Voir la rétention en fonction des risques et non des taux : Une autre façon pour les établissements de gérer le maintien des effectifs

Rapport préparé par Mark Conrad et Katherine Morris, Université York
pour le Conseil ontarien de la qualité de l'enseignement supérieur

Conseil ontarien
de la qualité de
l'enseignement supérieur
Un organisme du gouvernement de l'Ontario

Avertissement :

Les opinions exprimées dans ce rapport de recherche sont celles des auteurs et ne reflètent pas nécessairement le point de vue ou les politiques officielles du Conseil ontarien de la qualité de l'enseignement supérieur ou d'autres agences ou organismes qui ont offert leur soutien, financier ou autre, à ce projet.

Se référer au présent document comme suit :

Conrad, M et Morris, K. (2010) *Voir la rétention en fonction des risques et non des taux : Une autre façon pour les établissements de gérer le maintien des effectifs*. Toronto : Conseil ontarien de la qualité de l'enseignement supérieur.

Publié par :

Le Conseil ontarien de la qualité de l'enseignement supérieur

1, rue Yonge, bureau 2402
Toronto (Ontario) Canada
M5E 1E5
Téléphone : 416 212-3893
Télécopieur : 416 212-3899
Site Web : www.heqco.ca
Courriel : info@heqco.ca

© Imprimeur de la Reine pour l'Ontario, 2010

Table des matières

Résumé **2**

Ce qui plaide en faveur de l'adoption d'une nouvelle approche **6**

La persévérance globale et la rétention au niveau des établissements 6

Les taux de rétention des établissements sur le plan des taux bruts 8

Les taux de rétention des établissements sous forme de taux naturels 11

Une autre façon de voir la rétention : en fonction des risques et non des taux .. **17**

Aperçu de l'étude pilote : Estimation des risques d'interruption des études et prédiction des événements interrupteurs à l'Université York 18

Annexe A : Détails de l'étude pilote **26**

Données 26

Méthodes d'analyse 30

Résultats 34

Bibliographie **43**

Résumé

Par le passé, les mots « persévérance » et « rétention » (dans l'expression « rétention des étudiants », équivalente à « maintien des effectifs ») servaient de façon plus ou moins interchangeable à décrire le *fait* pour les étudiants de *rester* dans un cursus donné, en général au sein d'un même établissement et parfois dans le cadre d'un même programme d'études. Depuis quelques années, l'interprétation courante de persévérance dans ce contexte a évolué dans le sens de la *capacité* des étudiants de *poursuivre* leurs études postsecondaires et, à terme, d'obtenir un diplôme, sans égard à leurs éventuels changements de programmes ou même d'établissements ni à leurs absences temporaires mais complètes du milieu postsecondaire, le cas échéant. Or, l'idée fait son chemin en Ontario et ailleurs au Canada qu'une telle perspective globale de la persévérance aidera les gouvernements et les établissements à gérer un système d'éducation postsecondaire bien intégré et hautement fonctionnel, au sein duquel les étudiants peuvent se prévaloir d'un vaste choix de cheminements d'études les plus variés et d'itinéraires vers la réussite.

Ce serait toutefois une erreur de penser que ces préoccupations globales constituent l'angle principal sous lequel il convient d'aborder les résultats scolaires au niveau postsecondaire. En réalité, la notion de persévérance en tant que processus qui permet aux étudiants de surmonter des obstacles n'est intéressante que dans le contexte de décisions initiales de quitter un établissement donné et de ne pas y retourner. Le but principal d'une université quelle qu'elle soit devrait être d'améliorer le maintien de ses propres effectifs. Il est indispensable que toute stratégie visant à globalement rehausser les taux de persévérance mise avant tout et toujours sur l'amélioration des résultats *locaux* en matière de rétention des étudiants. Il y va de l'intérêt véritable des gouvernements et des universités de se doter de moyens d'évaluer l'efficacité des mesures de rétention mises en œuvre par les divers établissements, de les comparer et de les mettre en commun pour concrètement renforcer les résultats qu'elles produisent.

Malheureusement, deux méthodes couramment utilisées pour calculer le taux de rétention – celle des taux bruts et celle des taux naturels – présentent de graves défauts et ne peuvent guère être recommandées aux établissements d'enseignement postsecondaire de l'Ontario comme outils pour gérer leurs mesures de rétention.
La méthode des taux bruts est manifestement inadéquate. Le principal problème avec les taux bruts est qu'ils correspondent essentiellement à des mesures des résultats qui ne tiennent aucun compte des différences au niveau des intrants. Un établissement qui limite ses admissions à des étudiants très bien préparés sur les plans scolaire, financier et culturel à la vie universitaire qu'il leur propose peut s'attendre à être récompensé par des taux de réussite relativement élevés, et cela sans avoir à innover ni à trop investir dans ses mesures de rétention. Évaluer l'efficacité des mesures de rétention à la lumière des taux bruts favorise les établissements qui peuvent se permettre d'éliminer des risques potentiels liés à la rétention dès leur processus d'admission.
Une autre méthode répandue consiste à calculer le taux de rétention en établissant les écarts entre les taux bruts et les taux « escomptés » ou « naturels », puis en fondant les évaluations et les comparaisons sur ces écarts. Les taux de rétention naturels des établissements correspondent à la moyenne des probabilités estimatives de survenue

d’un événement (p. ex., toujours faire partie des effectifs après un an, obtenir son diplôme en l’espace de quatre ans) pour chaque membre d’une cohorte d’étudiants au sein d’un établissement. Un élément clé des modèles statistiques sur lesquels les estimations de probabilité sont fondées est le fait qu’il s’agit de modèles globaux, qui regroupent des données provenant de l’ensemble des établissements étudiés et qui aboutissent à un seul jeu de coefficients modèles appliqués à tous les établissements. Un autre aspect important des estimations de probabilité est qu’elles s’appuient sur des variables explicatives qui tiennent en général compte des seules caractéristiques préadmission des étudiants et parfois de certaines caractéristiques externes, telles que la taille de l’établissement, le domaine d’études et la question de savoir si l’établissement est une université urbaine accueillant en majorité des étudiants qui résident hors de son enceinte. Un établissement dont le taux brut excède son taux naturel est considéré obtenir de bons résultats en matière de rétention des étudiants, alors que la conclusion contraire est tirée pour un établissement dont le taux brut est inférieur au taux naturel. Cette méthode est en usage aux États-Unis, mais pas au Canada.

La méthode des taux naturels présente trois problèmes d’interprétation intrinsèques qui nuisent à son utilité : les interprétations normatives des taux naturels sont injustifiables; les attributions de la causalité – aux étudiants en ce qui concerne les taux naturels et aux établissements en ce qui concerne les écarts entre les taux naturels et les taux bruts – sont également injustifiables et peuvent s’avérer trompeuses; et enfin, une série unique de coefficients globaux n’est guère susceptible de fournir des caractérisations bien utiles des réalités en jeu dans chaque établissement. Un nombre important et croissant de recherches ancrent le processus de rétention dans le contexte local des établissements individuels, voire des étudiants individuels. L’accroissement des constatations tirées des recherches confirme de plus en plus que le système d’éducation postsecondaire n’est pas du tout homogène en ce qui a trait à l’ampleur ou à la nature des relations entre les facteurs qui influent sur la survenue d’événements déterminants en matière de rétention et la véritable survenue de tels événements. Non seulement les processus qui engendrent des événements marquants en matière de rétention agissent-ils à l’échelle locale, mais leur forme et leur intensité varient nettement d’un endroit à l’autre : ainsi, aucune caractérisation globale ne peut adéquatement résumer les conditions dans un établissement donné. La méthode des taux naturels semble se prêter à une analyse plus poussée et subtile, mais en réalité, il n’en est rien.

Une solution de rechange aux méthodes des taux bruts ou naturels est d’abandonner les analyses rétrospectives des taux de rétention en faveur d’analyses prospectives des risques liés à la rétention. En procédant de cette façon, les établissements utilisent des données historiques pour mettre au point des modèles statistiques des risques liés à la rétention qui touchent chaque étudiante ou étudiant individuellement. Ils se servent ensuite de ces modèles pour estimer, à l’égard de chaque personne parmi leurs effectifs, le « risque » (exprimé en tant que probabilité) de poursuite des études au-delà d’un certain délai.
Les modèles servant à l’analyse doivent inclure un vaste éventail de variables explicatives et pas juste celles qui ont trait aux caractéristiques préadmission des étudiants. Le but de l’exercice de modélisation est d’arriver à correctement estimer les risques réels liés à la rétention qui se présentent pour chaque étudiante ou étudiant. Il

est ensuite possible de cumuler ces probabilités pour obtenir une estimation de l'exposition à ces risques à l'échelle d'un établissement. Ces prévisions des risques d'ensemble pour les établissements diffèrent des taux naturels évoqués plus haut en ce sens qu'elles constituent des estimations du risque de survenue d'événements qui ne se sont pas encore produits (contrairement au taux naturels, qui sont rétrospectifs) et qu'elles visent à estimer de manière fiable le risque réel intégral, et non pas seulement la partie du risque qui est « attribuable à l'étudiante ou à l'étudiant ». La logique qui sous-tend l'interprétation des estimations des risques ne requiert aucune hypothèse normative ni causative. Au fil du temps, les résultats en matière de rétention des étudiants se font jour, et l'efficacité des mesures de rétention mises en œuvre par les établissements est évaluée en comparant les résultats réels aux risques estimés antérieurement.

Cette façon de procéder présente de nombreux avantages. Premièrement, elle permet d'éviter les problèmes associés à l'interprétation des taux naturels. Elle est centrée sur les résultats futurs en matière de rétention des étudiants et elle fournit des évaluations prospectives des risques propres à chaque étudiante ou étudiant que surviennent les événements qui produiront lesdits résultats. Cette façon de procéder aboutit donc à des renseignements fort utiles qu'un établissement peut aisément intégrer à ses mesures de rétention. Il est par ailleurs possible d'inclure dans cette analyse des données sur la participation d'étudiantes et d'étudiants individuels aux initiatives visant à favoriser leur rétention, que ce soit au stade de l'estimation des risques ou lors d'évaluations subséquentes de ces initiatives, voire les deux, selon le cas. Compte tenu de leur nature prospective, les estimations des risques peuvent servir d'intrants aux fins d'autres analyses, telles que les prévisions des effectifs. Enfin, vu que les estimations des risques sont exprimées sous forme de probabilités dont les valeurs se situent toujours entre 0 et 1 et dont l'interprétation est uniforme, il devrait être possible de les comparer directement d'un modèle à l'autre et d'un établissement à l'autre.

Nous avons examiné cette nouvelle façon de procéder dans le cadre d'une étude pilote basée sur les données relatives aux effectifs de premier cycle de l'Université York, située à Toronto, en Ontario. Nous nous sommes plus particulièrement servi d'une extension de l'analyse des probabilités de survie fondée sur les arbres de classification et de régression – à savoir d'un algorithme de « forêt de survie aléatoire » (*random survival forest ou* RSF) – pour analyser les données sur le délai de survenue d'un événement à l'origine d'une interruption des études touchant 83 593 personnes qui ont été admises pour la première fois, directement après l'école secondaire, à l'Université York, en première année d'un programme d'études de premier cycle entre l'automne 1996 et l'automne 2006. Aux fins de la présente étude pilote, le mot « interrompant » désigne toute personne qui ne s'est pas réinscrite à un établissement, que son absence des effectifs s'avère être temporaire ou permanente. « Interrompants » est donc un terme qui englobe à la fois les partants définitifs et temporaires, sans égard au statut que ces personnes pourront avoir ultérieurement au sein d'un quelconque établissement d'enseignement postsecondaire, le cas échéant. Le principal résultat de l'analyse est une série d'estimations de probabilité de survie pour chaque étudiante ou étudiant, en intervalles d'une année (c.-à-d., sa probabilité de survivre [poursuivre ses études] au-delà de la 1re année, sa probabilité de survivre au-delà de la 2e année, et ainsi de suite). Les probabilités de survie peuvent être interprétées comme des estimations des risques liés à la rétention, et les compléments

des probabilités de survie peuvent être interprétés comme des risques d'interruption des études. Les estimations de probabilité de survie ont servi de fondement à une classification binaire des étudiants en « interrompants » et « poursuivants », et elles ont été soumises à une évaluation de leur fiabilité et donc de leur utilité pour estimer l'exposition aux risques liés à la rétention au niveau des établissements.

Nous avons ainsi pu établir une classification prédictive unique des interrompants qui a correctement identifié un peu plus de 90 % des personnes qui ont interrompu leurs études durant la 1re année et plus de 35 % des interrompants en 2e année, et dont la spécificité prédictive était de 82 % (c.-à-d., dont le taux de résultats faussement positifs était de 18 %).

Les estimations de la probabilité de survie (c.-à-d., des risques liés à la rétention) des étudiants présentant le plus fort risque d'interrompre leurs études avant la fin de la 1re année se sont avérées plus élevées que les taux de survie réels (c.-à-d., les taux de rétention) et elles ne pouvaient de ce fait pas servir de fondement à l'évaluation de l'exposition aux risques d'interruption des études à l'échelle des établissements ou de l'efficacité des mesures de rétention. Nous sommes d'avis que l'utilisation d'un ensemble de données plus évolué permettrait vraisemblablement d'obtenir des estimations plus fiables, qui se prêteraient bien aux évaluations des risques courus en matière de rétention et des mesures prises pour y remédier. C'est pourquoi il serait indiqué de recueillir à l'avenir, d'une part, des données « de détection rapide » fondées sur les conditions initiales auxquelles les étudiants sont confrontés au moment d'entamer une année d'études et, d'autre part, des données « en cours d'études » qui deviennent disponibles à mesure que l'année avance. Les données de détection rapide pourraient fournir des indications quant aux buts des étudiants ou à leurs engagements, leur niveau de préparation et leur participation aux études, au même titre que des indications sur les éléments extérieurs qui demandent leur attention et qui leur prennent du temps. Elles pourraient aussi fournir des corrélats du capital social et culturel, voire même des mesures de l'intelligence émotionnelle (dont il a été dit qu'elle joue sur la capacité d'adaptation à la vie universitaire). Les données recueillies en cours d'études pourraient inclure des renseignements sur les premiers résultats des étudiants dans certains cours clés, indispensables à la poursuite des études, des indications sur l'intégration et l'engagement des étudiants, ou même de l'information concernant l'intention déclarée de ces derniers de persévérer aux études.

Un programme de recherches qualitatives par l'entremise d'entrevues ou de groupes de discussion visant à examiner de près le vécu collégial ou universitaire des personnes qui ont décidé d'arrêter leurs études pourrait aider les chercheurs qui se penchent sur le sujet à attribuer un ordre de priorité aux différentes solutions envisageables pour constituer un ensemble de données plus évolué pour la prédiction des risques liés à la rétention. Les lacunes que présentent à l'heure actuelle les données disponibles par rapport à ce qui est requis pour produire des estimations des risques précises sont un problème auquel il faudra sans nul doute s'attaquer. Une méthode prospective, fondée sur les risques liés à la rétention, présente de nombreux avantages, mais la rendre exploitable appellerait un travail énorme.

Ce qui plaide en faveur de l'adoption d'une nouvelle approche

La persévérance globale et la rétention au niveau des établissements

Dans leur étude des recherches sur le maintien et la diminution des effectifs étudiants dans les universités et les collèges, Grayson et Grayson (2003) font remarquer que la seule analyse des taux de maintien et de diminution associés à un établissement particulier peut nous « nous induire en erreur », surtout si l'idée est de comprendre la diminution et le maintien des effectifs globalement, à l'échelle du système. Il en est ainsi parce que bon nombre d'étudiants qui interrompent leurs études dans un établissement d'enseignement postsecondaire les poursuivent ultérieurement et bien souvent au sein d'un autre établissement. Dans les années qui ont suivi la publication de leur étude, plusieurs autres écrits sont parus qui rendent compte du suivi des inscriptions d'étudiants individuels dans l'ensemble des établissements, à l'échelle du système, à l'aide d'ensembles de données longitudinales (comme exemple récent, voir Parkin et Baldwin, 2009).

En dehors du fait qu'elles dressent un meilleur portrait de la diminution globale des effectifs des établissements d'enseignement postsecondaire – sur cinq ans, elles fixent la proportion d'étudiants qui interrompent leurs études à 10 %, contre 26 % lorsque le taux est calculé à partir d'ensembles de données relatives à des établissements individuels (Finnie et Qui, 2008) – ces études plus récentes ont contribué à l'acception du mot « persévérance » dans un sens très différent du mot « rétention ». Par le passé, les mots « persévérance » et « rétention » (dans l'expression « rétention des étudiants », équivalente à « maintien des effectifs ») servaient de façon plus ou moins interchangeable à décrire le *fait* pour les étudiants de *rester* dans un cursus donné, en général au sein d'un même établissement et parfois dans le cadre d'un même programme d'études. Depuis quelques années, l'interprétation courante de persévérance dans ce contexte a toutefois évolué dans le sens de la *capacité* des étudiants de *poursuivre* leurs études postsecondaires et, à terme, d'obtenir un diplôme, sans égard à leurs éventuels changements de programmes ou même d'établissements ni à leurs absences temporaires mais complètes du milieu postsecondaire, le cas échéant. La persévérance souligne le fait que, de nos jours, le cheminement des étudiants au palier postsecondaire varie et que certaines personnes semblent mieux à même que d'autres de procéder à des ajustements dans leur vie qui leur permettent de poursuivre leurs études. Ainsi, la persévérance semble être liée à la résilience, autrement dit à « la capacité de franchir les obstacles, de s'adapter au changement, de se remettre sur pied après un traumatisme ou de survivre et de s'épanouir dans l'adversité » (Fondation canadienne pour le développement de carrière, 2007; Parkin et Baldwin, 2009).

L'une des implications de cette nouvelle façon de voir les choses est que la mobilité des étudiants au sein du système d'éducation postsecondaire est une réalité incontournable dont les gouvernements comme les collèges et universités doivent tout simplement tenir compte. Une manière de composer efficacement avec la mobilité des étudiants est de bâtir un système d'éducation postsecondaire qui permette aux étudiants de faire leur choix parmi la pléthore de possibilités qui leur sont offertes en matière d'études, et cela à différentes étapes de celles-ci, pas juste au tout début. Des initiatives propices à renforcer la résilience des étudiants ou à réduire le degré de

résilience dont les étudiants doivent faire preuve (par l'atténuation de l'adversité et des obstacles inutiles associés à une réorientation de ses études postsecondaires en cours de route) pourraient aussi s'avérer utiles.

Ce serait toutefois une erreur de penser que ces préoccupations globales constituent l'angle principal sous lequel il convient d'aborder les résultats concernant le maintien et la diminution des effectifs ou la persévérance. En réalité, les préoccupations concernant la persévérance à l'échelle du système ne présentent un intérêt que dans le contexte de décisions initiales de quitter un établissement donné et de ne pas y retourner. Le premier plan d'attaque pour atteindre globalement des taux élevés de persévérance doit être pour les établissements d'aider les étudiants à atteindre leurs objectifs en matière d'études postsecondaires *sur place*. Comme l'affirment Parkin et Baldwin (2009), il appartient aux établissements d'enseignement postsecondaire de dépister les étudiants appartenant à des groupes susceptibles d'éprouver des difficultés en cours de route et de « leur offrir des programmes de soutien, conçus à leur intention et en fonction de leurs besoins, qui les aideront à s'adapter aux circonstances et à obtenir leur diplôme ». Parkin et Baldwin poussent leur raisonnement plus loin en disant que « les établissements (…) devront concentrer de plus en plus leurs interventions au niveau "micro", soit sur certains sous-ensembles de leurs effectifs étudiants. Les mesures qu'ils prendront à l'égard de ces groupes seront déterminantes pour l'essor du système d'enseignement postsecondaire canadien dans son ensemble. » Tout cela pour dire que si la décision d'une étudiante ou d'un étudiant d'interrompre ses études ne constitue aucunement un échec de la part de cette personne ou de l'établissement qu'elle fréquentait, la recherche laisse entendre qu'elle constitue bien une sorte d'échec.

Que les facteurs en jeu soient une carence d'ordre financier, culturel ou social, une préparation insuffisante ou autre chose, la décision d'interrompre ses études est souvent basée sur une cause profonde connue. Il ne suffit certainement pas de dire que pour s'adapter à ces cas de figure, la simple chose à faire serait de faciliter le choix d'autres cheminements au palier postsecondaire. Pareille solution ne ferait qu'engendrer une masse d'étudiants en errance permanente, qui se contenteraient d'expériences éducatives sans rapport avec leurs aspirations véritables, choisies seulement parce qu'elles sont à leur portée, compte tenu de leurs ressources financières, culturelles et personnelles. Il serait préférable que les établissements individuels s'efforcent d'y voir clair dans ce qui les distingue eux-mêmes, de comprendre leurs étudiants, et d'améliorer la manière dont ils communiquent avec ces derniers. Qui plus est, il serait bon que les gouvernements admettent qu'il est absolument nécessaire d'axer une fois pour toutes les efforts d'une stratégie visant à atteindre des taux de persévérance élevés sur l'amélioration des résultats *sur place* en matière de rétention des étudiants.

Dans ce contexte, un second extrait du rapport de Grayson et Grayson (2003) mérite qu'on s'y arrête. Au vu des constatations tirées des recherches sur l'identité des étudiants qui quittent leurs études et sur leurs motifs à cet égard, ces auteurs concluent comme suit :

> « [I]l peut être trompeur de généraliser à propos des étudiants qui abandonnent et des raisons qui motivent leur décision. Des facteurs comme

> la moyenne pondérée cumulative, qui dans certaines situations favorisent la persévérance, ont une influence nulle ou même négative dans d'autres circonstances. De toute évidence, le seul facteur qui agit à tout coup sur la rétention reste l'intention exprimée de l'étudiant de poursuivre ses études l'année suivante. Nous avons aussi des raisons de croire que l'intégration scolaire permet mieux d'expliquer la persévérance dans les établissements d'externes que dans les collèges et universités pourvus de résidences. Nous n'affirmons pas que nous devrions abandonner la possibilité de comprendre un jour les phénomènes de persévérance et de diminution des effectifs chez les étudiants; nous disons seulement que, jusqu'ici, les explications dont nous disposons sont spécifiques à chaque établissement. …. Dans des contextes institutionnels différents, des facteurs différents expliquent la diminution des effectifs. »

Un sujet sur lequel nous reviendrons plus en détail dans la prochaine section du présent document est la question de savoir « qui quitte et pourquoi ». Il semblerait que la meilleure source d'une réponse à cette question soient les établissements mêmes et les étudiants qui les fréquentent. Il n'est sans doute guère réaliste d'espérer trouver des réponses à la fois détaillées et d'application générale (p. ex., globale) à cette question, et les études qui ne tiennent pas compte de leurs retombées sur des établissements particuliers peuvent être difficiles à interpréter correctement et à bon escient.

Pour résumer, certains font valoir que malgré l'importance de comprendre la persévérance des étudiants de l'Ontario à l'échelle du système d'éducation postsecondaire – et d'adopter des politiques qui n'entravent pas indûment cette persévérance dans l'ensemble des établissements d'enseignement postsecondaire de la province – l'objectif premier de tout établissement devrait être d'améliorer le maintien de ses propres effectifs. Il y va de l'intérêt véritable des gouvernements et des universités de se doter de moyens d'évaluer l'efficacité des mesures de rétention mises en œuvre par les divers établissements, de les comparer et de les mettre en commun pour concrètement renforcer les résultats qu'elles produisent.

Les taux de rétention des établissements sur le plan des taux bruts

Par suite de l'adoption, en 1990, d'une nouvelle loi régissant le droit à l'information des étudiants et la sécurité sur les campus (*Student Right to Know and Campus Security Act - Public Law 101-542*), les collèges et universités des États-Unis ont été obligés de publier des statistiques sur le rendement de leurs effectifs (c.-à-d., leurs taux d'achèvement des études et d'obtention d'un diplôme). Astin (1993b, 1997) souligne que le libellé de cette loi trahit l'hypothèse sous-jacente que de connaître ces taux aiderait les étudiants à prendre des décisions éclairées quant à l'établissement auquel s'inscrire, supposément parce qu'un taux d'achèvement des études ou d'obtention d'un diplôme plus élevé est signe qu'un établissement fait du bon travail pour ce qui est de maintenir ses effectifs et d'aider ses étudiants à atteindre leurs objectifs. Suivant ce raisonnement, les établissements dont les taux sont plus bas sont considérés assez peu efficaces à cet égard. En d'autres mots, d'emblée, ce sont les taux bruts (également qualifiés de taux de rétention absolus, simples ou réels) qui ont servi à jauger l'efficacité et le rendement des mesures de rétention des établissements en

comparant ces derniers entre eux. La publication de taux bruts de rétention et d'autres éléments connexes (p. ex., la diminution des effectifs, l'achèvement des études et l'obtention d'un diplôme) est aujourd'hui monnaie courante aux États-Unis, au Canada et ailleurs, au même titre que les comparaisons subséquentes des établissements.

Néanmoins, le recours aux taux bruts pour comparer l'efficacité et le rendement des mesures de rétention des établissements est malavisé. Astin (1993a, 1993b, 1997) fait remarquer que les comparaisons des établissements fondées sur les taux bruts d'obtention d'un diplôme sont irrémédiablement confondues par des facteurs associés au genre d'étudiants qu'un établissement décide d'admettre. Il ressort de son analyse longitudinale des données relatives à 39 243 étudiants inscrits dans des programmes de quatre ans au sein de 129 collèges et universités que plus de la moitié des écarts entre les établissements sont statistiquement attribuables aux caractéristiques préadmission des étudiants, sans que les modèles statistiques ne tiennent compte de l'influence que les établissements peuvent avoir eu sur les résultats aux études (Astin, 1993b). Pareillement, Astin et Oseguera (2005) ont fait une étude longitudinale à l'aide de données d'une enquête menée en 1994 auprès d'étudiants de première année (*1994 CIRP Freshman Survey*) et conclu que les deux-tiers des différences entre les établissements concernant leurs taux d'obtention d'un diplôme étaient statistiquement attribuables aux caractéristiques préadmission des personnes admises dans les établissements pris en considération. Astin (1997) a par ailleurs noté l'importance de facteurs externes, tels que la taille de l'établissement, le domaine d'études et la question de savoir si l'établissement est une université urbaine accueillant en majorité des étudiants logés hors de son enceinte ou dans une résidence sur le campus. Il a toutefois pris garde de souligner que l'effet de ces facteurs sur les taux bruts peut varier du tout au tout. La principale conclusion que l'on peut tirer de ces études est qu'il serait malavisé et trompeur de comparer les taux bruts d'obtention d'un diplôme (et, par extension, les taux bruts de maintien ou de diminution des effectifs) de différents établissements sans tenir compte des caractéristiques préadmission des personnes admises à ces établissements.

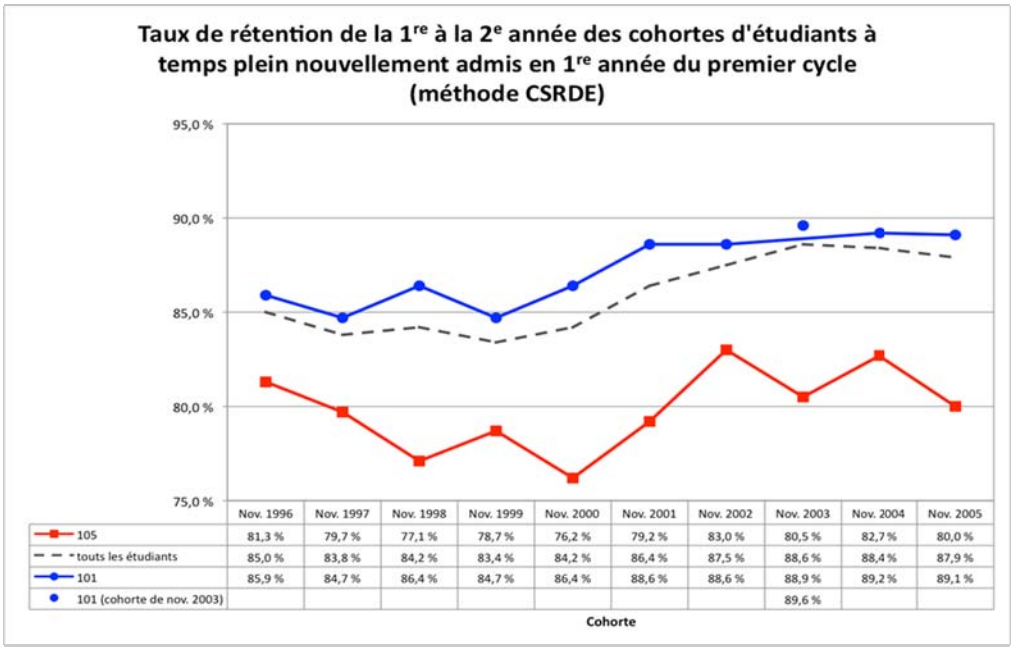
Le problème fondamental que pose l'analyse des taux bruts est que ces taux ne sont ni plus ni moins que des mesures des résultats non ajustées en fonction des variations au niveau des intrants, ce qui les rend manifestement inadéquats. Un établissement qui est en mesure de limiter ses admissions à des étudiants très bien préparés sur les plans scolaire, financier et culturel à la vie universitaire qu'il leur propose peut s'attendre à être récompensé par des taux de réussite relativement élevés, et cela sans avoir à investir grand-chose dans de bons programmes de rétention. Les établissements les plus sélectifs pourraient même (s'ils le souhaitaient) aller jusqu'à faire l'impair sur les besoins de leurs étudiants présentant des risques plus élevés et quand même afficher d'excellents taux de rétention globaux. Les établissements agissant de cette façon règlent tout simplement leurs problèmes de rétention en refusant d'admettre les étudiants qui présentent de plus grands risques à cet égard. Bien qu'une telle stratégie d'admissions puisse être valable pour certains établissements, son application universelle à l'échelle du système d'éducation postsecondaire est inimaginable, vu qu'elle serait indéniablement contraire aux objectifs de la politique publique, qui sont de rendre les études postsecondaires accessibles au plus grand nombre, sans restrictions sociales, économiques, culturelles ni (jusqu'à un certain point, s'entend) intellectuelles. En revanche, il existe aussi des

universités dont les politiques d'admission, le contexte socio-culturel et peut-être même le mandat contribuent à des taux de rétention plus bas. Il est fort possible que ces établissements fassent des efforts constants pour améliorer leurs programmes de rétention et même que ces efforts soient souvent couronnés de succès, mais quoi qu'ils fassent, ils ne pourront jamais espérer atteindre des taux bruts similaires à ceux des établissements qui ne sont dès le départ jamais exposés aux mêmes risques liés à la rétention. Lorsque les gouvernements et les étudiants potentiels se servent des taux bruts pour se faire une idée de l'efficacité relative des mesures de rétention des établissements et des résultats de celles-ci, ils n'apprécient guère à leur juste valeur les efforts déployés par les établissements qui ont une exposition aux risques bien plus considérable que d'autres. Bref, évaluer l'efficacité des programmes de rétention en se basant sur les taux bruts équivaut à récompenser les établissements qui sont capables de se délester de risques potentiels liés à la rétention durant le processus d'admission.

Il convient aussi de se rappeler que les taux bruts d'un établissement particulier peuvent changer au fil du temps, et ce en raison de l'évolution des caractéristiques de ses effectifs et non en raison de modifications apportées à ses mesures de rétention. L'Université York a par exemple noté dans le rapport qu'elle a présenté au gouvernement de l'Ontario en 2006 conformément au plan d'action pluriannuel de reddition de comptes que la dynamique de son taux de rétention de la première à la deuxième année se montrait en apparence sensible à la proportion d'étudiants admis directement après l'école secondaire (Université York, 2007). Le code administratif couramment attribué à ces étudiants est « 101 ». Au groupe des non-101 (qualifiés de « 105 ») correspond en général un taux de rétention plus faible et beaucoup plus dynamique qu'au groupe des 101 (voir la figure 1). Cet état de fait engendre des situations telles que celle que l'on a pu observer avec la cohorte d'étudiants admis en 2003. Cette cohorte incluait une plus forte proportion de 101 que d'habitude, avec pour conséquence que les 101 ont déterminé le taux de rétention global (représenté par la ligne en pointillés dans la figure) malgré une forte baisse du taux de rétention des non-101.

Figure 1

Taux de rétention des cohortes d'étudiants nouvellement admis (étudiants nouveaux à temps plein admis en 1re année d'études) à l'Université York entre l'automne 1996 et l'automne 2005, calculés suivant la méthode utilisée par le *Consortium for Student Retention Data Exchange* (CSRDE).



	Nov. 1996	Nov. 1997	Nov. 1998	Nov. 1999	Nov. 2000	Nov. 2001	Nov. 2002	Nov. 2003	Nov. 2004	Nov. 2005
105	81,3 %	79,7 %	77,1 %	78,7 %	76,2 %	79,2 %	83,0 %	80,5 %	82,7 %	80,0 %
touts les étudiants	85,0 %	83,8 %	84,2 %	83,4 %	84,2 %	86,4 %	87,5 %	88,6 %	88,4 %	87,9 %
101	85,9 %	84,7 %	86,4 %	84,7 %	86,4 %	88,6 %	88,6 %	88,9 %	89,2 %	89,1 %
101 (cohorte de nov. 2003)								89,6 %		

Source : Université York, 2007.

Les taux de rétention des établissements sous forme de taux naturels

Une autre méthode répandue consiste à calculer le taux de rétention en établissant les écarts entre les taux bruts et les taux « escomptés » ou « naturels », puis en fondant les évaluations et les comparaisons sur ces écarts. Les taux de rétention naturels des établissements correspondent à la moyenne des probabilités estimatives de survenue d'un événement (p. ex., toujours faire partie des effectifs après un an, obtenir son diplôme en l'espace de quatre ans) pour chaque membre d'une cohorte d'étudiants au sein d'un établissement. Les éléments clés des modèles statistiques sur lesquels les estimations de probabilité sont fondées sont (1) le fait qu'il s'agit de modèles globaux, qui regroupent des données provenant de l'ensemble des établissements étudiés et qui aboutissent à un seul jeu de coefficients modèles appliqués à tous les établissements; et (2) le fait que les probabilités s'appuient sur des variables explicatives qui tiennent en général compte des seules caractéristiques préadmission des étudiants et parfois de certaines caractéristiques externes, telles que la taille de l'établissement, le domaine d'études et la question de savoir si l'établissement est une université urbaine accueillant en majorité des étudiants qui résident hors de son enceinte (Astin, 1993a, 1993b, 1997; Higher Education Research Institute, 2003). Un établissement dont le taux brut excède son taux naturel est considéré obtenir de bons résultats en matière de rétention des étudiants, alors que la conclusion contraire est

tirée pour un établissement dont le taux brut est inférieur au taux naturel. Les adeptes du recours aux taux naturels font valoir que celui-ci permet d'éviter les écueils d'une analyse basée uniquement sur les taux bruts et qu'il fournit une « norme interne » (le taux de rétention naturel) par rapport à laquelle juger un établissement. Astin (1993b) dit qu'avec une telle méthode, un établissement est simplement « comparé à lui-même ». La comparaison des établissements repose ainsi sur l'écart entre les taux qu'ils s'attendaient à atteindre et les taux qu'ils ont réellement enregistrés.

La méthode des taux naturels semble présenter des avantages par rapport à la méthode incontestablement inadéquate des seuls taux bruts, et elle a d'ailleurs été mise en œuvre à plusieurs reprises. Astin a publié, seul (1997) et avec Oseguera (2005) des formules de calcul des taux naturels d'obtention d'un diplôme à l'intention des établissements d'enseignement postsecondaire américains et, depuis 1997, la publication *US News and World Report* diffuse des mesures du « rendement en matière d'obtention d'un diplôme » qui représentent essentiellement l'écart entre les taux d'obtention d'un diplôme bruts et naturels. Aucune formule des taux naturels n'a été mise au point à l'usage des établissements canadiens, une situation que Grayson et Grayson (2003) jugent « malheureuse ».

La méthode des taux naturels présente toutefois trois problèmes d'interprétation intrinsèques qui nuisent à son utilité : les interprétations normatives des taux naturels sont injustifiables et peuvent même être dommageables; les attributions de la causalité – aux étudiants en ce qui concerne les taux naturels et aux établissements en ce qui concerne les écarts entre les taux naturels et les taux bruts – sont également injustifiables et trompeuses; et enfin, une série unique de coefficients globaux n'est guère susceptible de fournir des caractérisations bien utiles des réalités en jeu dans chaque établissement.

Les taux naturels sont communément soumis à une interprétation normative. Les taux bruts qui sont très inférieurs à la norme sont considérés révélateurs d'une défaillance ou d'un échec (au niveau des mesures de rétention prises par les établissements) et ceux qui les excèdent de beaucoup sont vus comme des signes d'efficacité exceptionnelle ou d'excellence. Les analyses qui ont recours aux taux naturels semblent entériner cette hypothèse normative sans réserve. Les lecteurs sont confrontés à la question : « Que valent vos taux de rétention? », puis il leur est conseillé de se servir de la méthode des taux naturels pour y répondre (Astin, 1997; Higher Education Research Institute, 2003).

Qu'une telle hypothèse normative n'ait pas lieu d'être se comprend assez facilement. D'abord, rien n'empêche les défauts des mesures de rétention de se manifester, en partie, dans les taux naturels – et ils y feront leur apparition compte tenu de la covariance des effets de ces défauts et des caractéristiques préadmission des étudiants – et en partie dans les déviations des taux naturels. Deuxièmement, les taux naturels ne sont que des espérances statistiques, espérances qui changeront selon l'évolution de leurs ensembles de données sous-jacents. Une espérance statistique ne s'accompagne d'aucune sanction ni d'aucun avantage intrinsèque. Il se peut que des taux naturels relativement bas soient attendus pour les étudiants de l'un ou de l'autre sexe ou pour les membres d'un groupe ethnique particulier, mais sans que pareil taux ne soit pour autant souhaitable. Attribuer au taux naturel un statut neutre sur le plan

normatif ne nous paraît donc pas indiqué. L'interprétation des taux naturels qui convient en l'espèce est donc non-normative : les taux bruts qui tombent du côté opposé aux taux naturels devraient simplement être vus comme indicatifs de taux inférieurs aux attentes à l'échelle du système, et ceux qui tombent du côté opposé à la norme devraient être vus comme indicatifs de taux supérieurs à ces mêmes attentes.

Une question qu'il est impossible de passer sous silence est celle de savoir si les espérances sont, en elles-mêmes, suffisantes. Il est facile d'accepter notre énoncé antérieur, mais semblerait-il plus difficile d'y rester fidèle. En effet, nous estimons que la méthode des taux naturels appelle une interprétation normative; même si l'on prend garde à ne pas donner de formulation explicitement normative aux interprétations, les parties intéressées voudront à notre avis toujours savoir où se situe le taux naturel d'un établissement par rapport à d'autres et les établissements eux-mêmes seront toujours enclins à se targuer d'un taux naturel « hors norme », quelles que soient les conclusions que l'on puisse tirer de leurs accomplissements à cet égard. Dans un sens, les parties intéressées – exception faite des étudiants eux-mêmes – peuvent agir à leur guise avec les taux naturels.

Un autre élément qui nuit à la méthode des taux naturels est qu'elle implique que les étudiants (et, dans une bien moindre mesure, les facteurs externes, autrement dit l'environnement) sont les déterminants des taux naturels et les établissements les déterminants des écarts entre ces taux et les taux bruts. Un taux naturel est considéré être une « norme interne », un produit des étudiants admis par un établissement. Astin (1993b) laisse entendre que la méthode des taux naturels laisse les établissements libres de se demander : « Quelle est la qualité de nos résultats, compte tenu des étudiants que nous admettons? ». Les taux naturels sont générés par les étudiants et les établissements font passer ces taux de leurs valeurs naturelles à leurs valeurs brutes réelles. Ainsi, les différences entre les taux naturels et les taux bruts sont utilisées pour mesurer l'efficacité des mesures de rétention. Autrement dit, il est pris pour acquis que la relation entre les caractéristiques préadmission des étudiants et les résultats scolaires d'intérêt (disons, la rétention au-delà de la 1re année) – autrement dit, la relation telle qu'envisagée par le modèle de régression pour produire les taux naturels – est à la fois structurellement fiable et causale.

Pourtant, les modèles de régression ne font que décrire des relations fonctionnelles d'origine empirique n'utilisant que les caractéristiques préadmission comme variables explicatives; la structure et le paramétrage de ces modèles ne reflètent pas nécessairement les processus causaux à l'origine des résultats scolaires (pas plus que l'allongement des jupes ne cause vraiment une récession, par exemple). Comme nous l'avons dit plus tôt, les retombées des défauts des mesures de rétention à l'échelle des établissements (ou les effets de tout autre facteur dont un modèle ne tient pas compte) peuvent jouer sur le paramétrage d'un modèle dans la mesure où ces retombées varient manifestement au gré des caractéristiques préadmission des étudiants. De plus, la relation structurelle présumée entre les caractéristiques préadmission et les résultats est (implicitement) considérée être assez spécifique, tandis que les différences entre les taux naturels et les taux bruts sont attribuées sans réserve et sans conteste aux mesures de rétention prises par les établissements.

En réalité, les problèmes que pose la création de modèles qui attribuent la causalité d'un côté aux étudiants et de l'autre aux établissements vont bien plus loin que les points que nous venons de soulever. Un nombre croissant de recherches soulignent l'importance de la foule de relations qui existent entre l'étudiante ou l'étudiant, son établissement d'enseignement postsecondaire et son environnement socio-culturel plus général. Ces recherches décrivent l'abandon des études comme étant le résultat des *interactions* entre toutes les parties concernées, ce qui sous-entend qu'il serait contre-indiqué de tenter une simple division des risques liés à la rétention en deux groupes, les uns attribuables uniquement aux étudiants et les autres uniquement aux établissements. Ces recherches parlent souvent de la « compatibilité » entre un étudiant ou une étudiante et l'université ou le collège de son choix, et elles prônent la notion de capital social et culturel comme moyen de mieux comprendre la décision d'interrompre ses études. Les exemples les plus évidents de ce point de vue interactionniste se trouvent peut-être bien dans les nombreuses analyses qui utilisent la théorie de Tinto relative à l'intégration des étudiants (Tinto, 1975, 1993) ou qui se servent d'extrapolations de celle-ci (Cabrera, Nora et Castaneda, 1993; Sandler, 2000; Thomas, 2000) comme fondement théorique.

Selon la théorie de Tinto, les étudiants présentent différentes caractéristiques préadmission et entament leurs études postsecondaires avec des attentes et des objectifs variés, notamment au sujet de leur future carrière. L'établissement auquel ils se sont inscrits a ses propres attentes, buts et engagements. Les étudiants, l'établissement et leur environnement interagissent, et leur interaction détermine le vécu des étudiants, tant sur le plan de leurs études que sur le plan social. La question la plus importante concernant ce vécu est celle de savoir si les étudiants s'intègrent bien au milieu social et scolaire d'un établissement, car les étudiants décident d'interrompre leurs études ou de les poursuivre en fonction de leur propre sentiment d'intégration. Les étudiants qui ont l'impression que leur vécu à l'université ou au collège concorde bien avec leurs intentions, buts et engagements (ajustés) poursuivront vraisemblablement leurs études; par contre, ils seront plus susceptibles de les interrompre, si leurs attentes et la réalité ne cadrent pas. Un point clé de cette théorie – qui est parfois oublié dans son application – est que le vécu, l'intégration et les événements associés à la rétention (c.-à-d., l'interruption ou la poursuite des études) sont engendrés par un système dynamique d'interactions entre l'étudiante ou l'étudiant, l'établissement et l'environnement. Les événements associés à la rétention découlent de cette connexion. On peut dire que les caractéristiques préadmission « expliquent » les taux de rétention, mais cet argument n'est valable que dans le contexte d'une modélisation empirique. Les choses sont en réalité bien plus complexes, et c'est sans doute pour cela que l'on relève une si grande variation d'une étude à l'autre et d'un établissement à l'autre au niveau de l'ampleur et même de la direction des effets de caractéristiques préadmission précises.

Certaines études portant sur les étudiants de première génération (ÉPG), et plus exactement sur leur accès et leur persévérance aux études postsecondaires approfondissent cette question. La définition retenue pour ÉPG varie selon les fins auxquelles elle est employée, mais en général, cette notion englobe les étudiants dont les parents n'ont pas fait [un niveau spécifique] d'études postsecondaires. Cette définition a beau être simple, il est entendu que les raisons pour lesquelles les ÉPG peuvent vivre l'université différemment le sont beaucoup moins (comme exemple

récent d'une telle étude, voir Auclair et al., 2008). En fait, certains soutiendront que la notion d'ÉPG est une sorte de paramètre fictif ou de contenant qui rassemble d'innombrables facteurs à l'origine de *l'habitus* des étudiants. Tel que défini par Bourdieu (1977, 1990), *l'habitus* désigne toutes les dispositions et les compétences durables et transposables d'une personne, y compris ses schémas de perception, de connaissances et de pensée, de même que son inclination à interpréter son vécu et à agir de certaines façons. La durabilité renvoie à l'idée que *l'habitus* ne change que lentement au fil du temps – mais il change, en réponse aux expériences qu'une personne accumule tout au long de sa vie, à mesure qu'elle se fraie un chemin dans des domaines variés.

Un « domaine », dans le contexte de la théorie de Bourdieu, est pour sa part défini comme un champ d'action social, auquel se rapportent des normes et des attentes spécifiques, et sur lequel les individus se font concurrence pour obtenir la possession de divers types de capital. Les universités présentent toute une variété de tels domaines qui se chevauchent et dans lesquels les étudiants évoluent (les salles de classe, clubs, brasseries, dortoirs, salles de sports ou bureau du registraire ou de l'aide aux étudiants) – le capital en l'occurrence prend surtout des formes sociales et culturelles. L'argument pourrait être avancé que le principal problème pour bien des ÉPG est que leur *habitus* ne se transforme pas aisément en capital social et culturel conforme aux normes et aux attentes associées aux divers domaines de la vie universitaire. De ce point de vue, les ÉPG sont plus susceptibles que les autres étudiants de ressentir une séparation aiguë entre leur *habitus* et lesdits domaines et de se trouver par là même aux prises avec de larges déficits en capital social et culturel. Ces déficits sont interprétés comme signe d'une « incompatibilité » qui pèse particulièrement lourd dans la manière dont les ÉPG vivent leurs études universitaires. Le poids de ces déficits de capital, sans parler du traumatisme qui peut résulter de « l'abandon d'une identité sociale et de l'adoption d'une autre » (London, 1996), font qu'il est difficile pour ces ÉPG d'acquérir le capital social et culturel additionnel auquel les universités donnent accès. Autrement dit, « il faut dépenser du capital pour le faire fructifier » ou encore « les riches s'enrichissent, les pauvres non ».

De toute évidence, les ÉPG n'échouent pas toujours dans leur adaptation à la vie universitaire (Ishitani, 2006; Lohfink et Paulsen, 2005). Le fait est que l'*habitus* est loin de se limiter à ce qui distingue les ÉPG des non-ÉPG : au contraire, il englobe toutes les compétences et les dispositions, y compris celles qui contribuent à l'adaptabilité, à la résilience et à la persévérance (Duggan, 2002; Pascarella, Pierson, Wolniak et Terenzini, 2004). Bien d'autres concepts pouvant servir à organiser les idées relatives à la rétention viennent s'ajouter à celui d'ÉPG vis-à-vis de l'*habitus*. L'intelligence émotionnelle en est un exemple, et un autre le statut d'immigrant récent : ce dernier va gagner en importance pour les universités de la région du grand Toronto, vu que les immigrants récents forment une proportion croissante de la population de la région.

Les concepts d'*habitus*, de domaines, de capital culturel et social et d'incompatibilité mis de l'avant par Bourdieu – et la théorie sociale qui les imbrique – s'avèrent surtout utiles parce qu'ils aident les chercheurs à approfondir la réflexion sur l'idée que les événements associés à la rétention découlent d'interactions constantes entre les étudiants et leur établissement. Par ailleurs, ces concepts positionnent les établissements non pas comme agents, mais comme scènes de nombreux champs

d'action sociaux interreliés, formant chacun un corps social qui inclut quantité de personnes interagissant les unes avec les autres, et dont n'importe laquelle pourrait amener une étudiante ou un étudiant à porter un jugement différent sur son vécu universitaire et à décider de soit poursuivre ses études, soit les interrompre. Les politiques et les procédures officielles des établissements d'enseignement postsecondaire créent une partie des normes et des attentes auxquelles les étudiants doivent faire face, mais là encore, l'expérience qu'ils en font s'inscrit principalement dans le cadre de leurs interactions avec d'autres personnes – étudiants et membres du personnel enseignant ou administratif – qui ont chacune leur propre *habitus*. Le résultat net de ces interactions peut être surprenant et difficile à prédire. Lehmann (2007) constate, par exemple, que même des différences très mineures au niveau des caractéristiques de la population étudiante semblent avoir un effet majeur sur la manière dont les étudiants qui ont participé à son étude perçoivent leur vécu. La recherche de Lehmann vient aussi renforcer une idée bien connue de bon nombre de responsables de la gestion des effectifs au postsecondaire, à savoir que les étudiants ont tendance à relativiser leur vécu à l'université, y compris l'accumulation apparente de déficits sociaux et culturels, par rapport aux gains et aux pertes de capital social, culturel, voire financier, qu'ils peuvent s'attendre à réaliser ou à subir dans quantité de domaines autres que les études universitaires.

Donc en ce qui concerne les facteurs qui déclenchent les événements d'intérêt en matière de rétention (l'interruption ou la poursuite des études), il est vrai, comme diraient certains, que le diable est dans les détails du vécu de chaque étudiante ou étudiant. Il est bien sûr toujours possible de généraliser, mais la création de modèles de rétention assez fiables pour être utiles – le genre dont il faudrait pouvoir disposer pour faire des déclarations censées au sujet de l'efficacité des efforts déployés par les établissements pour maintenir leurs effectifs – nécessiterait sans doute la prise en compte d'une forte dose d'effets d'interaction, de non-linéarités et de données sur les étudiants. La mise au point de modèles simples qui permettraient de séparer correctement une partie des taux de rétention comme étant attribuable aux étudiants en se servant d'un nombre limité de caractéristiques préadmission semble de moins en moins faisable.

Pour terminer, il faut dire que la méthode des taux naturels repose sur des données totalisées relatives aux effectifs de nombreux établissements et sur une estimation à l'aide d'une seule série de coefficients modèles globaux. Les données sur des étudiantes et étudiants individuels sont ensuite entrées dans une formule paramétrée grâce à ces coefficients globaux, afin de calculer la probabilité de la survenue du résultat d'intérêt pour chacune de ces personnes. Les taux naturels d'un établissement particulier sont ensuite calculés en faisant la moyenne des probabilités établies pour ses étudiants. Cette méthode serait tout à fait valable si un seul jeu de coefficients globaux pouvait adéquatement refléter les conditions qui règnent réellement dans chaque établissement; sinon, on se trouve en présence d'un sophisme écologique, dans le sens où des inférences erronées sont faites au sujet de la nature d'objets et de relations à un niveau inférieur, celui de l'établissement, à partir de statistiques et de relations qui s'avèrent justes au niveau agrégé (en l'occurrence, à l'échelle du système).

Hélas, la mise en garde de Grayson et Grayson contre la formulation d'énoncés généraux (concernant, p. ex., tout le système) au sujet des étudiants qui interrompent leurs études et des motifs qui les poussent à agir ainsi est aussi pertinente aujourd'hui qu'elle l'était en 2003. Non seulement sommes-nous mal placés pour faire des énoncés globaux valables au sujet des partants et de leur motivation, il semble de plus en plus évident que nous ne pourrons jamais nous prononcer de manière définitive sur ce sujet, parce que les processus qui engendrent les décisions de quitter les études ou de les poursuivre (qui, ensemble, se traduisent par le taux de rétention) interviennent à un niveau bien plus local. Un nombre important et croissant de recherches ancrent le processus de rétention dans le contexte local des établissements individuels, voire des étudiants individuels. L'accroissement des constatations tirées des recherches confirme de plus en plus que le système d'éducation postsecondaire n'est pas du tout homogène en ce qui a trait à l'ampleur ou à la nature des relations entre les facteurs qui influent sur la survenue d'événements déterminants en matière de rétention et la véritable survenue de tels événements. Non seulement les processus qui engendrent des événements marquants en matière de rétention agissent-ils à l'échelle locale, mais leur forme et leur intensité varient nettement d'un endroit à l'autre : ainsi, aucune caractérisation globale ne peut adéquatement résumer les conditions dans un établissement donné. La méthode des taux naturels semble se prêter à une analyse plus poussée et subtile, mais en réalité, il n'en est rien.

Une autre façon de voir la rétention : en fonction des risques et non des taux

Une solution de rechange à la méthode des taux naturels est que les établissements utilisent des données historiques pour mettre au point des modèles statistiques des risques liés à la rétention qui touchent chaque étudiante ou étudiant individuellement. Ils se servent ensuite de ces modèles pour estimer, à l'égard de chaque personne parmi leurs effectifs, le « risque » (exprimé en tant que probabilité) de poursuite des études au-delà d'un certain délai. Les modèles servant à cette analyse doivent inclure un vaste éventail de variables explicatives et pas juste celles qui ont trait aux caractéristiques préadmission des étudiants. Le but de l'exercice de modélisation est d'arriver à correctement estimer les risques réels liés à la rétention qui se présentent pour chaque étudiante ou étudiant dans l'établissement en question. Il est ensuite possible de cumuler ces probabilités d'une manière ou d'une autre (p. ex., sous forme de moyenne pondérée en ETP) pour obtenir une estimation de l'exposition à ces risques à l'échelle d'un établissement. Ces prévisions des risques d'ensemble pour les établissements diffèrent des taux naturels évoqués plus haut en ce sens qu'elles constituent des estimations du risque de survenue d'événements qui ne se sont pas encore produits (contrairement au taux naturels, qui sont rétrospectifs) et qu'elles visent à estimer de manière fiable le risque réel intégral, et non pas seulement la partie du risque qui est « attribuable à l'étudiante ou à l'étudiant ». La logique qui sous-tend l'interprétation des estimations des risques ne requiert (et n'invite) aucune hypothèse normative ni causative. Au fil du temps, les résultats en matière de rétention des étudiants se font jour, et l'efficacité des mesures de rétention mises en œuvre par les établissements est évaluée en comparant les résultats réels aux risques estimés antérieurement.

Cette façon de procéder présente de nombreux avantages. Tout d'abord, elle permet d'éviter les trois problèmes associés à l'interprétation des taux naturels dont nous avons parlé dans la section précédente. Elle est centrée sur les résultats futurs aux études et elle fournit des évaluations prospectives des risques propres à chaque étudiante ou étudiant que surviennent les événements qui produiront lesdits résultats. Cette façon de procéder aboutit donc à des renseignements fort utiles qu'un établissement peut aisément intégrer à ses mesures de rétention. Il est par ailleurs possible d'inclure dans cette analyse des données sur la participation d'étudiantes et d'étudiants individuels aux initiatives visant à favoriser leur rétention, que ce soit au stade de l'estimation des risques ou lors d'évaluations subséquentes de ces initiatives, voire les deux, selon le cas. Compte tenu de leur nature prospective, les estimations des risques peuvent servir d'intrants aux fins d'autres analyses, telles que les prévisions des effectifs. Enfin, vu que les estimations des risques sont exprimées sous forme de probabilités dont les valeurs se situent toujours entre 0 et 1 et dont l'interprétation est uniforme, il devrait être possible de les comparer directement d'un modèle à l'autre et d'un établissement à l'autre.

Quelques remarques additionnelles s'imposent. Idéalement, la technique de modélisation statistique retenue pour établir les estimations des risques devrait se prêter à l'inclusion de quantité de variables explicatives, d'effets d'interaction, de colinéarités et de non-linéarités. Par ailleurs, au moment de choisir une technique de modélisation, il ne faut pas perdre de vue que la validité des prédictions est la clé de l'efficacité globale de la méthode employée. Enfin, les établissements doivent procéder aux évaluations des risques liés à la rétention au plus tôt, afin de se laisser le plus de temps possible pour leurs interventions en matière de rétention. Il s'agit donc de rassembler des données permettant de déceler les risques au plus vite, telles que les résultats d'enquête sur l'engagement et le degré de préparation des étudiants ou encore les premières notes obtenues par les étudiants.

Nous avons examiné cette nouvelle façon de procéder dans le cadre d'une étude pilote basée sur les données relatives aux effectifs de l'Université York, située à Toronto, en Ontario, dont le but était d'estimer les risques liés à la rétention des étudiants de premier cycle et de prédire les événements qui incitent ces derniers à interrompre leurs études, le cas échéant.

Aperçu de l'étude pilote : Estimation des risques d'interruption des études et prédiction des événements interrupteurs à l'Université York

Aux fins de la mise en œuvre de la méthode évoquée ci-dessus, nous nous sommes servi d'une extension des méthodes d'analyse des données de survie fondées sur les arbres de classification et de régression – et plus précisément d'un algorithme de « forêt de survie aléatoire » (*random survival forest ou* RSF) (Ishwaran et Kogalur, 2007; Ishwaran, Kogalur, Blackstone et Lauer, 2008) – pour analyser les données sur le délai de survenue d'un événement à l'origine d'une interruption des études touchant les personnes qui ont été admises pour la première fois, directement après l'école secondaire, à l'Université York, en première année d'un programme d'études de premier cycle entre l'automne 1996 et l'automne 2006. Recourir à des techniques d'analyse des données de survie pour examiner les risques liés à la rétention n'a rien

de nouveau (DesJardins, McCall, Ahlburg et Moye, 2002; Ishitani et DesJardins, 2003; Ishitani et Snider, 2006; Radcliffe, Huesman, Kellogg et Jones-White, 2009). Ce que notre approche a de plus inusité, c'est l'utilisation d'ensembles d'arbres aux fins de l'analyse des données de survie, et pourtant, cette façon de procéder présente divers avantages, dont non le moindre est que ce genre de technique fonctionne en général extrêmement bien (Berk, 2006; Breiman, 2001). Ces techniques sont aussi d'un emploi assez souple pour pouvoir s'adapter aux non-linéarités et à d'autres aspects d'ensembles de données vastes et complexes (Ishwaran, Kogalur, Blackstone et Lauer, 2008), et notamment à la prise en considération courante de nombreuses sous-catégories de la classification générale. Aux fins de la présente étude pilote, le mot « interrompant » désigne toute personne qui n'est pas réinscrite à l'université (au 1er novembre, la date de recensement officielle), que son absence des effectifs s'avère être permanente ou non, et sans égard au statut que cette personne pourra avoir ultérieurement dans un autre établissement d'enseignement postsecondaire, le cas échéant. Les interrompants incluent les « changeants », autrement dit les personnes qui poursuivent leurs études postsecondaires à un moment ou à un autre – mais dans un autre établissement – et les « partants définitifs » qui abandonnent leurs études pour de bon. Dans le contexte d'un établissement qui cherche à réduire le nombre d'étudiants qui le quittent, l'événement clé est l'interruption des études, quelle que soit la décision prise ultérieurement par l'interrompant au sujet de la reprise de ses études.

Le principal résultat de l'analyse est une série d'estimations de probabilité de survie pour chaque étudiante ou étudiant, en intervalles d'une année (c.-à-d., sa probabilité de survivre [poursuivre ses études] au-delà de la 1re année, sa probabilité de survivre au-delà de la 2e année, et ainsi de suite). Les probabilités de survie peuvent être interprétées comme des estimations des risques liés à la rétention, et les compléments des probabilités de survie peuvent être interprétés comme des risques d'interruption des études. Nous avons visuellement comparé les estimations de probabilité de survie aux taux de rétention réels, afin d'évaluer leur fiabilité et donc leur utilité pour estimer l'exposition aux risques liés à la rétention au niveau des établissements. Nous nous sommes aussi servi des estimations de probabilité de survie comme fondement d'une classification binaire des étudiants en « interrompants » et « poursuivants », les premiers étant ceux dont les estimations de probabilité de survie sont tombées sous un certain seuil et qui, selon les prédictions, interrompraient leurs études et les seconds ceux dont ces estimations ont dépassé ce seuil et qui donc les poursuivraient. La validité des prédictions de ces classifications a été évaluée à l'aide de grilles de correction (tableaux croisés des prédictions par comparaison aux résultats réels observés). Les pages qui suivent résument les principaux résultats de l'étude pilote, qui est décrite plus en détail à l'annexe A.

Les probabilités estimatives de survie au-delà de la 2e année nous ont servi à prédire les interrompants, peu importe l'année durant laquelle l'interruption des études s'est produite dans la réalité (voir la figure 2) pour les étudiants inclus dans un ensemble de données d'essai. (Les données d'essai n'ont pas été utilisées pour bâtir les modèles de prévision, ce qui nous a permis d'évaluer le pouvoir prédictif du modèle de façon objective.) Moyennant une valeur seuil de 0,67, nous avons pu prédire correctement 1 458, ou 90,3 %, des interrompants en 1re année recensés dans l'ensemble de données d'essai (représentés par des points rouges dans la figure 2) et 363, ou 35,3 %, des interrompants en 2e année (représentés par des points verts). La

spécificité de prédiction globale de la classification s'est établie à 82,4 % (ce qui correspond à un taux global de résultats faussement positifs de 17,6 %; cf. le tableau 1).

Figure 2

Estimations d'ensemble de la probabilité de survie au-delà de 2 années d'études, selon le statut d'interrompant. Les points rouges correspondent à des événements interrupteurs survenant la 1re année, les points verts à des événements interrupteurs survenant la 2e année et les points jaunes à des événements interrupteurs survenant après la 2e année. Les points bleus correspondent à des étudiants qui poursuivent leurs études sans interruption jusqu'à l'obtention d'un diplôme. La ligne rouge représente un seuil de probabilité de survie de 0,67, valeur en-deçà de laquelle les prédictions veulent que les étudiants seront des interrompants. Les observations sont censurées pour tenir compte du fait que les étudiants toujours inscrits (et qui ne sont pas en voie d'obtenir leur diplôme) à la fin de la période à l'étude sont exclus de la figure, vu que le statut réel de ces étudiants au début de l'automne 2007 (c.-à-d., l'année suivant immédiatement la fin de la période à l'étude) est inconnu. Du bruit aléatoire a été ajouté à la répartition des points le long de l'axe d'abscisses pour réduire les chevauchements de points excessifs.

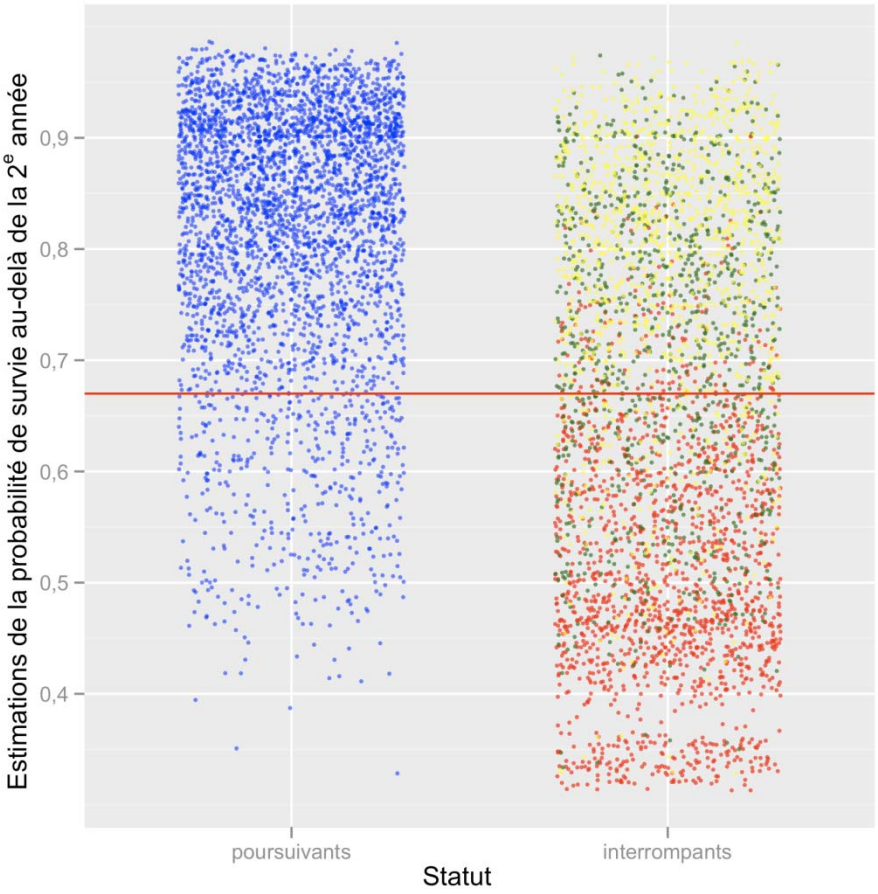


Tableau 1

Grille de correction pour la classification des étudiants en deux groupes : les « interrompants » et les « poursuivants ». Cette classification est fondée sur une valeur seuil de 0,67 appliquée aux estimations d'ensemble de la probabilité de survie au-delà de 2 ans pour les étudiants inclus dans l'ensemble de données d'essai. Les observations sont censurées pour tenir compte du fait que les étudiants toujours inscrits (et qui ne sont pas en voie d'obtenir leur diplôme) à la fin de la période à l'étude sont exclus de la grille de correction, vu que le statut réel de ces étudiants au début de l'automne 2007 (c.-à-d., l'année suivant immédiatement la fin de la période à l'étude) est inconnu.

		Prédictions		
		Poursuivants	Interrompants	Total
Observations	Poursuivants	3 127	428	3 555
	Interrompants	1 808	1 998	3 806
	Total	4 935	2 426	7 361

Ainsi, malgré les limitations des données utilisées aux fins de cette étude pilote (les délais ont, p. ex., été mesurés en années d'études et non en période d'études de 4 mois ou en mois individuels; peu de variables explicatives avaient trait à l'état de préparation et aux exigences d'ordre social, culturel et personnel; et aucune donnée sur les premières notes obtenues par les étudiants n'était disponible), nous sommes parvenus à une classification efficace des interrompants en 1re année. Ce résultat donne à penser que les personnes chargées de gérer la rétention auraient intérêt à envisager ce type de méthode prospective des risques liés à la rétention comme source de renseignements utiles, d'autant plus que les maigres ressources consacrées à la gestion de la rétention sont en général axées sur les étudiants de 1re année qui risquent de décrocher, vu que les interruptions des études sont plus fréquentes durant la première année et avant l'admission en 2e année que durant toute autre année prise individuellement.

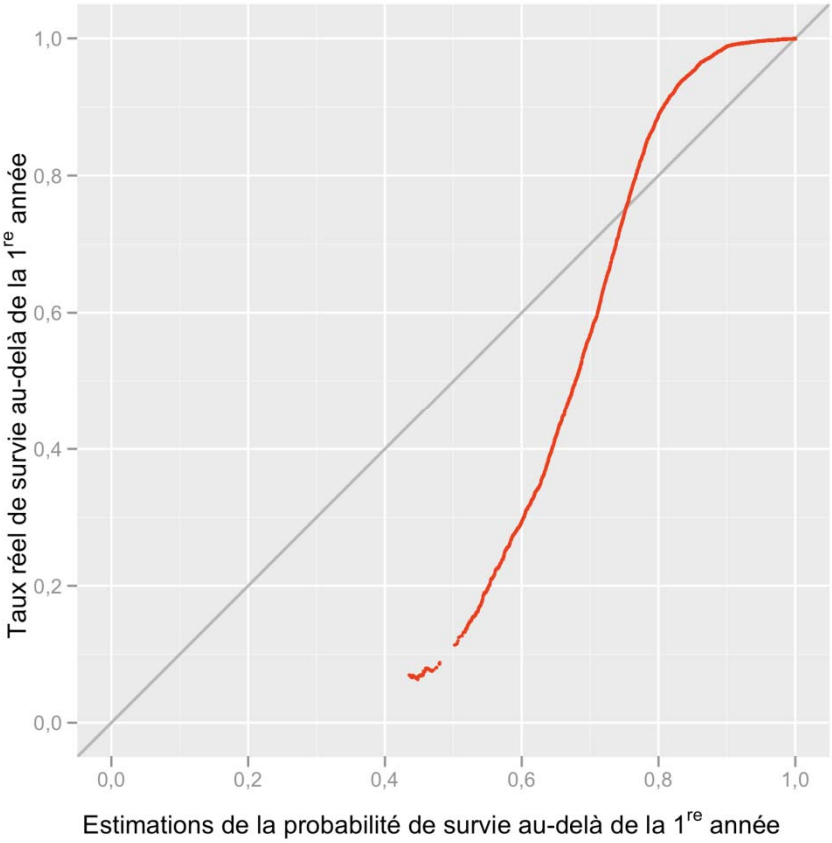
Les résultats de la classification concernant les interrompants potentiels en 2e année ou toute année ultérieure, bien qu'ils soient utiles, sont certainement moins satisfaisants que ceux qui ont trait aux interrompants en première année, en particulier parce que de nombreux étudiants additionnels interrompent leurs études durant la deuxième année. Il semble toutefois juste de s'attendre à ce que la prise en compte de variables explicatives relatives aux progrès scolaires des étudiants, aux exigences personnelles et financières auxquelles ces derniers doivent satisfaire et même à leurs intentions de persévérer aux études (ou pas) améliore la prédiction des interruptions dans les années d'études plus avancées.

On pourrait aussi s'attendre à ce qu'un ensemble de données plus évolué puisse produire des estimations de probabilité de survie plus fiables, en particulier pour les étudiants qui présentent le plus fort risque d'interrompre leurs études. Il est évident que les estimations de probabilité de survie que notre étude pilote nous a permis d'obtenir ne sont pas d'une fiabilité suffisante pour pouvoir servir de base à l'évaluation de l'exposition des établissements aux risques d'interruption. La figure 3 illustre la relation entre les probabilités estimées de survie (c.-à-d., de poursuite des études) au-delà de la 1re année d'études et les taux réels d'une survie de cette durée pour les étudiants inclus dans l'ensemble de données d'essai. Idéalement, les probabilités estimées et les taux réels devraient être quasiment identiques. Globalement, les estimations de probabilité de survie se comportent comme on pouvait s'y attendre : les valeurs de probabilité les plus élevées sont associées à des taux réels supérieurs – ce résultat est important, vu que nous lui devons l'efficacité de la prédiction des interrompants en première année.

D'un autre côté, la figure 3 montre clairement que les estimations de la probabilité de survie *ne correspondent guère* aux taux réels, sauf dans deux petits segments de la figure (l'un est centré sur la valeur 0,75 et l'autre sur la valeur 1,0). En particulier, les étudiants qui présentent le plus grand risque d'interrompre leurs études ont des probabilités de survie estimatives nettement plus élevées que leurs taux de survie réels localisés. Cela veut dire que les risques d'interruption ont été sous-estimés pour ces étudiants. Un plus grand degré de fiabilité serait requis pour pouvoir utiliser les estimations de probabilité de survie à bon escient comme intrants aux fins des modèles de prévision des effectifs, des mesures de l'exposition des établissements aux risques liés à la rétention ou encore des évaluations des mesures de rétention. Les travaux entrepris à la suite de la présente étude pilote détermineront si le recours à des ensembles de données plus évolué permettrait de réaliser des gains de fiabilité suffisants.

Figure 3

Représentation des taux réels, localisés, de survie au-delà de la 1re année d'études, selon la probabilité estimée d'une telle durée de survie. La ligne noire représente l'égalité entre les taux réels et les probabilités estimées.



Les travaux futurs devraient en particulier porter sur l'obtention d'estimations de probabilité de survie au-delà de la 1re année d'études assez fiables, sachant qu'une partie des efforts fournis dans ce sens pourront peut-être aussi servir à améliorer les estimations de survie au-delà de la 2e année, voire sur une durée plus longue. Obtenir des estimations plus précises nécessitera sans doute la collecte de données « de détection rapide » fondées sur les conditions initiales auxquelles les étudiants sont confrontés au moment d'entamer une année d'études et, d'autre part, des données « en cours d'études » qui deviennent disponibles à mesure que l'année avance. Les données de détection rapide pourraient fournir des indications quant aux buts des étudiants ou à leurs engagements, leur niveau de préparation et leur participation aux études, au même titre que des indications sur les éléments extérieurs qui demandent leur attention et qui leur prennent du temps. Elles pourraient aussi fournir des corrélats du capital social et culturel, voire même des mesures de l'intelligence émotionnelle (dont il a été dit qu'elle joue sur la capacité d'adaptation à la vie universitaire). Les données recueillies en cours d'études pourraient inclure des renseignements sur les premiers résultats des étudiants dans certains cours clés, indispensables à la

poursuite des études, des indications sur l'intégration et l'engagement des étudiants, ou même de l'information concernant l'intention déclarée de ces derniers de persévérer aux études.

En vue d'aider les chercheurs qui se penchent sur le sujet à attribuer un ordre de priorité aux différentes solutions envisageables pour constituer un ensemble de données plus évolué aux fins de la prédiction des risques liés à la rétention, nous suggérons d'entreprendre un premier programme de recherches qualitatives par l'entremise d'entrevues ou de groupes de discussion dont l'objet serait d'examiner de près le vécu universitaire des personnes qui ont décidé d'arrêter leurs études.

La possibilité d'utiliser un seul algorithme pour produire des estimations des risques sur des intervalles variés est l'un des principaux avantages qui plaide en faveur de l'utilisation de techniques d'analyse de survie pour analyser les risques liés à la rétention. (L'autre avantage majeur de ces techniques d'analyse est qu'elles se prêtent à l'utilisation de données récentes, quoique censurées.) D'un point de vue pratique, une analyse de survie qui peut se contenter d'intrants provenant d'un seul ensemble de données et produisant des estimations des risques liés à la rétention pour tous les étudiants, quelle que soit leur année d'études, est bien plus facile à effectuer qu'une analyse qui nécessite plusieurs ensembles de données distincts et une exécution séparée pour chaque cohorte d'étudiants. Ainsi, s'il est vrai que l'utilisation d'une technique d'analyse de survie n'est pas indispensable pour l'application de la méthode prospective des risques liés à la rétention décrite dans le présent rapport, elle serait néanmoins avantageuse.

Pour revenir à la question de l'évaluation de l'efficacité des mesures de rétention prises par les établissements, le présent rapport esquisse une méthode fondée sur la comparaison, d'une part, des estimations prospectives des risques liés à la rétention touchant les étudiants et, d'autre part, des résultats réels en matière de rétention des étudiants à l'échelle de ces derniers, une fois que les données sur ces résultats sont disponibles. Cette façon de procéder présente elle aussi plusieurs avantages : elle n'exige aucune interprétation normative ou causative des quantités de survie d'origine empirique; elle n'exige aucun modèle de rétention global, ce qui limite les risques de sophisme écologique (et donc de comparaisons erronées entre les établissements); et elle laisse entrevoir la possibilité d'obtenir des renseignements prévisionnels d'une grande utilité pratique que les établissements pourraient intégrer à leurs mesures visant à améliorer le maintien de leurs effectifs. À condition de recenser attentivement les interventions en matière de rétention qui touchent les étudiants individuellement, il serait par ailleurs possible de comparer les risques estimatifs aux résultats réels pour se faire une idée de l'efficacité d'interventions précises. Les mesures de rétention elles-mêmes pourraient ainsi être vues comme étant des hypothèses de travail, constamment mises à l'essai et modifiées au fil du temps dans un processus de gestion adaptative et d'amélioration continue. De plus, il serait possible d'incorporer les données recueillies aux analyses de survie effectuées les années subséquentes, de sorte à rendre compte en permanence de l'efficacité des mesures de rétention. Enfin, vu que les estimations des risques sont exprimées sous forme de probabilités dont les valeurs se situent toujours entre 0 et 1 et dont l'interprétation est uniforme, il devrait être possible de les comparer directement d'un modèle à l'autre et d'un établissement à l'autre. Pleinement mise en œuvre, la méthode préconisée dans le présent rapport

intègre donc la conception des mesures de rétention et leur évaluation, tout comme elle intègre aussi l'évaluation des mesures de rétention et la prévision des effectifs.

Il semblerait toutefois que pour pleinement réaliser les avantages potentiels de cette nouvelle méthode, il serait indispensable d'utiliser dès le départ un ensemble de données sur les étudiants assez évolué. Les lacunes que présentent à l'heure actuelle les données disponibles par rapport à ce qui est requis pour produire des estimations des risques précises sont un problème auquel il faudra sans nul doute s'attaquer. La méthode mise de l'avant ici présente de nombreux avantages, mais pour la rendre pleinement exploitable, des recherches qualitatives et une longue et minutieuse collecte de données s'imposent.

Annexe A : Détails de l'étude pilote

Données

Nous avons compilé les données sur le délai de survenue d'un événement pour tous les étudiants qui ont été admis pour la première fois, directement après l'école secondaire, à l'Université York, en première année d'un programme d'études de premier cycle durant la période à l'étude, à savoir entre l'automne 1996 et l'automne 2006. L'ensemble de données ainsi constitué comporte un enregistrement pour chacune des 83 593 personnes répondant à ces critères qui se sont inscrites à York durant la période prise en considération. Seuls les étudiants dont l'inscription était en règle à la date officielle de recensement des effectifs à l'automne (soit le 1er novembre) sont considérés inscrits pour l'année d'études correspondante.

L'ensemble de données inclut deux éléments de données qui forment un « couplet délai-événement ». Le premier élément de ce couplet est le nombre d'années d'études écoulées entre la date d'admission d'une étudiante ou d'un étudiant et la première survenue de l'événement d'intérêt, en l'occurrence une interruption des études. Aux fins de la présente étude pilote, le mot « interrompant » désigne toute personne qui n'est pas réinscrite à l'université (au 1er novembre, la date de recensement officielle), que son absence des effectifs s'avère être permanente ou non. « Interrompants » est donc un terme qui englobe à la fois les partants définitifs et temporaires, sans égard au statut que ces personnes pourront avoir ultérieurement, à l'Université York ou au sein de tout autre établissement d'enseignement postsecondaire, le cas échéant. En ce qui concerne les personnes qui interrompent leurs études, les reprennent puis les interrompent à nouveau, seul le délai jusqu'à la survenue du premier événement interrupteur est retenu. Une personne qui interrompt ses études est considérée avoir fait partie des effectifs durant toute l'année d'études antérieure. En conséquence, pour une personne admise à l'automne 2000 qui ne serait pas revenue l'année suivante – autrement dit, qui n'aurait pas fait partie des effectifs inscrits au 1er novembre 2001 – le délai de survenue de l'événement serait de 1 année d'études. (Cette façon de procéder s'expliquer par la résolution temporelle des données disponibles aux fins de l'étude pilote; elle n'est aucunement indispensable à l'analyse en général. En réalité, les étudiants peuvent évidemment interrompre leurs études à tout moment; les processus de génération des interrompants sont continus dans le temps, et non pas discontinus.)

Les ensembles de données incluent des observations censurées, la censuration ayant lieu soit parce que des étudiants étaient toujours inscrits à la fin de la période à l'étude, soit parce qu'ils se sont antérieurement « retirés de l'étude » par suite de l'obtention de leur diplôme. (Nous reviendrons sur cette seconde source de données censurées dans la prochaine section.) Pour les données censurées, l'élément relatif au délai de survenue de l'événement correspond au nombre d'années d'études durant lesquelles une personne était « en observation ». Les personnes qui étudieraient pendant quatre ans, par exemple, avant d'obtenir leur diplôme engendreraient une durée jusqu'à la censuration de 4 années d'études. Le statut des étudiants toujours inscrits à la fin de la période à l'étude ne serait, lui, par définition, pas observé au début de l'année d'études suivant immédiatement la fin de la période prise en considération, et de ce fait, un

incrément presque – mais pas tout à fait – intégral est ajouté à la durée jusqu'à la censuration pour ces étudiants en ce qui concerne la dernière année d'observation. La durée jusqu'à la censuration retenue pour les étudiants qui étaient inscrits les trois dernières années de la période à l'étude est par exemple de 2,9 années, et non de 3. Un incrément égal à une année complète est ajouté uniquement si le statut de l'étudiante ou de l'étudiant est observé le 1er novembre de l'année suivante, et pareille observation est impossible après la fin de la période à l'étude. La proportion de la dernière année qui est « tronquée » est négligeable, vu que nous utilisons dans notre analyse l'estimateur de Nelson-Aalen pour établir la fonction de risque cumulé. La seule chose qui importe est qu'un incrément partiel, et non complet, est ajouté à l'égard des étudiants inscrits de façon constante jusqu'à la fin de la période à l'étude pour la dernière année d'observation. Le deuxième élément du couplet délai-événement est un indicateur qui signale si l'observation est censurée ou s'il s'agit d'un réel événement interrupteur.

Le couplet délai-événement est la variable de réponse. Les autres éléments de l'ensemble de données sont des variables explicatives tirées des renseignements d'admission et de l'information biographique, financière et scolaire que l'Université York compile pour chaque étudiante et étudiant. Plusieurs de ces variables explicatives sont représentées par deux éléments de données, l'un étant leur valeur durant la première année d'études d'une personne, et l'autre sa valeur durant son année d'études la plus récente, autrement dit celle qui a immédiatement précédé l'événement interrupteur ou la censuration. Le tableau 2 énumère les éléments de données retenus comme variables explicatives aux fins des modèles statistiques décrits dans le présent rapport.

Tableau 2

Liste des éléments de données utilisés comme variables explicatives aux fins des modèles statistiques décrits dans le présent document

Année initiale :	catégorie de moyennes aux crédits du DESO (60-69,9 %, 70-74,9 %, 75-80,9 %, 80-89,9 %, 90-100 %)
Année initiale :	étudiante ou étudiant à temps plein (au moins 80 % d'une pleine charge de cours)
Année initiale :	étudiante ou étudiant à temps plein (au moins 60 % d'une pleine charge de cours)
Année initiale :	âge (en années)
Année initiale :	logement en résidence ou à l'extérieur
Année initiale :	nationalité canadienne ou étrangère
Année initiale :	admission tout de suite après l'école secondaire ou ultérieurement (étudiante ou étudiant de catégorie 101 ou 105)
Année courante :	indicateur d'interruption d'études antérieure
Année courante :	étudiante ou étudiant ayant redoublé
Année courante :	nationalité canadienne ou étrangère
Année courante :	indicateur de changement de statut en matière de nationalité (0 = aucun changement depuis l'année initiale; 1 = changement)
Année courante :	logement en résidence ou à l'extérieur
Année courante :	programme d'études générales ou spécialisées
Année courante :	nombre cumulé des 1res majeures (1 au départ, augmente avec chaque changement de 1re majeure)
Année courante :	nombre cumulé des 2es majeures (1 au départ, augmente avec chaque changement de 2e majeure)
Année courante :	faculté d'origine
Année courante :	indicateur de changement de faculté (0 = aucun changement depuis l'année initiale; 1 = changement)
Année courante :	diplôme visé (BA, BSc, etc.)
Année courante :	indicateur de changement de diplôme visé (0 = aucun changement depuis l'année initiale; 1 = changement)
Année courante :	étudiante ou étudiant à temps plein (au moins 80 % d'une pleine charge de cours)
Année courante :	étudiante ou étudiant à temps plein (au moins 80 % d'une pleine charge de cours)
Année courante :	valeur en équivalence au temps plein pour l'exercice
Année courante :	1re majeure
Année courante :	nombre de bourses
Année courante :	montant des bourses
Année courante :	nombre d'autres aides financières aux études
Année courante :	montant des autres aides financières aux études

Parmi les 83 593 étudiants inclus dans l'ensemble de données, 26 % sont restés inscrits jusqu'à la fin de la période à l'étude, 36 % sont restés inscrits jusqu'à l'obtention de leur diplôme et 38 % ont interrompu leurs études (voir la figure 4). Qui plus est, nous savons que 20 % des 31 850 étudiants qui ont interrompu leurs études ont fini par retourner à l'Université York, et que 35 % des étudiants qui sont retournés à

York ont fini par obtenir leur diplôme. Il convient de noter que l'ensemble de données inclut des personnes admises à l'Université York en qualité d'étudiants à temps partiel, ce qui explique qu'il engendre un nombre proportionnellement plus élevé d'interrompants que ne le ferait un ensemble de données composé uniquement d'étudiants à temps plein. Parmi les étudiants pris en considération dans cette étude pilote, seules 34 % des personnes admises à York comme étudiants à temps plein (avec au moins 80 % d'une pleine charge de cours) ont à un moment ou à un autre interrompu leurs études, par comparaison à 60 % de celles qui ont été admises comme étudiants à temps partiel. En ce qui concerne les étudiants admis entre l'automne 1996 et l'automne 2000, 40 % d'entre eux ont interrompu leurs études une première fois durant la première année et 65 % avant la fin de la deuxième année (voir la figure 5). Les étudiants qui ont interrompu leurs études la première année sont moins susceptibles d'être retournés à l'Université York que les interrompants ultérieurs : à peine 21 % des personnes qui ont interrompu leurs études la première année les ont reprises à York, tandis que 35, 37 et 28 % des personnes qui ont interrompu leurs études à l'Université York pour la première fois durant la seconde, troisième ou quatrième année, respectivement, y sont retournées étudier.

Figure 4

Cheminement après leur admission initiale des étudiants inclus dans l'ensemble de données de l'étude pilote aux fins de l'analyse des risques liés à la rétention.

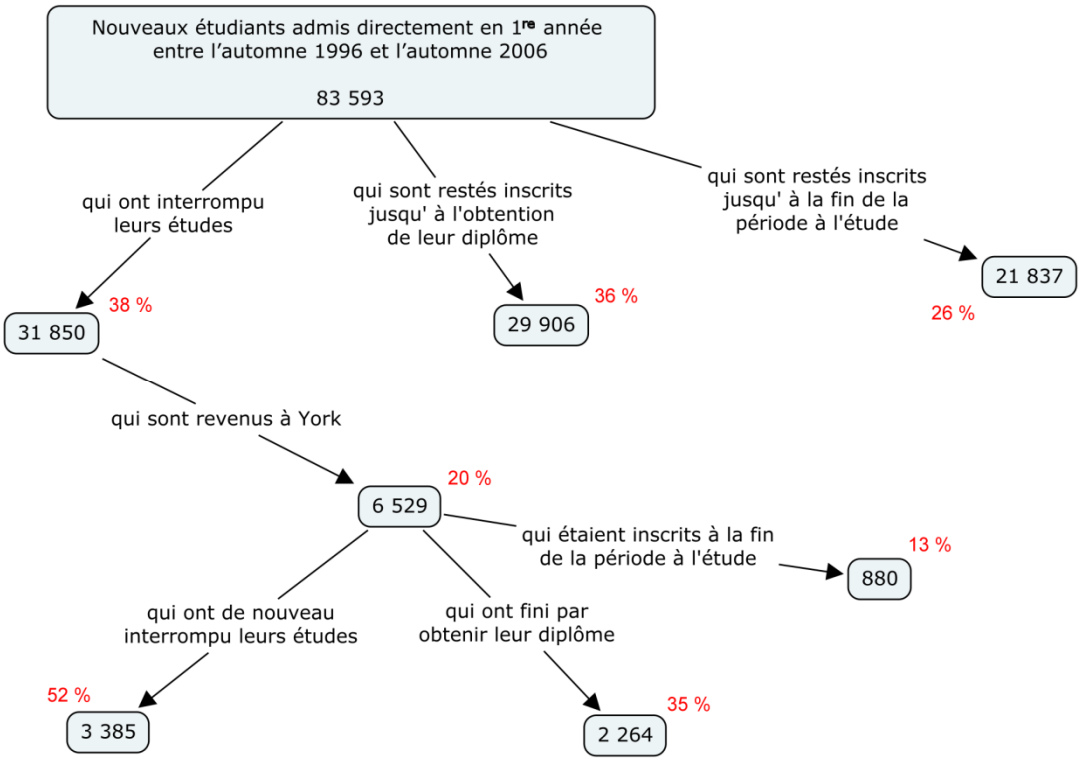
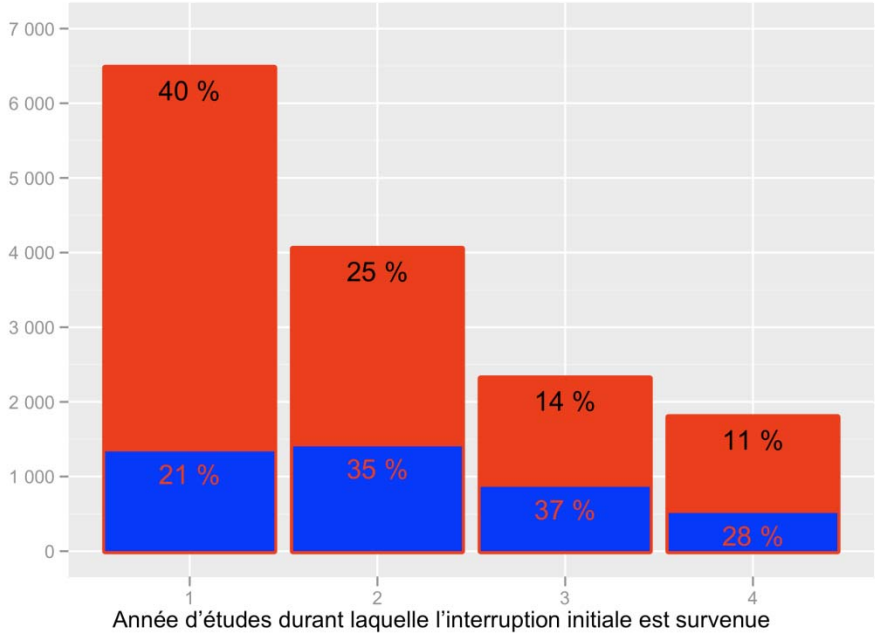


Figure 5

Histogramme des interruptions initiales, selon l'année d'études durant laquelle elles sont survenues, durant les 4 années suivant l'admission des étudiants concernés à un programme d'études de premier cycle. Les données ont trait aux étudiants inclus dans l'ensemble de données de l'étude pilote admis entre l'automne 1996 et l'automne 2000. Les barres rouges représentent le nombre d'interruptions initiales survenues dans une année (le pourcentage d'interrompants – y compris ceux qui ont interrompu leurs études au-delà des 4 premières années – étant indiqué en caractères noirs); les barres bleues représentent le nombre d'interrompants recensés dans une année qui ont fini par se réinscrire à l'Université York (le pourcentage de réinscriptions étant indiqué en caractères rouges).



Méthodes d'analyse

Nous avons analysé l'ensemble de données sur le délai de survenue d'un événement à l'aide de la technique de la « forêt de survie aléatoire » (*random survival forest ou* « RSF »), autrement dit d'une méthode fondée sur un ensemble d'arbres de survie utilisée pour l'analyse des données sur le délai de survenue d'un événement censurées à droite (Ishwaran et Kogalur, 2007; Ishwaran, Kogalur, Blackstone et Lauer, 2008). Les méthodes qui reposent sur les arbres de survie sont une extension des méthodes d'analyse des données de survie (c.-à-d., les données sur le délai de survenue d'un événement) fondées sur les arbres de classification et de régression. Les méthodes des ensembles d'arbres supposent la combinaison, d'une façon prédéterminée, des valeurs ajustées tirées d'un grand nombre de tentatives de croissance d'arbres. Un algorithme stochastique produit un large ensemble d'arbres (d'où le mot « forêt ») dont le produit est combiné. L'idée est qu'une procédure faible (qui produit, p. ex., des arbres de survie individuels dont la validité des prédictions est faible) devient forte lorsqu'elle fonctionne « en comité », et le fait est que les méthodes

des ensembles engendrent bien souvent d'excellents résultats (Berk, 2006; Breiman, 2001). L'algorithme RSF utilisé dans la présente analyse est résumé ci-dessous, d'après Ishwaran, Kogalur, Blackstone et Lauer (2008) :

1) Prélever de nombreux échantillons à partir de l'ensemble de données initial complet. Chaque échantillon bootstrap exclut environ 37 % des enregistrements qui font partie des données originales. (Les données exclues pour un échantillon particulier sont dites « hors du sac » -- *out-of-the-bag ou « oob »* -- pour l'échantillon en question.)
2) Construire un arbre de survie binaire pour chaque échantillon bootstrap. À chaque nœud de l'arbre, choisir au hasard plusieurs variables explicatives admissibles. Le nœud est fissionné à l'aide de la variable admissible qui maximise les écarts de survie entre les nœuds filles, en fonction d'un critère de survie préétabli. En maximisant les écarts de survie, l'arbre sépare les cas dissemblants. À mesure que l'arbre croît et que les cas dissemblants sont séparés, les chances de survie de chaque nœud deviennent plus uniformes.
3) Faire croître chaque arbre jusqu'à sa pleine taille, sous réserve de la contrainte qu'à chaque nœud terminal doivent correspondre non moins d'un certain nombre préétabli d'événements uniques, non censurés.
4) Calculer une fonction de risque cumulé (« FRC »), à l'aide de l'estimateur de Nelson-Aalen, pour chaque nœud terminal de chaque arbre.
5) Calculer une FRC d'ensemble pour chaque enregistrement en faisant « tomber » les enregistrements du haut de chaque arbre. Vu que les arbres sont binaires, chaque enregistrement tombe inévitablement sur un seul nœud terminal par arbre. La FRC associée au nœud en question est attribuée à l'enregistrement. La FRC d'ensemble pour l'enregistrement correspond à la moyenne de toutes les FRC attribuées à cet enregistrement dans tous les arbres.
6) Les prédictions issues de la FRC sont biaisées, parce que les données utilisées pour produire les prédictions ont aussi servi à construire les arbres au départ. Il s'agit de produire des prédictions FRC pratiquement sans biais en faisant tomber des arbres uniquement les enregistrements qui étaient hors du sac et qui n'ont donc pas contribué au processus de croissance. Ces estimations presque impartiales sont qualifiées de FRC d'ensemble hors du sac.
7) Il est ensuite possible de faire tomber de nouvelles données des arbres pour obtenir des prédictions FRC d'ensemble pour ces nouvelles données.

Nous avons retenu cette technique, parce qu'elle est simple à mettre en œuvre dans un contexte de prédictions prospectives, la validité de ses prédictions se compare favorablement à celles obtenues à l'aide d'autres méthodes (en particulier en présence de nombreuses variables explicatives) et elle se prête à la prise en compte des effets d'interaction, des non-linéarités touchant la relation entre les variables explicatives et les variables de réponse ou encore des colinéarités entre les variables explicatives. Le principal résultat qu'engendre cette technique, lorsqu'elle est employée dans un contexte prévisionnel avec de nouvelles données, est un ensemble prévu de fonctions de risque cumulé (« FRC »), une pour chaque enregistrement dans le nouvel ensemble de données. Ces FRC sont aisément converties en fonctions de probabilité

de survie à l'aide des relations bien connues entre les différentes quantités utilisées pour décrire les données de survie (Klein et Moeschberger, 2003). En particulier, l'exponentielle négative d'une valeur de risque cumulé est considérée l'équivalent de sa valeur de probabilité de survie d'ensemble. Dans le cadre de la présente étude pilote, ces probabilités de survie peuvent être interprétées comme des estimations des risques liés à la rétention au niveau de chaque étudiante ou étudiant. Les compléments des probabilités de survie peuvent être interprétés comme des prédictions des risques d'interruption des études touchant chaque étudiante ou étudiant.

Pour les besoins de l'étude pilote, nous avons prélevé dans l'ensemble global de 83 593 enregistrements cinq échantillons distincts de 10 000 enregistrements pour former cinq « ensembles d'apprentissage ». Chaque ensemble d'apprentissage a été prélevé sans remplacement de l'ensemble de données complet initial, ce qui signifie que les données relatives à une étudiante ou à un étudiant particulier peuvent se retrouver dans plus d'un ensemble d'apprentissage, sans toutefois figurer plus d'une fois au sein d'un même ensemble d'apprentissage. Nous avons appliqué l'algorithme RSF à chacun des cinq ensembles d'apprentissage et comparé les résultats. Ensuite, nous avons sélectionné les résultats de l'un des cinq ensembles d'apprentissage aux fins d'une analyse plus poussée : nous avons converti les estimations FRC hors-du-sac de l'ensemble d'apprentissage en estimations de probabilité de survie hors-du-sac; celles-ci, ainsi que les valeurs réelles, connues, relatives au délai de survenue d'un événement ont été entrées dans un algorithme sous forme d'arbre binaire de classification, afin de repérer rapidement les valeurs seuils des probabilités de survie en-deçà desquelles il serait possible de classifier les étudiants comme étant « à risque » d'interrompre leurs études. Puis nous avons prélevé un « ensemble d'essai » composé de 10 000 enregistrements, dont aucun ne fait partie de l'ensemble d'apprentissage sélectionné, à partir des 83 593 enregistrements initiaux. Nous avons produit des estimations de probabilité de survie d'ensemble pour chaque enregistrement compris dans l'ensemble d'essai en faisant tomber les enregistrements de l'ensemble d'arbres construits à l'aide des données d'apprentissage, puis nous nous sommes servi de ces estimations, en conjonction avec les valeurs seuils établies au préalable, pour classifier les étudiants en deux catégories, d'une part les « interrompants » et de l'autre les « poursuivants ». La validité des prédictions ainsi obtenues a été évaluée moyennant la comparaison des classifications établies au statut réel et connu des étudiants représentés dans l'ensemble d'essai en termes d'interruption des études. (Dans le cas d'une étude définitive, les ensembles d'arbres de la forêt de survie aléatoire (RSF) et les valeurs seuils de survie seraient appliquées à des données entièrement nouvelles, pour lesquelles aucun délai de survenue d'un événement n'aurait encore été observé, de manière à produire des prédictions prospectives.)

L'évaluation de l'exactitude des estimations de probabilité de survie d'ensemble produites pour les données d'essai a eu lieu moyennant la comparaison de ces estimations aux taux de survie réels, « localisés ». Le calcul des taux réels, locaux, s'est fait à l'aide d'un algorithme à fenêtre mobile : pour déterminer le taux réel associé à une valeur de probabilité de survie particulière, toutes les observations dans l'ensemble de données d'essai ont d'abord été pondérées en fonction de leur distance euclidienne de cette valeur (que nous qualifierons de valeur cible). Une pondération de

zéro a été attribuée aux observations associées à une probabilité de survie estimative dont la distance de la valeur cible était ±0,1. Les autres observations moins distantes ont été pondérées d'un poids égal au complément de leur distance de la valeur cible élevé à la 4e puissance (afin de mettre en relief les observations très proches de la valeur cible). Prenons par exemple une probabilité de survie cible de 0,5 : dans ce cas, toutes les observations assorties d'une probabilité de survie estimative supérieure à 0,6 ou inférieure à 0,4 auraient une pondération de zéro. Une observation à laquelle correspond une estimation de probabilité de survie de 0,525 se trouve à 0,025 de distance de la valeur cible de 0,5. Le complément de cette distance est 0,975, et la pondération attribuée à cette observation serait de $0{,}975^4 = 0{,}9037$. La moyenne pondérée d'une variable binaire faisant ressortir les étudiants qui poursuivent leurs études au-delà de la première année a ensuite été calculée et interprétée comme étant le taux de survie réel localisé associé à la probabilité de survie cible. Nous avons calculé un tel taux pour chaque valeur estimative de probabilité de survie d'ensemble générée à l'aide des données d'essai, puis nous avons rapporté sur un graphique une comparaison des estimations de probabilité et des taux réels localisés.

Une application complémentaire du logiciel statistique R (R Development Core Team, 2008), baptisée « randomSurvivalForest 3.5.1 » (Ishwaran et Kogalur, 2007), constitue la mise en œuvre de référence de la technique RSF et a été utilisée dans la présente étude pilote. Le tableau 3 présente les paramètres retenus, tels que requis par l'application randomSurvivalForest. Bien que cette application permette de saisir des données manquantes, nous avons décidé de simplement exclure des analyses faites dans le cadre de l'étude pilote les enregistrements incomplets. Nous avons utilisé le programme « rpart » du logiciel statistique R (Therneau, Atkinson et transfert sur R par Brian Ripley, 2009) pour obtenir les classifications binaires.

Tel qu'indiqué précédemment, la présente étude a engendré des données censurées, et ce de deux façons. Soit une étudiante ou un étudiant était toujours inscrit à la fin de la période sur laquelle a porté la collecte de données, soit la personne s'est « retirée » de l'étude parce qu'elle a obtenu son diplôme. Traiter la seconde catégorie d'étudiants comme des observations censurées est tout à fait logique, si l'on considère que l'obtention d'un diplôme est hors de la portée de la présente analyse et qu'elle est sans rapport avec le processus qui mène à l'interruption des études (vu que le délai jusqu'à l'obtention d'un diplôme ne fournit en principe aucune information sur la durée durant laquelle une personne serait restée inscrite à l'université si elle n'avait pas obtenu son diplôme). Ceci revient à imaginer un système universitaire où les études ne mènent jamais à un diplôme et où tous les étudiants poursuivent leurs études jusqu'au jour où ils décident de les interrompre (Klein et Moeschberger, 2003). Nous appellerons cette formulation particulière du problème de modélisation du délai jusqu'à la survenue de l'événement la « formulation d'un risque unique ». À vrai dire, il s'agit là de la formulation du problème que nous avons retenue aux fins de notre étude pilote.

D'un autre côté, l'analyse effectuée dans le cadre de cette étude pilote équivaut presque à une « formulation des risques concurrents », qui reconnaît la survenue possible de plusieurs types d'événements – en l'occurrence, soit une interruption initiale des études ou l'obtention d'un diplôme – mais seulement un des types (à savoir, l'interruption) est jugé intéressant, compte tenu de la question étudiée et du délai restreint pris en considération dans l'interprétation des résultats, soit les deux

premières années d'études. Cette restriction du délai est voulue pour des raisons pratiques, sachant, comme nous le notions plus tôt, que la plupart des interrompants sont recensés au début de la troisième année d'études et que les activités visant à améliorer la rétention des étudiants interviennent en général tôt (avant la deuxième année). Comme l'un des deux types d'événements – l'obtention d'un diplôme – ne se produit pratiquement jamais avant la fin de la troisième année d'études, le recours à une formulation des risques concurrents mène à des résultats presque identiques à ceux qui découlent d'une formulation d'un risque unique, compte tenu de la restriction du délai précitée.

Deux quantités sont souvent utilisées dans la formulation des risques concurrents, à savoir le complément de la fonction de survie par cause et la fonction de cumul des événements par cause, lesquelles sont identiques pour un type d'événement particulier lorsqu'elles sont calculées pour des moments dans le temps antérieurs à celui où s'est produit l'un ou l'autre des autres types d'événements concurrents (Klein et Moeschberger, 2003). Dans le contexte qui nous intéresse, les quantités relatives aux interrompants seront égales les unes aux autres à l'issue des calculs portant sur des moments antérieurs à toute obtention de diplôme. Vu que la fonction de survie par cause pour un type d'événement particulier est, sur le plan des calculs, identique à la fonction de survie fondée sur la formulation d'un risque unique lié à ce même type d'événement, il est clair que la formulation d'un risque unique produira des résultats très approchants de ceux de la formulation des risques concurrents, compte tenu une fois encore du délai restreint examiné.

Quoi qu'il en soit, nous n'avions pas d'autre choix que d'adopter la formulation d'un risque unique pour l'étude pilote : au moment de notre analyse, aucun logiciel n'était encore disponible qui aurait permis la mise en œuvre d'une méthode d'ensemble d'arbres de survie fondée sur des formulations de risques concurrents, même si la plus récente actualisation majeure de l'algorithme RSF offre désormais des fonctionnalités complètes d'analyse de données sur des risques concurrents (Ishwaran et Kogalur, 2010). Bien que les résultats présentés aient en partie rapport à des moments ultérieurs à la deuxième année d'études, ils mettent l'accent sur les résultats qui ont rapport à des interruptions d'études survenues avant la troisième année, donc avant la majorité des obtentions de diplômes.

Résultats

Nous avons exécuté l'algorithme RSF une fois pour chacun des cinq ensembles d'apprentissage : le tableau 3 fournit des renseignements sommaires à cet égard. Durant chacune de ces exécutions, les taux d'erreur des prédictions s'est stabilisé après la construction d'environ 400 arbres de survie, et nous avons arrêté chaque exécution lorsque le nombre d'arbres construits a atteint 600. La mesure globale de la validité des prédictions découlant de l'exécution de l'algorithme RSF est obtenue à l'aide de l'indice de concordance de Harrell (indice C), dont les détails algorithmiques sont décrits par Harrell, Califf, Pryor, Lee et Rosati (1982). Ishwaran, Kogalur, Blackstone et Lauer (2008) fournissent une explication plus poussée de l'emploi de l'indice C avec l'algorithme RSF. Les valeurs de l'indice C vont de 0 à 1, 0,5 correspondant à une validité des prédictions qui n'est pas meilleure qu'une simple conjecture et 0,0 correspondant à des prédictions sans erreur.

Tableau 3

Renseignements sommaires concernant chaque exécution de l'algorithme de la forêt de survie aléatoire. Les données exécutées consistaient en 10 000 enregistrements sélectionnés au hasard (sans remplacement) prélevés sur l'ensemble complet des données relatives aux 83 593 enregistrements des délais jusqu'à la survenue de l'événement.

	Exécution 1	Exécution 2	Exécution 3	Exécution 4	Exécution 5
Nombre d'échantillons bootstrap (arbres par forêt)	600 échantillons bootstrap				
Nombre total de variables explicatives	27 variables				
Nombre de variables explicatives sélectionnées au hasard à chaque fission de nœud	5 variables				
Règle de fission	aléatoire, selon Mantel-Haenszel				
Nombre de points de fission aléatoire	1 point de fission				
Taille minimale du nœud terminal	3 enregistrements				
Taille de l'échantillon après suppression des enregistrements incomplets	9 925	9 918	9 923	9 930	9 906
Nombre réel d'interrompants	3 835	3 762	3 845	3 809	3 716
Nombre moyen de nœuds terminaux	219	204	210	217	209
Taux d'erreur estimatif des prédictions (indice C)	19,2 %	18,2 %	20,1 %	19,1 %	19,2 %

Les valeurs de l'indice C à l'issue des cinq exécutions de l'algorithme RSF, exprimées en pourcentage (tableau 3), se situent entre 18,2 et 20,1 %. Les valeurs de l'indice C comprises dans cette fourchette sont certainement plus fiables qu'une simple conjecture, mais elles ne sont pas très basses. Il convient toutefois de noter que les valeurs de l'indice C dénotent la validité globale des prédictions concernant les événements interrupteurs qui se produisent à tout moment de la période à l'étude et qu'elles ne fournissent aucun renseignement sur des prédictions relatives à un moment spécifique. Dans la pratique, il serait intéressant de connaître les probabilités d'une interruption des études après une, voire deux années d'études – d'où l'utilité d'évaluer la validité des prédictions de l' algorithme RSF pour aborder ces questions plus précises.

La figure 6 présente les résultats de l'analyse de l'ensemble d'essai, qui illustrent la relation entre les estimations d'ensemble de la probabilité de survie (c.-à-d., de la poursuite des études) au-delà de la première année et le nombre réel d'interrompants

recensés jusqu'à la fin de cette première année. Il est clair d'après cette figure que les estimations de probabilité de survie aboutissent à une distinction bien nette entre les personnes qui interrompent leurs études avant la fin de la première année et celles qui les poursuivent au-delà. En effet, un seuil de probabilité de survie d'ensemble de 0,72 garantit une bonne validité des prédictions lorsqu'il est appliqué à l'ensemble d'essai. Une grille de correction (tableau 4) fournissant plus de détails sur la validité des prédictions montre un taux d'erreur global des prédictions de 9,2 %, une sensibilité des prédictions de 85,3 % (qui signifie que 14,7 % des interrompants réels en première année ont été incorrectement classifiés comme poursuivants) et une spécificité des prédictions de 69,6 % (qui signifie que 30,4 % des interrompants en première année prédits ont en réalité poursuivi leurs études sans interruption et se sont inscrits en deuxième année). Seuls 8,0 % des étudiants qui ont poursuivi leurs études au-delà de la première année ont été classifiés par erreur comme les ayant interrompues la première année.

Figure 6

Estimations d'ensemble de la probabilité de survie au-delà de la 1re année, selon le statut d'interrompant en 1re année. Les points rouges représentent les interrompants, quelle que soit l'année d'études durant laquelle l'interruption est survenue; les points bleus représentent les poursuivants. La ligne rouge correspond au seuil de probabilité de survie de 0,72, valeur en-deçà de laquelle les prédictions veulent que les étudiants interrompront leurs études la 1re année. Les observations sont censurées pour tenir compte du fait que les étudiants toujours inscrits en première année à la fin de la période à l'étude sont exclus de la grille de correction, vu que le statut réel de ces étudiants de 1re année au début de l'année d'études suivante (c.-à-d., l'année suivant immédiatement la fin de la période à l'étude) est inconnu. Du bruit aléatoire a été ajouté à la répartition des points le long de l'axe d'abscisses pour réduire les chevauchements de points excessifs.

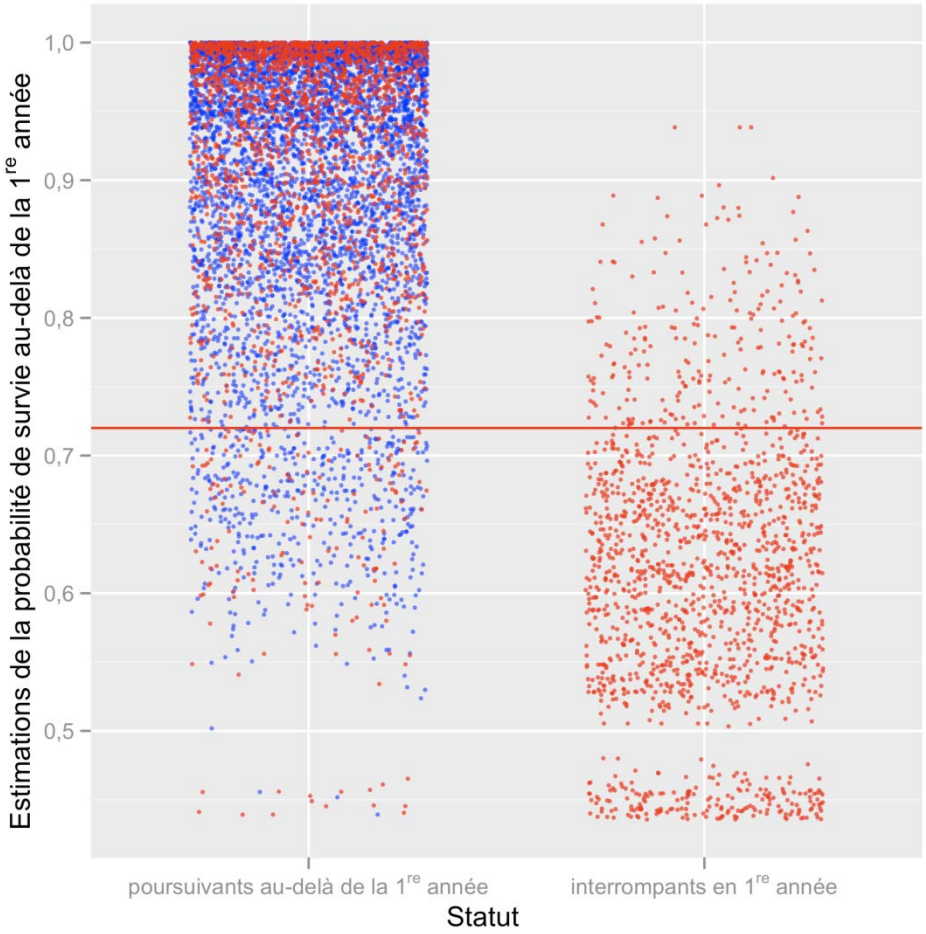


Tableau 4

Grille de correction pour la classification des étudiants en deux groupes : les « interrompants en 1re année » et les « poursuivants ». Cette classification est fondée sur une valeur seuil de 0,72 appliquée aux estimations d'ensemble de la probabilité de survie au-delà de la 1re année pour les étudiants inclus dans l'ensemble de données d'essai. Les observations sont censurées pour tenir compte du fait que les étudiants toujours inscrits en première année à la fin de la période à l'étude sont exclus de la grille de correction, vu que le statut réel de ces étudiants de 1re année au début de l'année d'études suivante (c.-à-d., l'année suivant immédiatement la fin de la période à l'étude) est inconnu.

		Prédictions		
		Poursuivants	Interrompants en 1re année	Total
Observations	Poursuivants	6 911	601	7 512
	Interrompants en 1re année	238	1 376	1 614
	Total	7 149	1 977	9 126

Les points dans la colonne de gauche de la figure 6 incluent un grand nombre d'interrompants durant les années supérieures (colorés en rouge), dont seule une faible proportion est captée de façon accessoire dans la classification des interrompants en 1re année (c.-à-d., se trouve sous le seuil de probabilité de survie au-delà de la 1re année de 0,72). Comme environ 25 % de toutes les interruptions initiales surviennent durant la deuxième année et que 60 % interviennent durant les années supérieures (y compris la deuxième), il est aussi intéressant de procéder à une évaluation de la validité des prédictions relatives aux classifications des interrompants en 2e année et durant les années supérieures. Aux fins de cette évaluation, nous avons produit une classification qui prédit les interrompants – sans égard à l'année durant laquelle l'interruption survient – fondée sur une estimation des probabilités de survie au-delà de deux ans. À priori, les résultats de cette classification (voir la figure 7 et le tableau 5) ne sont guère impressionnants, car ils présentent un taux d'erreur global des prédictions de 30,4 % et une sensibilité des prédictions de 52,5 % (ce qui signifie que 47,5 % des interrompants réels ont été incorrectement classifiés comme poursuivants). La spécificité des prédictions par contre, soit 82,4 %, est respectable (elle signifie que seuls 17,6 % des interrompants prévus ont en fait poursuivi leurs études jusqu'à l'obtention d'un diplôme). Par ailleurs, 12 % des étudiants qui ont poursuivi leurs études sans interruption jusqu'à l'obtention d'un diplôme ont été classifiés par erreur comme les ayant interrompues.

Ce qui rend les résultats de cette classification particulièrement attrayants est toutefois le fait que non moins de 1 458, ou 90,3 %, des interrompants en 1re année recensés dans l'ensemble de données d'essai (représentés par des points rouges dans la figure 7) et 363, ou 35,3 %, des interrompants en 2e année (représentés par des points verts) sont prédits correctement. Ces résultats, combinés à la spécificité globale relativement élevée des prédictions (82,4 %) donnent à penser qu'il devrait être

possible de créer des modèles de prédiction des interruptions suffisamment puissants pour pouvoir être utilisés dans un contexte opérationnel en vue de trier les risques liés à la rétention des étudiants (et en particulier des étudiants de 1re année) et d'orienter les interventions visant à améliorer cette rétention en conséquence.

Figure 7

Estimations d'ensemble de la probabilité de survie au-delà de 2 années d'études, selon le statut d'interrompant. Les points rouges correspondent à des événements interrupteurs survenant la 1re année, les points verts à des événements interrupteurs survenant la 2e année et les points jaunes à des événements interrupteurs survenant après la 2e année. Les points bleus correspondent à des étudiants qui poursuivent leurs études sans interruption jusqu'à l'obtention d'un diplôme. La ligne rouge représente un seuil de probabilité de survie de 0,67, valeur en-deçà de laquelle les prédictions veulent que les étudiants seront des interrompants. Les observations sont censurées pour tenir compte du fait que les étudiants toujours inscrits (et qui ne sont pas en voie d'obtenir leur diplôme) à la fin de la période à l'étude sont exclus de la figure, vu que le statut réel de ces étudiants au début de l'automne 2007 (c.-à-d., l'année suivant immédiatement la fin de la période à l'étude) est inconnu. Du bruit aléatoire a été ajouté à la répartition des points le long de l'axe d'abscisses pour réduire les chevauchements de points excessifs.

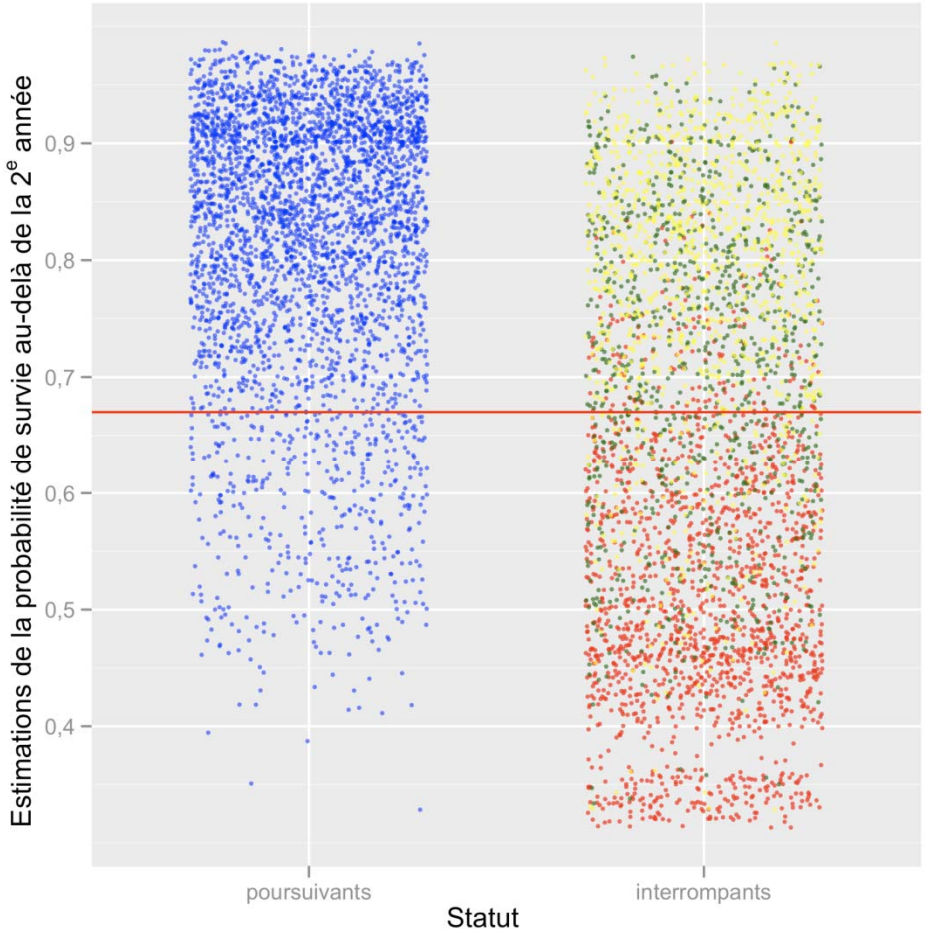


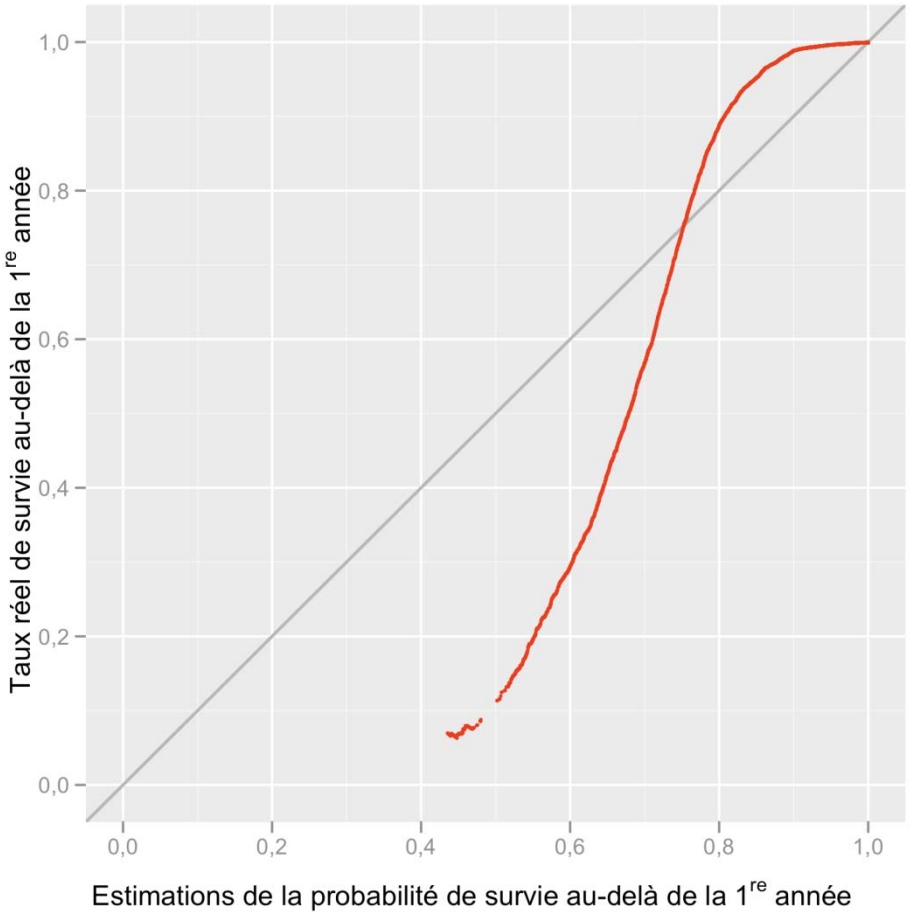
Tableau 5

Grille de correction pour la classification des étudiants en deux groupes : les « interrompants » et les « poursuivants ». Cette classification est fondée sur une valeur seuil de 0,67 appliquée aux estimations d'ensemble de la probabilité de survie au-delà de 2 ans pour les étudiants inclus dans l'ensemble de données d'essai. Les observations sont censurées pour tenir compte du fait que les étudiants toujours inscrits (et qui ne sont pas en voie d'obtenir leur diplôme) à la fin de la période à l'étude sont exclus de la grille de correction, vu que le statut réel de ces étudiants au début de l'automne 2007 (c.-à-d., l'année suivant immédiatement la fin de la période à l'étude) est inconnu.

		Prédictions		
		Poursuivants	Interrompants	Total
Observations	Poursuivants	3 127	428	3 555
	Interrompants	1 808	1 998	3 806
	Total	4 935	2 426	7 361

La figure 8 illustre la relation entre les estimations de la probabilité de survie au-delà de la 1re année d'études et les taux de survie réels passé la 1re année. Globalement, les estimations de probabilité de survie se comportent comme on pouvait s'y attendre : les valeurs de probabilité les plus élevées sont associées à des taux réels supérieurs, se prêtant à une classification binaire efficace des étudiants en « poursuivants » et en « interrompants ». Par contre, la répartition des points montre clairement que les estimations de la probabilité de survie générées par l'algorithme RSF ne correspondent guère aux taux réels, sauf dans deux petits segments de la figure (l'un est centré sur la valeur 0,75 et l'autre sur la valeur 1,0). En particulier, les étudiants qui présentent le plus grand risque d'interrompre leurs études ont des probabilités de survie estimatives nettement plus élevées que leurs taux de survie réels localisés. Cet état de fait réduit l'utilité des probabilités estimées aux fins de l'évaluation tant des profils des risques liés à la rétention établis par les établissements d'enseignement que de l'efficacité des mesures de rétention, bien que l'on puisse raisonnablement supposer que la création et l'utilisation d'ensembles de données plus évolués que ceux que nous avons employés dans cette étude pilote permettraient d'obtenir des estimations plus fiables des probabilités de survie (autrement dit des risques liés à la rétention).

Figure 8
Représentation des taux réels, localisés, de survie au-delà de la 1re année d'études, selon la probabilité estimée d'une telle durée de survie. Le calcul des taux réels, locaux, s'est fait à l'aide d'un algorithme à fenêtre mobile : pour déterminer le taux réel associé à une valeur de probabilité de survie cible, toutes les observations ont d'abord été pondérées en fonction de leur distance euclidienne de cette valeur. Une pondération de zéro a été attribuée aux observations associées à une probabilité de survie estimative dont la distance de la valeur cible était ±0,1. Les autres observations moins distantes ont été pondérées d'un poids égal au complément de leur distance de la valeur cible élevé à la 4e puissance. La moyenne pondérée d'une variable binaire faisant ressortir les étudiants qui poursuivent leurs études au-delà de la 1re année a ensuite été calculée et interprétée comme étant le taux de survie réel localisé associé à la probabilité de survie cible. Un tel taux a été calculé pour chaque valeur de probabilité de survie générée à l'aide de l'ensemble de données d'essai. La ligne noire représente l'égalité entre les taux réels et les probabilités estimées.



Bibliographie

ASTIN, A.W. 1993a. *What matters in college?: Four critical years revisited*. San Francisco, maison d'édition Jossey-Bass.

ASTIN, A.W. 1993b. « College retention rates are often misleading », *Chronicle of Higher Education,* vol. 22, p. A48.

ASTIN, A.W. 1997. « How "good" is your institution's retention rate? », *Research in Higher Education,* vol. *38,* p. 647-658.

ASTIN, A. et L. OSEGUERA. 2005. *Degree attainment rates at American colleges and universities: Effects of race, gender, and institutional type* (éd. rév.), Los Angeles, Higher Education Research Institute.

AUCLAIR, R., P. BÉLANGER, P. DORAY, P. GALLIEN, A. GROLEAU, L. MASON et P. MERCIER. 2008. *Transitions - Les étudiants de première génération : un concept prometteur?,* note de recherche, Fondation canadienne des bourses d'études du millénaire, Montréal.

BERK, R.A. 2006. An introduction to ensemble methods for data analysis. *Sociological Methods and Research,* vol. 34, p. 263-295.

BOURDIEU, P. 1977. *Outline of a theory of practice,* Cambridge et New York, Presses universitaires de Cambridge.

BOURDIEU, P. 1990. *The logic of practice*, Stanford, Presses universitaires de Stanford.

BREIMAN, L. 2001. Random forests, *Machine Learning,* vol. 45, p. 5-32.

CABRERA, A. F., A. NORA et M. B. CASTANEDA, M.B. 1993. College persistence: Structural equations modeling test of an integrated model of student retention, *Journal of Higher Education,* vol. 64, p. 123-139.

DESJARDINS, S. L., B. P. MCCALL, D. A. AHLBURG et M. J. MOYE. 2002. Adding a timing light to the "tool box", *Research in Higher Education,* vol. 43*,* p. 83-114.

DUGGAN, M. B. 2002. *The effect of social capital on the first-year persistence of first generation college students*, dissertation en vue de l'obtention d'un doctorat en éducation, Boston, University of Massachusetts.

FINNIE, R. et H. T. QUI. 2008. *The patterns of persistence in post-secondary education in Canada: Evidence from the YITS-B dataset,* document de travail du projet MESA (Measuring the Effectiveness of Student Aid) de mesure de l'efficacité de l'aide aux études, Educational Policy Institute, Université Queen's, Kingston.

FONDATION CANADIENNE POUR LE DÉVELOPPEMENT DE CARRIÈRE. 2007. *Appliquer le concept de résilience au développement de carrière : leçons à retenir pour le développement de programmes d'études,* Montréal, Fondation canadienne des bourses d'études du millénaire.

GRAYSON, J. P, et K. GRAYSON. 2003. *Les recherches sur le maintien et la diminution des effectifs étudiants,* Montréal, Fondation canadienne des bourses d'études du millénaire.

HARRELL, F. E., JR., R. M. CALIFF, D.B. PRYOR, K. L. LEE et R.A. ROSATI. 1982. Evaluating the yield of medical tests. *Journal of the American Medical Association,* vol. 247*,* p. 2543-2546.

HIGHER EDUCATION RESEARCH INSTITUTE. 2003. *How "good" is your retention rate? Using the CIRP survey to evaluate undergraduate persistence*, Los Angeles, Higher Education Research Institute.

ISHITANI, T. T. 2006. Studying attrition and degree completion behavior among first-generation college students in the United States, *The Journal of Higher Education,* vol. 77*,* p. 861-885.

ISHITANI, T. T. et S. L. DESJARDINS. 2003. A longitudinal investigation of dropout from college in the United States, *Journal of College Student Retention,* vol. 4*,* p. 173-201.

ISHITANI, T. T. et K. G. SNIDER. 2006. Longitudinal effects of college preparation programs on college retention, *IR Applications,* vol. 9, p. 1-9.

ISHWARAN, H. et U. B. KOGALUR. 2007. Random survival forests for R, *R News,*vol. 7*,* p. 25-31.

ISHWARAN, H. et U. B. KOGALUR. 2010, *RandomSurvivalForest: Ishwaran and Kogalur's Random Survival Forest,* http://CRAN.R-project.org/package=randomSurvivalForest.

ISHWARAN, H., U. B. KOGALUR, E. H. BLACKSTONE et M. S. LAUER. 2008. Random survival forests, *The Annals of Applied Statistics,* vol. *2,* p. 841-860.

KLEIN, J. P. et M. L. MOESCHBERGER. 2003. Nonparametric estimation of basic quantities for right-censored and left-truncated data, dans : *Survival analysis: Techniques for censored and truncated data,* 2e éd., p. 91-138, New York, Springer.

LEHMANN, W. 2007. "I just didn't feel like I fit in": The role of habitus in university dropout decisions, *The Canadian Journal of Higher Education,* vol. 37*,* p. 89-110.

LOHFINK, M. M. et M. B. PAULSEN. 2005. Comparing the determinants of persistence for first-generation and continuing-generation students, *Journal of College Student Development,* vol. 46*,* p. 409-428.

LONDON, H. B. 1996. How college affects first-generation students, *About Campus,* vol. 1, no. 5, p. 9-13.

PARKIN, A. ET N. BALDWIN. 2009. *La persévérance dans les études postsecondaires au Canada : dernières percées,* note de recherche, Montréal, Fondation canadienne des bourses d'études du millénaire.

PASCARELLA, E. T., C. T. PIERSON, G. C. WOLNIAK et P. T. TERENZINI. 2004. First-generation college students: Additional evidence on college experiences and outcomes, *Journal of Higher Education,* vol. 75*,*p. 249-285.

R DEVELOPMENT CORE TEAM. 2008. *R: A language and environment for statistical computing,* Vienne, R Foundation for Statistical Computing.

RADCLIFFE, P. M., R. L. HUESMAN JR., J. P. KELLOGG et D. R. JONES-WHITE. 2009. Identifying students at risk: Utilizing survival analysis to study student-athlete attrition, *IR Applications,* vol. 21, p. 1-15*.*

SANDLER, M. E. 2000. Career decision-making self-efficacy, perceived stress, and an integrated model of student persistence: A structural model of finances, attitudes, behavior, and career development, *Research in Higher Education,* vol. 41, p. 537-580.

THERNEAU, T. M. et B. ATKINSON, transféré sur R par Brian Ripley. 2009. *Rpart: Recursive partitioning and regression trees*, Vienne, R Foundation for Statistical Computing.

THOMAS, S. L. 2000. Ties that bind: A social network approach to understanding student integration and persistence, *Journal of Higher Education,* vol. 71*,* p. 591-615.

TINTO, V. 1975. Dropout from higher education: A theoretical synthesis of recent research, *Review of Educational Research,*vol. 45, p. 89-125.

TINTO, V. 1993. *Leaving college: Rethinking the causes and cures of student attrition,* 2e éd., Chicago, Presses universitaires de Chicago.

UNIVERSITÉ YORK. 2007. Rapport 2006-2007 dans le cadre de l'entente pluriannuelle de reddition de comptes (2006-07 multi-year accountability agreement report-back), Université York, Toronto.