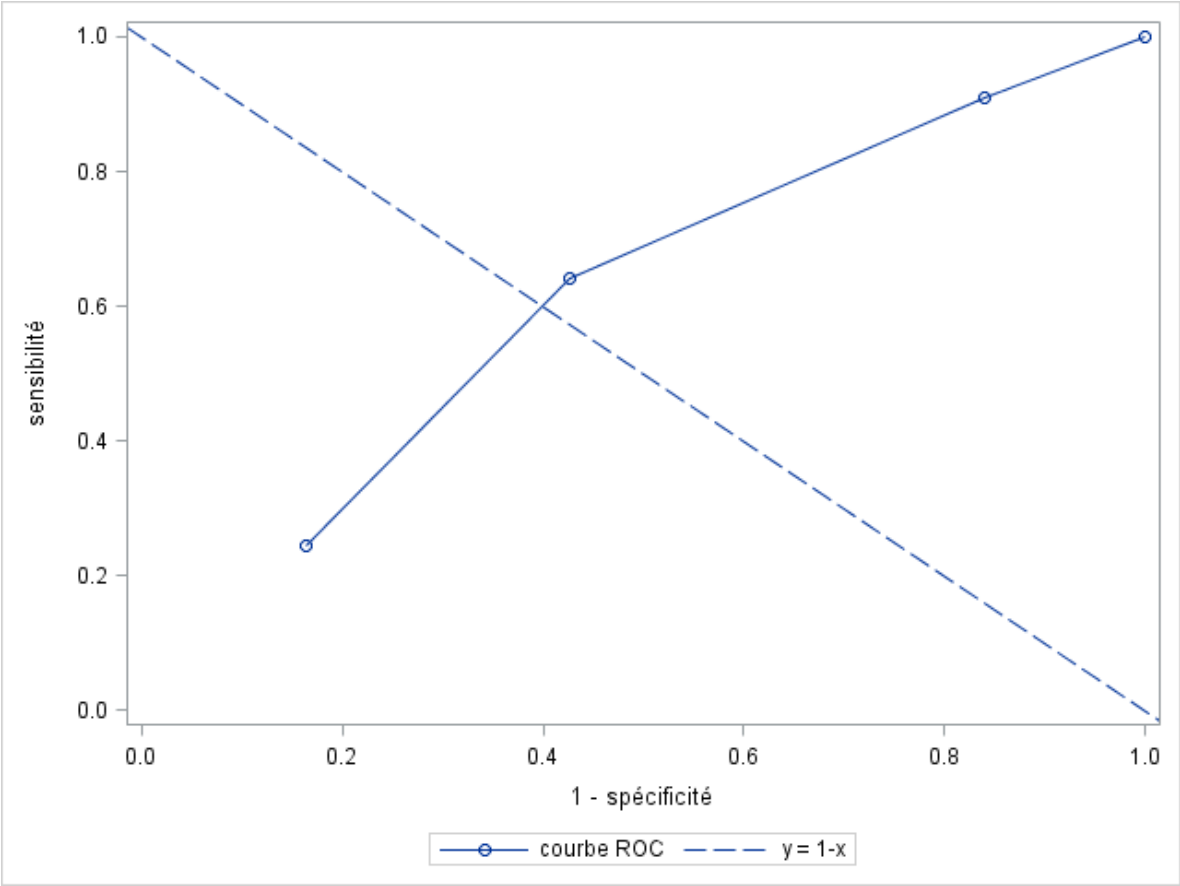
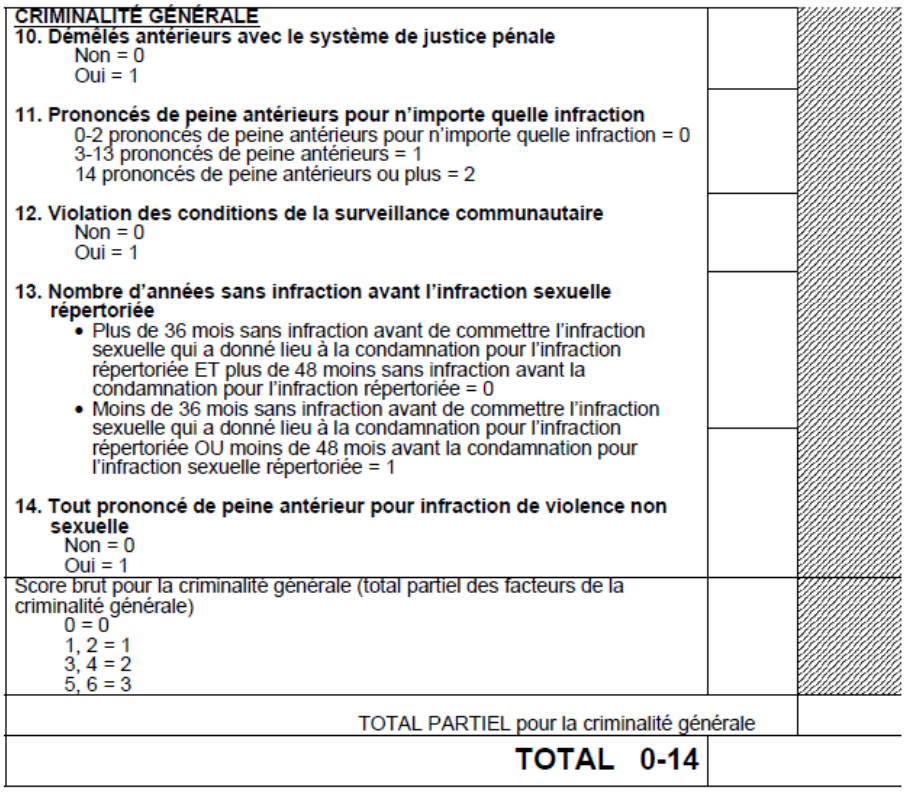


FIGURE 10 – Courbe Roc de la Statique-99

Annexe D : Grille d'évaluation Statique-2002

Annexe I
Formulaire de codage de la Statique-2002

CODAGE DE LA STATIQUE-2002		
FACTEURS	Score brut	Score partiel
ÂGE **1. Âge à la mise en liberté** 50 ans ou plus = 0 35 à 49,9 ans = 1 25 à 34,9 ans = 2 18 à 24,9 ans = 3		
PERSISTANCE DES INFRACTIONS SEXUELLES **2. Prononcés de peine antérieurs pour infractions sexuelles** Aucune date de prononcé de peine pour infractions sexuelles = 0 1 = 1 2, 3 = 2 4 ou plus = 3 **3. Toute arrestation à l'adolescence pour une infraction sexuelle et condamnation à l'âge adulte pour une infraction sexuelle distincte** Aucune arrestation pour une infraction sexuelle avant l'âge de 18 ans = 0 Arrestation avant l'âge de 18 ans et condamnation après l'âge de 18 ans = 1 **4. Fréquence des infractions sexuelles** Moins d'un prononcé de peine tous les 15 ans = 0 Un prononcé de peine ou plus tous les 15 ans = 1		
Score brut pour la persistance (total partiel des infractions sexuelles) 0 = 0 1 = 1 2, 3 = 2 4, 5 = 3		
SCORE PARTIEL pour la persistance des infractions sexuelles		
INTÉRÊTS SEXUELS DÉVIANTS **5. Tout prononcé de peine pour infractions sexuelles sans contact** Non = 0 Oui = 1 **6. Toute victime de sexe masculin** Non = 0 Oui = 1 **7. Victimes jeunes sans lien de parenté avec le délinquant:** Il n'y a pas plus de deux victimes de moins de 12 ans, dont l'une sans lien de parenté = 0 Plus de deux victimes de moins de 12 ans, dont l'une doit être sans lien de parenté = 1		
TOTAL PARTIEL pour les intérêts sexuels déviants		
RELATION AVEC LES VICTIMES **8. Toute victime sans lien de parenté avec le délinquant** Non = 0 Oui = 1 **9. Toute victime inconnue** Non = 0 Oui = 1		
TOTAL PARTIEL pour la relation avec les victimes		



CRIMINALITÉ GÉNÉRALE

10. Démêlés antérieurs avec le système de justice pénale
Non = 0
Oui = 1

11. Prononcés de peine antérieurs pour n'importe quelle infraction
0-2 prononcés de peine antérieurs pour n'importe quelle infraction = 0
3-13 prononcés de peine antérieurs = 1
14 prononcés de peine antérieurs ou plus = 2

12. Violation des conditions de la surveillance communautaire
Non = 0
Oui = 1

13. Nombre d'années sans infraction avant l'infraction sexuelle répertoriée
- Plus de 36 mois sans infraction avant de commettre l'infraction sexuelle qui a donné lieu à la condamnation pour l'infraction répertoriée ET plus de 48 moins sans infraction avant la condamnation pour l'infraction répertoriée = 0
- Moins de 36 mois sans infraction avant de commettre l'infraction sexuelle qui a donné lieu à la condamnation pour l'infraction répertoriée OU moins de 48 mois avant la condamnation pour l'infraction sexuelle répertoriée = 1

14. Tout prononcé de peine antérieur pour infraction de violence non sexuelle
Non = 0
Oui = 1

Score brut pour la criminalité générale (total partiel des facteurs de la criminalité générale)
0 = 0
1, 2 = 1
3, 4 = 2
5, 6 = 3

TOTAL PARTIEL pour la criminalité générale

TOTAL 0-14

Traduction des scores de la Statique-2002 en catégories de risque

0-2 = faible risque
3-4 = risque faible à moyen
5-6 = risque moyen
7-8 = risque moyen à élevé
9+ = risque élevé

FIGURE 11 – Grille d'évaluation pour la Statique-2002 et barème de conversion des pointages [14])

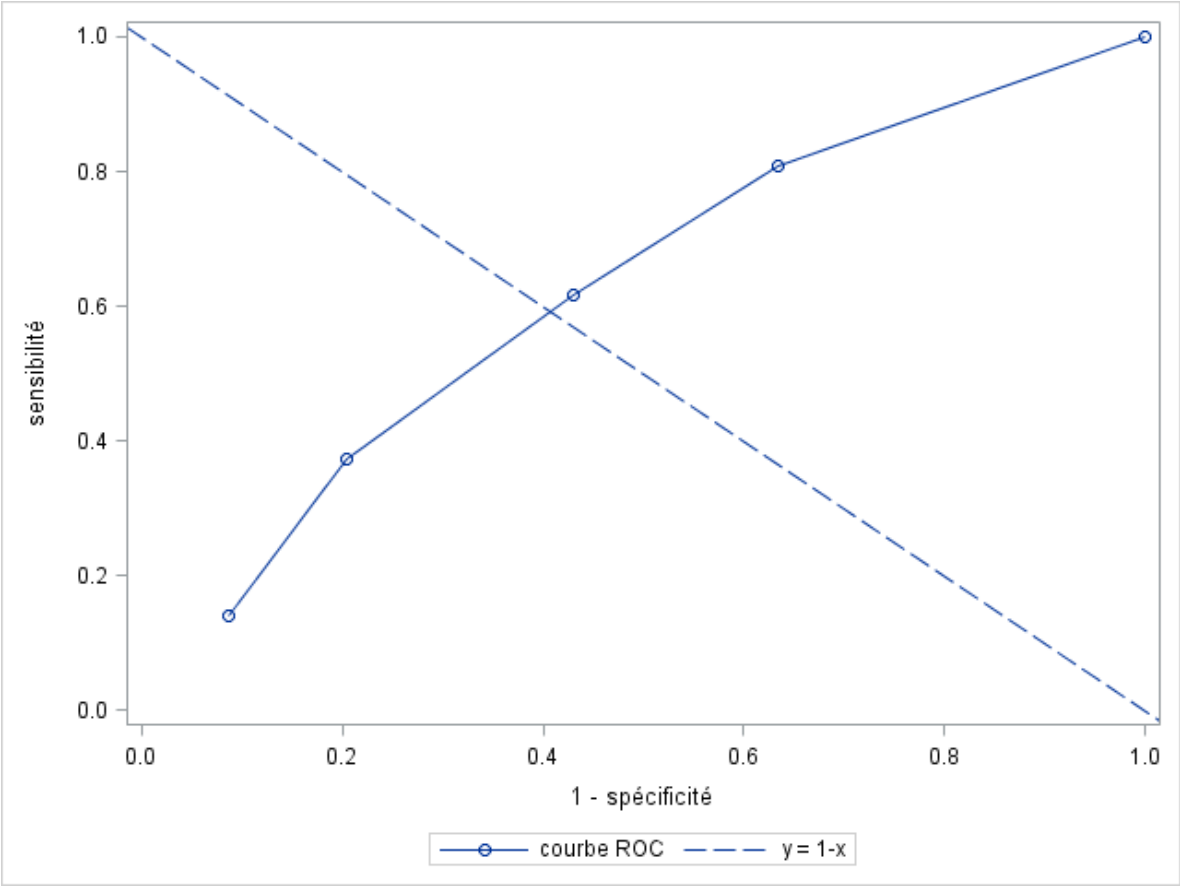


FIGURE 12 – Courbe Roc de la Statique-2002

Annexe E : Variables utilisées pour l'objectif 1

Tableau 25 – Variables disponibles contenant de l'information unique pour la prédiction de la récidive violente ou sexuelle

Variable	Explication en français	Type de la variable
age	Âge à la libération (variable créée)	Numerique continue
breach	Bris des conditions de la surveillance communautaire	Dichotomique
education1	Études secondaires terminées	Dichotomique
ethnic	Origine ethnique (0 pour caucasien, 1 sinon)	Dichotomique
extra	Présence d'au moins une victime sans lien de parenté	Dichotomique
indexvio	Présence de condamnations répertoriées avec violence non sexuelle	Dichotomique
lives	Cohabitation avec un conjoint pendant deux années consécutives	Dichotomique
males	Présence d'au moins une victime de sexe masculin dans le passé	Dichotomique
notouch	Présence de prononcé de peine pour des infractions sexuelles sans contact	Dichotomique
noxcon	Nombre de peines sans chef d'accusation de nature violente ou sexuelle	Numérique discrète
période à risque	Durée de la période de suivi (mois)	Numérique continue
prichany	Présence de démêlés antérieurs avec le système de justice pénale	Dichotomique
prichsex	Nombre de chefs d'accusation antérieurs pour infractions sexuelles	Numérique discrète
privio99	Condamnation antérieure pour violence non sexuelle pour la Statique-99	Dichotomique
proant	Nombre de peines antérieures	Numérique discrète
sexrate	Fréquence des infractions sexuelles - Un prononcé de peine ou plus tous les 15 ans	Dichotomique
sexcon	Nombre de peines avec au moins un chef d'accusation de nature sexuelle	Numérique discrète
stranger	Présence d'au moins une victime qui était un inconnu	Dichotomique
twolt12	Présence de victimes jeunes sans lien de parenté avec l'individu	Dichotomique
viocon	Nombre de peines avec au moins un chef d'accusation pour violence	Numérique discrète

Annexe F : Transformation de la forme fonctionnelle

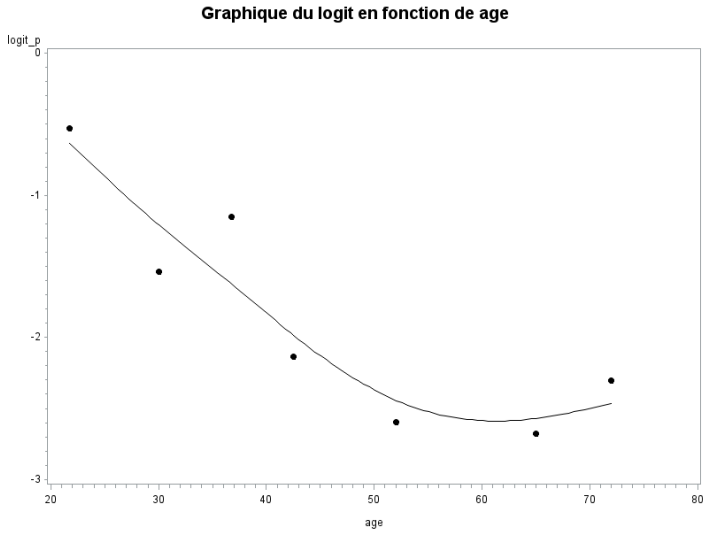


FIGURE 13 – Graphique du logit(p) en fonction de la variable explicative age non transformée où p est la proportion des individus qui ont récidivé

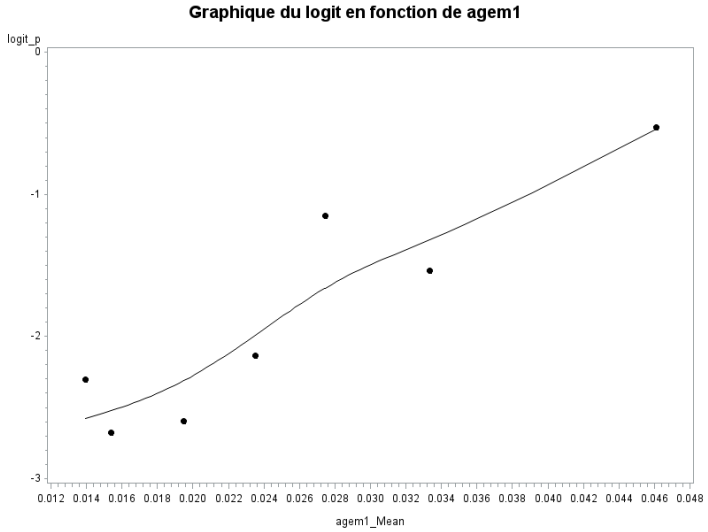


FIGURE 14 – Graphique du logit(p) en fonction de la variable explicative agem1 (Inverse de la variable age) où p est la proportion des individus qui ont récidivé

Annexe G : Représentations graphiques des groupes selon certaines variables

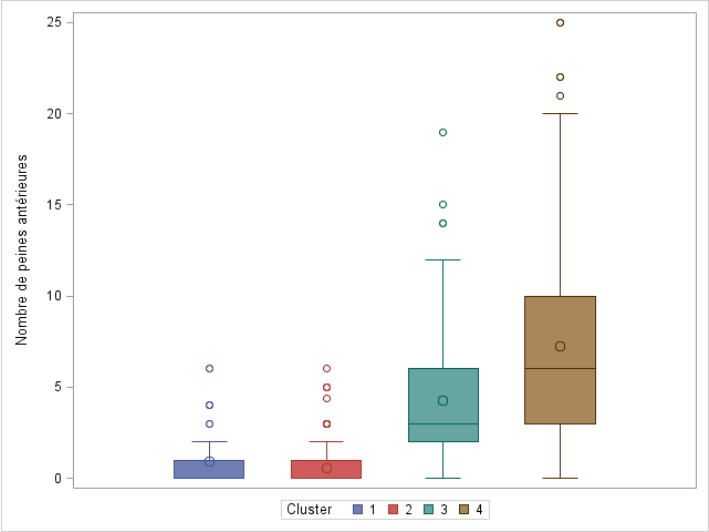


FIGURE 15 – Répartition du nombre de peines antérieures toutes causes confondues

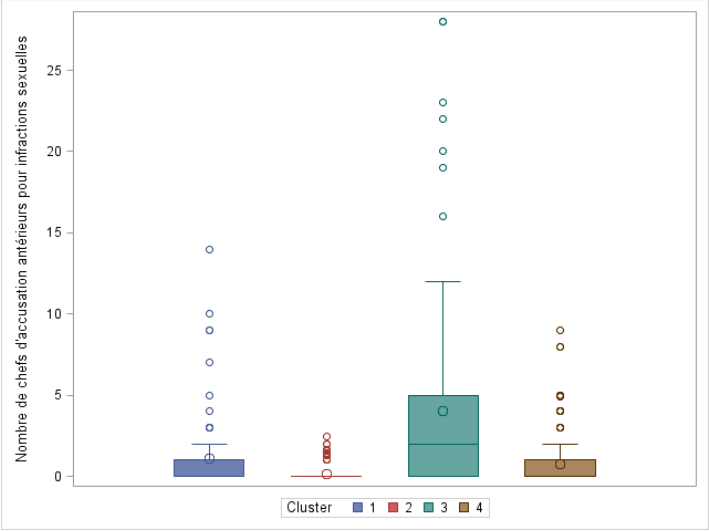


FIGURE 16 – Répartition du nombre de chefs d'accusation antérieurs pour infraction sexuelles

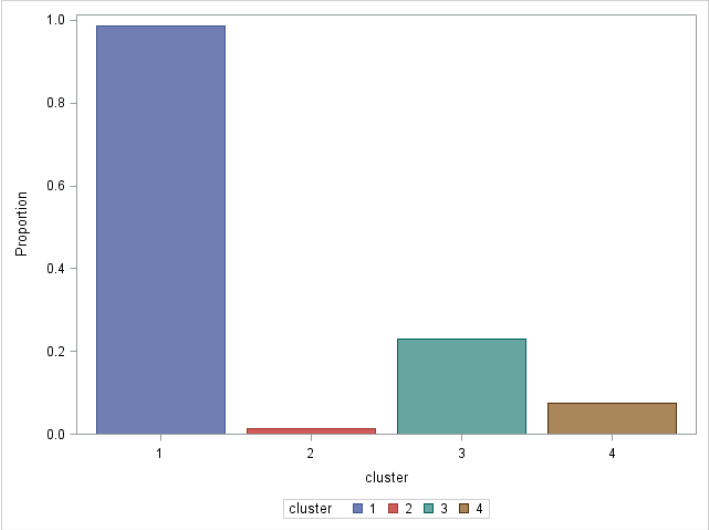


FIGURE 17 – Proportion d'individus ayant au moins une victime de sexe masculin